Muñoz-Mas, R.; Gil-Martínez, E.; Oliva-Paterna, FJ.; Belda, E.; Martinez-Capel, F. (2019). Tree-based ensembles unveil the microhabitat suitability for the invasive bleak (Alburnus alburnus L.) and pumpkinseed (Lepomis gibbosus L.): Introducing XGBoost to eco-informatics. Ecological Informatics. 53:1-12. https://doi.org/10.1016/j.ecoinf.2019.100974

Tree-based ensembles unveil the microhabitat suitability for the invasive bleak (*Alburnus alburnus* L.) and pumpkinseed (*Lepomis gibbosus* L.): introducing XGBoost to eco-informatics

Rafael Muñoz-Mas[12*], Enric Gil-Martínez[1], Francisco J. Oliva-Paterna[3], Eduardo J. Belda[1], Francisco Martínez-Capel[1]

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

[2]GRECO, Institute of Aquatic Ecology, University of Girona, 17003 Girona, Catalunya, Spain.

[3]Departamento de Zoología y Antropología Física, Universidad de Murcia, Avda. Teniente Flomesta 5, 30003 Murcia, Spain.

*Correspondence to: Rafael Muñoz-Mas, e-mail: rafa.m.mas@gmail.com, voice: +34 622098521

Keywords

Conditional random forests

eXtreme Gradient Boosting machine

Gradient boosting machine

Oblique random forests

Random forests

SMOTE

Abstract

Random Forests (RFs) and Gradient Boosting Machines (GBMs) are popular approaches for habitat suitability modelling in environmental flow assessment. However, both present some limitations theoretically solved by alternative tree-based ensemble techniques (*e.g.* conditional RFs or oblique RFs). Among them, eXtreme Gradient Boosting machines (XGBoost) has proven to be another promising technique that mixes subroutines developed for RFs and GBMs. To inspect the capabilities of these alternative techniques, RFs and GBMs were compared with: conditional RFs, oblique RFs and XGBoost by modelling, at the micro-scale, the habitat suitability for the invasive bleak (*Alburnus alburnus* L.) and pumpkinseed (*Lepomis gibbosus* L.). XGBoost outperformed the other approaches, particularly conditional

and oblique RFs, although there were no statistical differences with standard RFs and GBMs. The partial dependence plots highlighted the lacustrine origins of pumpkinseed and the preference for lentic habitats of bleak. However, the latter depicted a larger tolerance for rapid microhabitats found in run-type river segments, which is likely to hinder the management of flow regimes to control its invasion. The difference in the computational burden and, especially, the characteristics of datasets on microhabitat use (low data prevalence and high overlapping between categories) led us to conclude that, in the short term, XGBoost is not destined to replace properly optimised RFs and GBMs in the process of habitat suitability modelling at the micro-scale.

1 Introduction

Biological invasions have increased in recent decades due to globalization and human activities, which is currently one of the main threats to freshwater biodiversity (Clusa et al., 2018). Flow regimes are leading factors in controlling fish population dynamics (Poff et al., 1997). Therefore, in several Mediterranean ecosystems, the alteration of the natural intra- and inter-annual flow variations, together with the creation of lentic habitats related to flow regulation infrastructures (e.g. reservoirs and weirs), have favoured the establishment of non-native invasive species (Clavero et al., 2013; Muñoz-Mas et al., 2016a; Ribeiro and Collares-Pereira, 2010). In Iberian rivers, two of the most prominent invasive fish species, which have been favoured by the anthropic modification of flow regimes, are the bleak (*Alburnus alburnus* L., 1758) and pumpkinseed (*Lepomis gibbosus* L., 1758) (Ilhéu et al., 2014).

The bleak inhabits open waters of lakes and medium-to-large rivers conforming large aggregations in backwaters and other still waters (Kottelat and Freyhof, 2007). In accordance, bleak has been considered as a limnophilic fish (Harby et al., 2007), although in others studies, performed in Mediterranean streams, the species has been categorized as eurytopic because it dwelled preferably run-type habitats with appreciable flow velocity (Masó et al., 2016; Muñoz-Mas et al., 2016d). The same is applicable to the pumpkinseed, which despite the lacustrine origins has been found to occur more often than expected in lotic habitats (Vilizzi et al., 2012).

64 Both species may have benefited from river regulation (Muñoz-Mas et al., 2016d; Vilizzi et
65 al., 2012). In accordance, environmental flows could be designed to counteract the
66 proliferation of these unwanted fish species (Acreman et al., 2014). However, effective
67 management strategies can only be developed with a good knowledge of how different
68 management alternatives affect invasive species (Thomsen et al., 2014). It is therefore
69 necessary to understand the critical habitat requirements to design better conservation and
70 management plans (Fukuda and De Baets, 2016). Ecological modelling has demonstrated to
71 be a supportive approach to develop spatial and temporal projections of the ecosystem
72 status under different flow regimes, resulting in more effective and less uncertain
73 management decisions (Stoffels et al., 2018).

74 In this regard, micro-scale habitat suitability models combining hydraulic variables such as
75 flow velocity or water depth are among the most popular for environmental flow
76 assessment (Nguyen et al., 2018). Although the univariate habitat suitability criteria is living
77 a sort of revival (see e.g. Gobeyn et al., 2017), the discipline has evolved significantly since
78 the initial concept developed by Waters (1976). In accordance, machine learning techniques
79 are currently used as tools for modelling the habitat suitability as well as for revealing
80 important environmental predictors and specific habitats required by the target species
81 (Fukuda and De Baets, 2016). However, modellers went one step beyond and, currently,
82 they advocate the use of machine learning techniques based on model ensembles instead
83 on single models (Clavero et al., 2017; Muñoz-Mas et al., 2016b). The main idea behind this

84 approach is to aggregate diverse models to obtain combined predictions outperforming
85 that of any single component (Bourel et al., 2017; Marmion et al., 2008; Ren et al., 2016).

86 The use of model ensemble has been favoured by two independent issues. On the one
87 hand, the interaction between data and modelling approach – which can render different
88 results (Fukuda et al., 2014; Muñoz-Mas et al., 2016b) – make it difficult to select the most
89 appropriate methodology, especially nowadays when a vast number of techniques is
90 available (Thuiller et al., 2009). On the other hand, model ensembles have mathematical
91 characteristics that lead to superior performance (see e.g. Dietterich, 2000; Ren et al., 2016
92 for thorough descriptions).

93 In ensemble modelling, model diversity is paramount to improve accuracy but also to
94 prevent over-fitting (Ren et al., 2016). Diversity can be induced by training each single
95 model on different aspects of the training dataset (data diversity) (Brown et al., 2005) or by
96 employing different modelling techniques and/or hyper-parameter settings (Muñoz-Mas et
97 al., 2016b; Ren et al., 2016). A number of approaches exist to induce data diversity (Brown
98 et al., 2005). However, the most famous approaches are *bagging* (Breiman, 1996) and
99 *boosting* (Freund and Schapire, 1997), whose popularity has increased hand in hand with
100 the popularity of tree-based approaches (Ren et al., 2016) because decision trees provide
101 different results when the training dataset is varied (Breiman, 2001).

102 The most important technique resulting from the combination of *bagging* and decision
103 trees is Random Forests (RFs) (Breiman, 2001). It consists of ensembles of Classification And

104 Regression Trees (CARTs) (Breiman et al., 1984) that are trained by resampling with
105 replacement the input samples (*n*) to develop each decision tree, and while developing each
106 decision tree the input predictor variables or features (*p*) are also resampled at every split,
107 data partition or node (Breiman, 2001).

108 Nevertheless, over the years a number of limitations of RFs arose, which triggered the
109 development of a myriad of alternative implementations. Among them, one of the most
110 popular approaches have been the conditional RFs (Hothorn et al., 2005; Strobl et al., 2008,
111 2007). This alternative approach solved the variable-selection bias of the original
112 implementation towards variables offering many potential cut-points, which artificially
113 increases the usefulness of variables that are continuous or with many categories (Strobl et
114 al., 2007).

115 Another relevant aspect of the individual trees of the forest is the orientation of the splits,
116 which are typically carried out with axis-parallel planes that may be sub-optimal (Muñoz-
117 Mas et al., 2016a) and ecologically unreliable because they render stair-like decision
118 surfaces (Menze et al., 2011). Breiman (1984) proposed the use of multivariate or oblique
119 splits, which drove the development of oblique RFs (Menze et al., 2011). The existence of
120 oblique RFs has been acknowledged in a number of studies (e.g. Fukuda et al., 2014;
121 Muñoz-Mas et al., 2016a). However, to the best of our knowledge they have not been used
122 to develop habitat suitability models so far.

123 Within the *boosting*-based group of ensembles, AdaBoost (Freund and Schapire, 1997) and
124 particularly the superseding Gradient Boosting Machines (GBMs) (Friedman, 2002, 2001)
125 enjoy great popularity, especially after the working guide published by Elith et al. (2008).
126 However, while b*agging*-based ensembles reduce the variance of the aggregated
127 predictions, *boosting*-based approaches reduce the bias and only subsidiarily, the variance
128 (Ren et al., 2016). This difference is caused by the differences in the resampling strategy:
129 *bagging* performs resampling (with replacement) and is based on constant resampling
130 probabilities (i.e. each sample is used to train similar number of trees) whereas *boosting*
131 iteratively varies the resampling probabilities (see Fig. 3 in methods for a graphical
132 depiction). This implies that *boosting*-based ensembles are built in a sequential manner
133 gradually increasing the emphasis on the observations poorly modelled by adding trees
134 until the data are completely over-fitted (Elith et al., 2008; Gómez-Ríos et al., 2017).
135 Developers were soon aware of this phenomenon. Therefore, the different *boosting*-based
136 approaches included subroutines to prevent over-fitting (e.g. the shrinkage parameter of
137 GBMs) (Gómez-Ríos et al., 2017).

138 This process of refinement has not discontinued and it triggered the development of
139 additional subroutines to reduce over-fitting – some of them based on the RFs approach –
140 which ended up in the development of the eXtreme Gradient Boosting machines (XGBoost)
141 (Chen and Guestrin, 2016). XGBoost has turned out to be one the most promising
142 algorithms to classify both binary and multi-class datasets, especially in the former case
143 (Gómez-Ríos et al., 2017).

144 To the best of our knowledge XGBoost has never been tested with ecological datasets.
145 Therefore, to increase the knowledge about the environmental limitations for bleak
146 (*Alburnus alburnus* L.) and pumpkinseed (*Lepomis gibbosus* L.), the habitat suitability for
147 these invasive species has been modelled with five state-of-the-art tree-based ensemble
148 approaches. The tree-based approaches compared were: standard RFs (Breiman, 2001),
149 GBMs (Friedman, 2002, 2001), conditional RFs (Hothorn et al., 2005; Strobl et al., 2008,
150 2007), oblique RFs (Menze et al., 2011) and XGBoost (Chen and Guestrin, 2016). The
151 remainder of the paper is structured as follows: section 2 describes the training datasets,
152 the approach followed to optimise each tree-based ensemble and the particularities of the
153 compared techniques. This section also describes how the models were compared. In
154 section 3, the accuracy of the five different tree-based approaches and the reliability of the
155 modelled habitat suitability are presented. In section 4, the results are discussed, and
156 integrated with current literature; finally, the conclusions and caveats are provided in the
157 section 5.

158

159 # 2 Methods

160 ## 2.1 Data collection

161 The survey was performed in a segment of the Jucar River located downstream of the Tous
162 Dam where both species co-occur (Fig. 1). The data collection was performed in 2017 at the
163 microhabitat scale by systematically sampling areas of few m^2 with homogeneous depth,

velocity, substrate and cover. Each of these small areas was a sampling unit where presence/absence and the environmental variables were recorded. The survey was performed during high water temperature and low flows (Q = 8.74 m^3/s) to mimic summertime conditions (i.e. the period of lower flows in non-flow-regulated Mediterranean rivers) that, given the irrigation purpose of the upstream reservoir and latitude of the study site, occur in September. The data collection was performed in a river segment that encompasses areas of deep pools and relatively shallow rapids and runs, all of them with optimum visibility. Therefore, the microhabitat study was conducted by underwater observation (snorkelling) and the sampling balanced the areas of deep-slow and shallow-fast flow types (Muñoz-Mas et al., 2016c, 2012). Microhabitat characteristics for the presence data were measured at fish locations (i.e. where bleak and pumpkinseed individuals of body length > 5 cm were observed). On the contrary, the absence data were collected following a systematic sampling approach (Bovee, 1986), which consists of measuring the microhabitat characteristics over a regular grid.

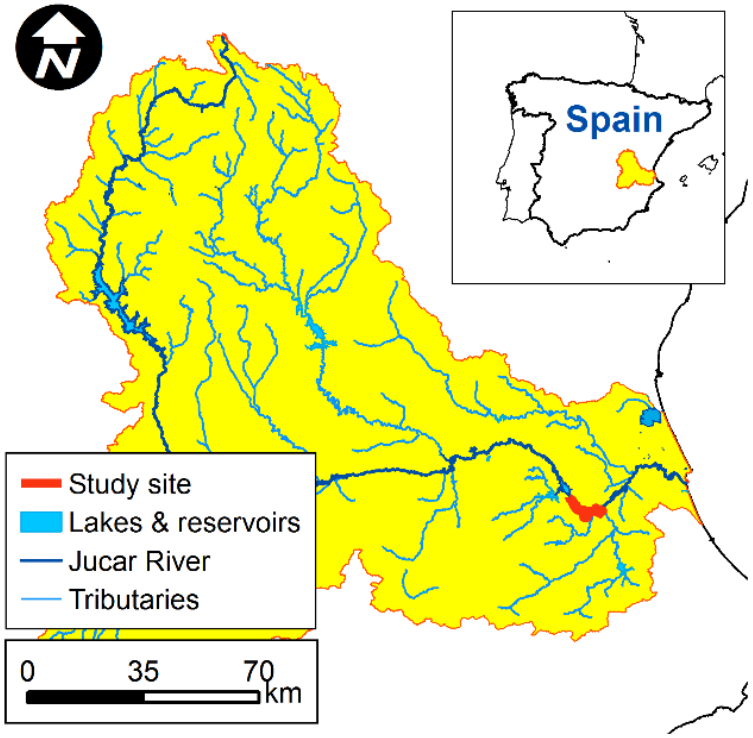


Fig. 1. Location of the river segment (in red) in the Jucar River Basin (in yellow) where the microhabitat preferences of the invasive bleak (*A. alburnus*, L.) and pumpkinseed (*L. gibbosus* L.) were studied.

Three hydraulic variables (depth, mean flow velocity and substrate composition) were used to characterise each sampling unit or sampled microhabitat (Muñoz-Mas et al., 2012). In addition, the presence and dimensions of five cover types were scored. They encompassed the concept of structural (*e.g.* boulders, log jams) and escape (*e.g.* aquatic vegetation) cover and corresponded to reeds, aquatic vegetation, shade, rocks and log jams and small woody debris (Muñoz-Mas et al., 2016c). The scale used was that described by Muñoz-Mas et al. (2016c), which ranges from 0 (absent) to 3 (massive). Depth (m) was measured with a wading rod (to the nearest cm). Mean flow velocity (m/s) was measured with an electromagnetic current meter (Valeport®, United Kingdom). The percentage of each

substrate class was visually estimated following a simplification of the American Geophysical Union size scale, which corresponds to, silt (Ø ≤ 62 µm), sand (62 µm > Ø ≤ 2 mm), fine gravel (2 > Ø ≤ 8 mm), gravel (8 > Ø ≤ 64 mm), cobble (64 > Ø ≤ 256 mm), boulder (Ø > 256 mm) and bedrock (Muñoz-Mas et al., 2012); these percentages were aggregated into the dimensionless substrate index (Mouton et al., 2011), which ranges between 0 (silt) and 8 (bedrock). In the end, bleak was observed in 98 microhabitat and pumpkinseed in 42 whereas the available/unoccupied conditions (sampling units where none of the target species was observed) were measured in 1258 microhabitats. In accordance, the data prevalence (i.e. the proportion of presences in the dataset) for each species corresponded to 0.07 and 0.03 respectively for bleak and pumpkinseed (Fig. 2).

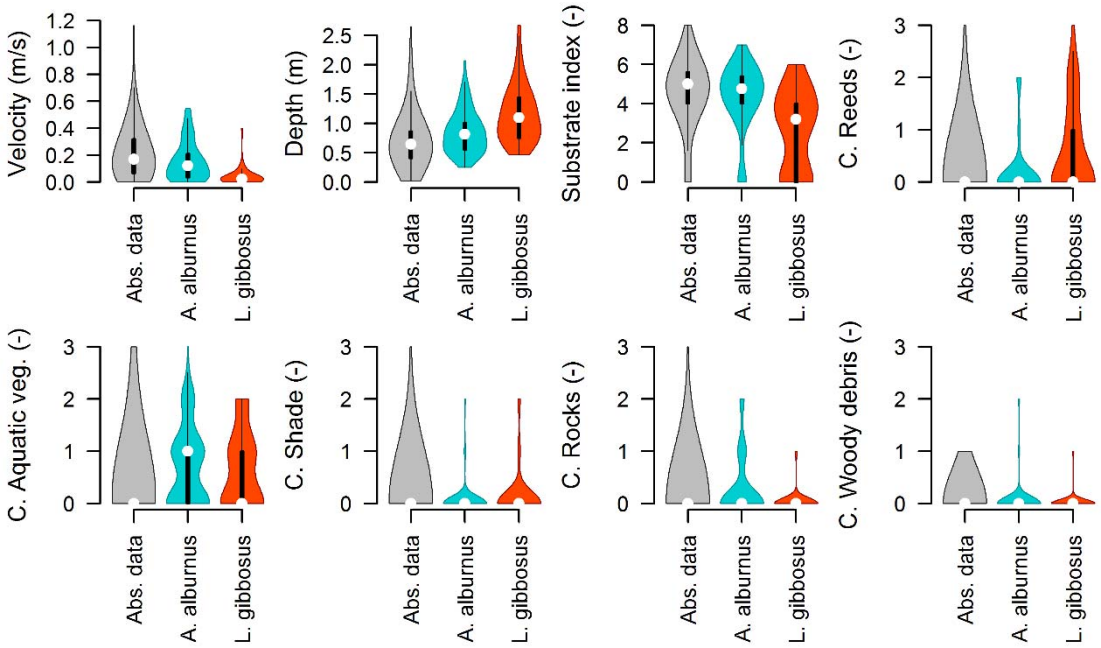


Fig. 2. Violin plots of the absence data (Abs.) and the microhabitats occupied (presence data) by the invasive bleak (*A. alburnus*, L.) and pumpkinseed (*L. gibbosus* L.).

2.2 Data preparation for cross-validation - Synthetic Minority Over-sampling TEchnique (SMOTE)

The parameters that maximise tree-based ensembles performance (*i.e.* hyper-parameters) were sought performing repeated *k*-fold, a type of cross-validation that implies partitioning several times the dataset into *k* equal-sized parts. For each of these repetitions, the tree-based ensemble is trained *k* times using *k*−1 parts and, each time, the performance is evaluated using the *k* part held out. However, even after cross-validation, low data prevalence (i.e. imbalance between the presence and absence classes) may impact model performance and reliability (Fukuda and De Baets, 2016). Several solutions to deal with this problem have been proposed and implemented, both for individual models and for ensemble approaches. These can be summarised into: i) data resampling (i.e. over-sampling and/or under-sampling), ii) algorithm modification or iii) cost-sensitive learning (López et al., 2013). Lamentably, the implementation of these approaches is uneven. For instance, the current implementation of RFs in the *R* package *randomForest* (Liaw and Wiener, 2002) allows data resampling and class weighting, whereas that to develop oblique RFs (i.e. *obliqueRF* - Menze et al., 2012) does not allow any of those solutions. In accordance, for each dataset involved in the ensemble training (*i.e.* each part of the complete dataset used for model training), we applied the Synthetic Minority Oversampling TEchnique (SMOTE) (Chawla et al., 2002) whereas the datasets used for ensemble validation (*i.e.* each part of the complete dataset held out and used to test model performance) remained unaltered.

226 This technique is one of the most renowned resampling approaches to deal with low
227 prevalence datasets (López et al., 2013) and it is independent to the software package
228 involved in the development of each kind of tree-based ensemble. SMOTE simultaneously
229 over-samples the minority class (i.e. presence data) and under-samples the majority class
230 (absence data) to get the desired prevalence – in this case 0.5 – (Chawla et al., 2002).
231 However, SMOTE instead of simply resampling with replacement the presence data creates
232 synthetic data. Therefore, the presence data is over-sampled by taking each presence data
233 and introducing synthetic additional data along the line segments joining any/all of the *k*
234 nearest neighbours (López et al., 2013).

235 The implementation of SMOTE was that included within the *R* package *DMwR* (Torgo,
236 2010). In order to accelerate model training within the $3 \times 3\ cross - validation$ (nine
237 models), we selected parameters that ended up with balanced datasets of inferior
238 dimensions (i.e. $N_{SMOTE} < N_{Original}$, and $n_{presence} \approx n_{absence}$). Therefore, the number of new
239 instances generated for each presence (i.e. the parameter named *perc.over*) corresponded
240 to 20 by the ratio between the number of absence data and that of presence data (i.e.
241 255±2 for bleak and 596 for pumpkinseed); the parameter *k* corresponded to 10% of the
242 number of presence data; and the number of absence data randomly selected for each
243 *smoted* observation (i.e. the parameter named *perc.under*) was 1.25. These parameter
244 settings rendered balanced training datasets with 359±3 and 343 data (presence and
245 absence) respectively for bleak and pumpkinseed. Violin plots depicting the similarity

246 between the distributions of the original datasets and those generated by SMOTE can be
247 found in Appendix A.

248

249 ## 2.3 Tree-based ensembles optimisation

250 In the context of tree-based ensembles optimisation, two elements are recommended to be
251 adjusted to maximise model performance and to prevent over-fitting: the hyper-parameter
252 settings and the set of feature variables involved in the ensemble training (Martinez-de-
253 Pison et al., 2017). The hyper-parameters drive the tree-based ensemble growth (*e.g.* by
254 controlling the maximum number of terminal nodes in each decision tree or the number
255 trees in the forest) and they interact with the feature variables. In accordance, the optimal
256 hyper-parameters and feature variable set was simultaneously sought following a *wrapper*
257 approach involving cross-validation and the Genetic Algorithm (GA) (Holland, 1992)
258 implemented within the *R* package *rgenoud* (Mebane Jr and Sekhon, 2011). This approach
259 has proved markedly proficient to concomitantly search the optimal hyper-parameters and
260 feature variable set (see e.g. Martinez-de-Pison et al., 2017).

261 Optimisations based on GAs consist of initially generating a population of feasible solutions
262 (i.e. random combinations of variable sets and parameters within the feasible range). Then,
263 the performance criteria for each combination (chromosome/phenotype) is calculated, and
264 the search of the better solution (evolution) takes place by more frequently crossing
265 (reproducing) those individuals with better performance, but regularly altering (mutating)

266 the remaining individuals to enhance the sampling of the potential combinations of
267 parameters and variables (Muñoz-Mas et al., 2016a).

268 Following previous experiences (Muñoz-Mas et al., 2018, 2016a), the optimisation was
269 performed following a repeated k-fold scheme – described in section 2.2 – but the
270 performance criterion maximised exclusively the mean True Skill Statistic (TSS) (Allouche et
271 al., 2006) because the training datasets presented optimal data prevalence (i.e. 0.5). The
272 TSS proved good behaviour on low prevalence datasets (Somodi et al., 2017) and it consists
273 of the sum of Sensitivity (*Sn*) and Specificity (*Sp*) minus one (i.e. $TSS = Sn + Sp - 1$). The
274 *Sn* corresponds to the ratio of presences correctly classified and *Sp* corresponds to the ratio
275 of absences correctly classified (see Mouton et al., 2010 for additional details about
276 performance criteria).

277 The nine different operators that govern the optimisation performed with *genoud* (Mebane
278 Jr & Sekhon, 2011) were selected to avoid premature convergence, as previously indicated
279 in Muñoz-Mas et al. (2017). The population size and number of generations varied in
280 accordance with the number of parameters and variables involved in the optimization,
281 after:

282 $$N_{population} = N_{generations} = 10 \times (\#\, Parametes + 8)$$

283 where $\#\, Parametes$ varied for each tree-based ensemble approach (Piotrowski, 2017) and
284 the 8 corresponded with the eight environmental variables used as predictor variables in
285 the models: velocity, depth, substrate and the five different cover types (reeds, aquatic

286 vegetation, shade, rocks and woody debris). Finally, the optimisation halted after
287 $\#\,Parametes + 8$ generations without improvement. The chromosomes were composed
288 of integers; the first part, which varied in length, encompassed the tree-based ensemble
289 parameters whereas the second part was composed of an 8-bit string equalling the number
290 of variables within the training dataset (see Muñoz-Mas et al., 2016a for additional details
291 on chromosome structure). In order to properly sample the searching space a Latin
292 hypercube, as implemented in the *R* package *lhs* (Carnell, 2016), was used to initialise the
293 population instead of using the random generator because it reduces the computational
294 burden and facilitates the algorithm convergence to satisfactory solutions (Knowles and
295 Hughes, 2005). Once the best hyper-parameters and input variables were determined, a
296 single tree-based ensemble was trained employing the entire dataset (i.e. without cross-
297 validation) to perform additional analyses (Muñoz-Mas et al., 2016a).

298

299 **2.3.1 Random Forests – RFs**

300 Random Forests (RFs) (Breiman, 2001) is the first exponent of the *bagging* group of tree-
301 based ensemble approaches (Fig. 3). The RFs model was developed employing the *R*
302 package *randomForest* (Liaw and Wiener, 2002). In accordance, in addition to the ultimate
303 variable set (Muñoz-Mas et al., 2016a), four elements required optimisation to prevent
304 over-fitting, namely the parameters constraining the tree growth (*nodesize* and *maxnodes*)
305 (Muñoz-Mas et al., 2016a), the percentage of random samples used to train each individual

tree of the forest (*sampsize*) and the randomness introduced in every recursive binary split (*mtry*) (Strobl et al., 2009). The range tested for each parameter were based on the dimensions (i.e. $n \times p$) of the training datasets (Table 1). Finally, the number of trees was set to 250, although the stabilization of the error in the *out-of-bag* (i.e. the dataset held out of each decision tree) was inspected to ascertain the adequacy of this number.

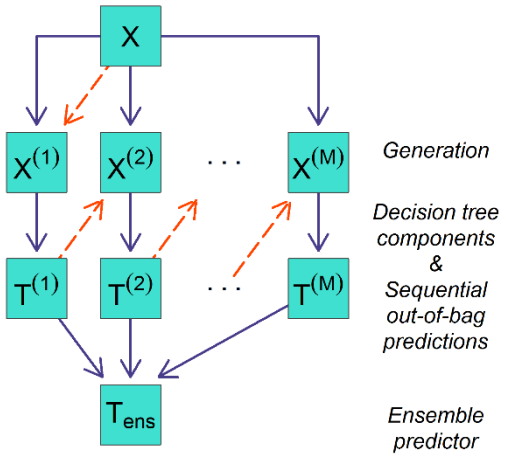


Fig. 3. Framework for *bagging* and *boosting* based decision tree ensemble construction, *X* is the original dataset, $X^{(i)}, i \in \{1, \ldots, M\}$ are the generated datasets with M equal to the number of decision trees in the ensemble, $T^{(i)}$ are the individual decision trees of the ensemble (trained by resampling the original dataset *X*) and T_{ens} is the final ensemble predictor (i.e. that combining every decision tree). The dashed arrows in the generation of the $T^{(i)}$ denote *boosting* related ensemble framework, whose predictions are sequentially used to update the resampling probabilities of the data used to train the subsequent decision trees (adapted from Ren et al., 2016).

2.3.2 Gradient Boosting Machines – GBMs

Gradient Boosting Machines (GMBs) are included within the *boosting* group of techniques (Friedman, 2002, 2001) that is built in a sequential manner by increasingly focusing on the observations more difficult to predict (Elith et al., 2008; Ren et al., 2016) (Fig. 3). The GBMs

were developed employing the *R* package *gbm* (Greenwell et al., 2019). In accordance, five parameters were optimised (Table 1), namely:

1. *n.trees*: number of *boosting* rounds or trees in the model.
2. *shrinkage*: learning rate or step-size reduction. It ranges between 0 and 1 and higher values preclude over-fitting. It is closely linked to the optimal value of *n.trees* (Ridgeway, 2007). Therefore, lower values of *shrinkage* may require larger values of *n.trees* in order to get adequate performance.
3. interaction.depth: maximum depth of variable interactions.
4. *n.minobsinnode*: minimum number of observations in the terminal nodes of the trees.
5. *bag.fraction*: the fraction of the training dataset randomly selected to propose the next tree in the expansion (i.e. resampling in *n*). This introduces randomness into the model.

The selected distribution for the outputs was *adaboost* and the ultimate number of trees involved in further predictions, which is usually inferior to the total number of trained trees (i.e. *n.trees*), was that minimising the out-of-bag estimate of the improvement in predictive performance because external cross-validation was performed (Ridgeway, 2007).

2.3.3 Conditional Random Forests – Conditional RFs

Conditional RFs (Hothorn et al., 2006; Strobl et al., 2007) are a member of the *bagging* group of tree-based ensemble approaches that solved the bias of the original RFs towards

variables continuous or with a large number of categories (Strobl et al., 2007). To prevent this behaviour conditional RFs perform a *permutation test* (Strasser and Weber, 1999) – under the null hypothesis of independence – to selected variables for additional splits and to determine when the tree growth must stop (Hothorn et al., 2006). The Conditional RFs model was developed employing the *R* package *party* (Hothorn et al., 2010). The default hyper-parameter settings controlling the unbiased tree growth are based on previous experiences (Strobl et al., 2007) and, in the package vignette, its modification is discouraged (Hothorn et al., 2015). Therefore, in addition to the ultimate variable set, only the parameter controlling the randomness introduced in every recursive binary split (*mtry*) was optimised. The tested values ranged between 1 and 8 in accordance with the maximum number of variables potentially included in the model (Table 1) and the number of individual trees of the forest was also set to 250.

2.3.4 Oblique Random Forests – Oblique RFs

Oblique RFs are a member of the *bagging* group of tree-based ensemble approaches, which instead of orthogonal trees with thresholds on individual features at every split, use oblique trees separating the feature space by unrestrictedly oriented hyperplanes (Menze et al., 2011). In accordance, they do not render the typical stair-like decision surfaces of other RFs approaches, which permits inferring smoother decision surfaces (Menze et al., 2011) (Fig. 4). The oblique RFs model was developed employing the *R* package *obliqueRF* (Menze et al.,

2012). Therefore, in addition to the ultimate variable set, only two parameters can be adjusted to prevent over-fitting namely, *mtry*, which controls the randomness introduced in every recursive binary split, and *training_method*, which indicate the multivariate models for binary splits used in each node. The tested values for *mtry* also ranged between 1 and 8 but the parameter *training_method* was set to *log* (i.e. logistic regression) in order to rank the variable importance (Table 1). Finally, the number of individual trees of the forest was set to 250.

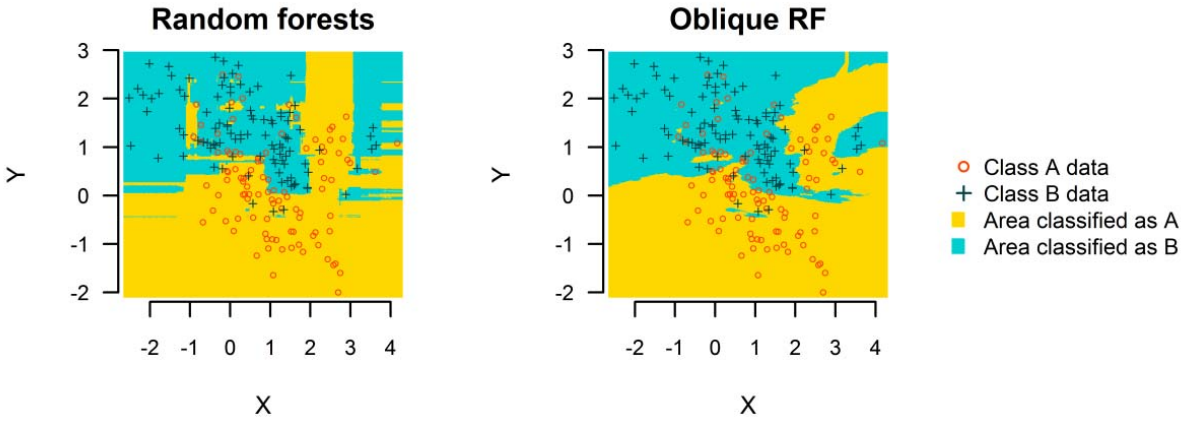


Fig. 4. Visualization of the typical stair-like decision border obtained with an axis-parallel tree-based ensemble approach (i.e. Random Forests - RFs) and smooth decision border obtained with oblique RFs (adapted from Menze et al., 2011).

2.3.5 eXtreme Gradient Boosting machines - XGBoost

EXtreme Gradient Boosting machines (XGBoost) are a kind of *boosting*-based ensemble machine leaning technique (Chen and Guestrin, 2016). To prevent over-fitting, XGBoost and its implementation within the homonymous *R* package (Chen et al., 2017) include a number

of specific routines; some of them formerly envisaged for random forests and other modelling techniques. XGBoost is a new tree-based ensemble technique; thus, the nine parameters finally optimised and the ranges employed slightly varied compared to previous studies (Gómez-Ríos et al., 2017; Martinez-de-Pison et al., 2017; Xiao et al., 2017). Table 1 includes a summary of the nine parameters optimised, whose impact is described in the XGBoost manual (Chen et al., 2017), and they are:

6. *nrounds*: number of *boosting* rounds or trees in the model.
7. *eta*: step size shrinkage used in each update to prevents over-fitting. It is analogous to the *shrinkage* parameter described for GBMs. It typically lies between 0.01 - 0.3.
8. *gamma*: a pseudo-regularization parameter determining the minimum loss reduction required to make further partitions on each individual tree of the forest. It ranges from 0 to ∞. Larger values (up to 20) may prevent over-fitting, although when too large it may impede an adequate performance. The parameter *gamma* brings improvement when shallow trees are desired (small values of *max_depth*).
9. *min_child_weight*: minimum sum of weighted data needed in child nodes to perform a further partition. It is usually one. Larger values may prevent over-fitting.
10. *max_depth*: maximum depth/partitions allowed in each individual tree of the forest; 0 indicates unlimited number of partitions.
11. *subsample*: percentage of random samples used to train each individual tree of the forest. It is analogous to the *bag.fraction* parameter described for GBMs. It typically lies between 0.5 - 0.8.
12. *colsample_bytree*: parameter controlling the randomness introduced in every recursive binary split (i.e. resampling in *p*). It is equivalent to random forests *mtry*, although it is specified as a percentage. It typically lies between 0.5 - 0.9.

13. *alpha*: L1 regularization term on weights (analogous to *Lasso* regression). It ranges from 0 to ∞. Larger values prevent over-fitting. In addition to shrinkage, enabling alpha also results in feature selection. In accordance, it is more useful on high dimensional datasets.
14. *lambda*: L2 regularization term on weights (analogous to *Ridge* regression). It ranges from 1 to ∞. Larger values prevent over-fitting.

Table 1. Name, range, accuracy and description of the optimised parameters for the five alternative tree-based ensemble approaches. Additional parameters and the *R* packages employed are indicated in the corresponding sections.

Method	Parameter	Range	Accuracy	Description
Random Forests (RFs)	mtry	[1, 8]	1	Number of variables randomly sampled as candidates at each split
	nodesize	[1, 50]	1	Minimum number of samples at each terminal node/leaf
	maxnodes	{2,...,100, ∞}	1	Maximum number of terminal nodes/leaves
	sampsize	[0.5, 1]	0.01	Number of samples randomly sampled to train each tree
Gradient Boosting Machines	n.trees	[10, 5000]	1	Number of trees
	shrinkage	[0.01, 0.4]	0.01	Shrinkage parameter/learning rate
	interaction.depth	[1, 8]	1	Maximum depth of variable interactions
	n.minobsinnode	[1, 50]	1	Minimum number of samples at each terminal node/leaf
	bag.fraction	[0.5, 0.99]	0.01	Number of samples randomly sampled to train each tree
Conditional RFs	mtry	[1, 8]	1	Number of variables randomly sampled as candidates at each split
Oblique RFs	mtry	[1, 8]	1	Number of variables randomly sampled as candidates at each split
XGBoost	nrounds	[10, 5000]	1	Number of trees
	eta	[0.01, 1]	0.01	Shrinkage parameter/learning rate
	gamma	[0, 50]	0.05	Minimum loss reduction to permit additional partitions
	min_child_weight	[1, 50]	1	Minimum number of samples at each terminal node/leaf
	max_depth	{1,...,100, ∞}	1	Maximum number of terminal nodes/leaves
	subsample	[0.5, 1]	0.01	Number of samples randomly sampled to train each tree
	colsample_bytree	[0.5, 1]	0.01	Number of variables randomly sampled as candidates at each split
	alpha	[0.0, 50]	0.05	L1 regularization term on weights
	lambda	[1.0, 50]	0.05	L2 regularization term on weights

2.4 Tree-based ensemble comparison and ecological significance

To determine statistical differences between the performance criteria obtained with the five tree-based ensemble approaches, the non-parametric Friedman aligned ranks test (Friedman, 1940) was calculated employing the values of the TSS calculated for the nine validation datasets obtained during the cross-validation (García et al., 2010; García and Herrera, 2008). The *p-values* of the test were adjusted applying the Bergmann and Hommel correction (Bergmann and Hommel, 1988), as previously indicated in García and Herrera (2008), and the results were graphically characterised employing the function *drawAlgorithmGraph,* implemented in the *R* package *scmamp* (Calvo and Santafé, 2016). This function plots a graph where the tree-based ensemble approaches are the nodes and they appear linked when the null hypothesis of being equal cannot be rejected. Finally, the component performance criteria (i.e. *Sn* and *Sp*) and the accuracy or Correctly Classified Instances (CCI) were also inspected.

The variable importance was examined employing the *R* functions implemented in the corresponding packages (i.e. *importance*, *varimp*, *importance* and *xgb.importance*). These packages render the variable importance in different scales. Therefore, in order to facilitate an adequate comparison, the resulting importance was standardised dividing the values by the value of the largest importance. Then, mean variable importance and confidence intervals were compared.

436 Finally, the modelled relationship between the selected variables and the habitat suitability
437 for bleak and pumpkinseed was graphically characterised with partial dependence plots
438 (Friedman, 2001). Partial dependence plots depict the average of the response variable
439 *versus* the inspected variable and account for the effects of the remaining variables within
440 the model by averaging their effects (Muñoz-Mas et al., 2018). The partial dependence plots
441 were calculated adapting the code appearing in the *R* package *randomForests* (Liaw and
442 Wiener, 2002).

443

444 # 3 Results

445 ## 3.1 Best hyper-parameter settings

446 The best tree-based ensembles for each species were obtained with different input
447 variables and hyper-parameters (Table 2). The Random Forests (RFs) *mtry* was generally low
448 but different for each species, although the optimal number of samples per terminal node
449 (*nodesize*) coincided. Conversely, the tree depth (*maxnodes*) and per cent of the training
450 dataset resampled (*sampsize*) markedly differed. The values of *shrinkage* (learning rate) for
451 the Gradient Boosting Machines (GBMs) were similar, but those of *n.trees* were not; the
452 one for the pumpkinseed was markedly smaller. The *interaction.depth* and, particularly, the
453 *n.minobsinnode* were significantly different for both species. Conversely, the values of
454 *bag.fraction* were of intermediate magnitude in both cases. The *mtry* for conditional and
455 oblique RFs presented similar value for each species, although those of oblique RFs were

456 smaller. Regarding the eXtreme Gradient Boosting machines (XGBoost), intermediate values
457 of *nrounds* (number of trees) and *eta* (learning rate) were obtained for both species. The
458 loss reduction to permit additional partitions (*gamma*) and minimum number of samples
459 per terminal node (*min_child_weight*) differed. Conversely, the maximum depth of the trees
460 (max_depth) and the parameters governing resampling in $n \times p$ (*subsample* and
461 *colsample_bytree*) were similar. The regularization parameters (*alpha* and *lambda*) were
462 twofold higher for bleak than they were for pumpkinseed.

463

464 Table 2. Best parameters obtained for the five different tree-based ensemble approaches
465 obtained after $3 \times 3\ cross - validation$.

Method	**Parameter**	**Bleak (*A. alburnus* L.)**	**Pumpkinseed (*L. gibbosus* L.)**
Random Forests (RFs)	*mtry*	1	2
	nodesize	9	9
	maxnodes	23	5
	sampsize	273 (≈76%)	186 (≈54%)
Gradient Boosting Machines/Boosted Regression Trees	n.trees	3076	815
	shrinkage	0.16	0.13
	interaction.depth	4	1
	n.minobsinnode	4	46
	bag.fraction	0.87	0.73
Conditional RFs	*mtry*	6	3
Oblique RFs	*mtry*	4	2
XGBoost	*nrounds*	2499	3781
	eta	0.34	0.55
	gamma	0.25	13.45
	min_child_weight	1	12
	max_depth	35	39
	subsample	0.83	0.80
	colsample_bytree	0.90	0.78
	alpha	7.10	3.65
	lambda	30.60	16.65

466

3.2 Performance of the tree-based ensembles

XGBoost outperformed the other approaches (i.e. achieved statistically higher values of the True Skill Statistic – TSS), particularly the conditional and oblique RFs (Fig. 5 & Table 3). However, the corrected Friedman aligned ranks test indicated no statistical difference between XGBoost, RFs and GBMs because the mean TSS obtained with RFs for bleak was higher than that obtained with XGBoost while GBMs achieved high values for both species (Table 3). The statistical test indicated no significant difference between the conditional and oblique RFs. Consequently, these two are connected in Fig. 5.

The five tree based approaches were over-predictive (specificity ≤ sensitivity) both for the training and validation datasets (Table 3). Oblique RFs presented the highest performance in the four criteria during the training phase. However, this was not the general case in the validation phase. Conditional random forests presented intermediate performance in both cases; training and validation.

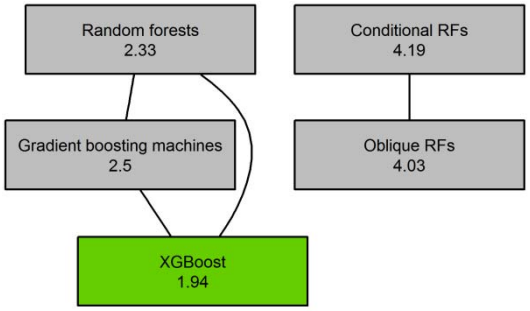


Fig. 5. Results of the corrected non-parametric Friedman aligned ranks test comparing the performance (nine values of TSS per species obtained during the cross-validation) of the five tree-based ensemble approaches. Green background highlights the tree-based ensemble approach with the best performance. The approaches appear linked when the null hypothesis of being equal was not rejected (no statistical difference).

Table 3. Performance criteria and confidence intervals to evaluate the five tree-based ensemble techniques: accuracy or Correctly Classified Instances (CCI), True Skill Statistics (TSS), Sensitivity (Sn) and Specificity (Sp), obtained during the 3×3 cross-validation (nine models). The best results for the objective criterion (i.e. mean TSS) appear in bold.

		Bleak (*A. alburnus* L.)				**Pumpkinseed (*L. gibbosus* L.)**			
		CCI	**TSS**	**Sn**	**Sp**	**CCI**	**TSS**	**Sn**	**Sp**
Training	Random forests (RFs)	0.83±0.01	0.65±0.02	0.87±0.01	0.78±0.01	0.86±0.01	0.71±0.02	0.91±0.01	0.80±0.01
	Gradient Boosting Machines	0.82±0.01	0.63±0.01	0.86±0.01	0.77±0.01	0.86±0.01	0.73±0.02	0.91±0.01	0.81±0.01
	Conditional RFs	0.83±0.01	0.65±0.01	0.87±0.01	0.77±0.02	0.86±0.01	0.73±0.02	0.91±0.01	0.82±0.01
	Oblique RFs	0.98±0.00	0.96±0.00	1.00±0.00	0.96±0.01	0.98±0.00	0.96±0.00	1.00±0.00	0.96±0.01
	XGBoost	0.80±0.01	0.58±0.01	0.86±0.01	0.72±0.01	0.85±0.01	0.70±0.02	0.90±0.01	0.80±0.01
Validation	Random forests (RFs)	0.72±0.01	**0.51±0.02**	0.80±0.02	0.71±0.01	0.79±0.01	0.66±0.03	0.87±0.03	0.79±0.01
	Gradient Boosting Machines	0.71±0.01	0.48±0.03	0.78±0.04	0.71±0.01	0.81±0.01	0.65±0.03	0.84±0.04	0.81±0.01
	Conditional RFs	0.70±0.01	0.44±0.02	0.75±0.01	0.69±0.01	0.80±0.01	0.61±0.03	0.81±0.04	0.80±0.01
	Oblique RFs	0.72±0.01	0.45±0.02	0.73±0.02	0.72±0.01	0.72±0.01	0.45±0.02	0.73±0.02	0.72±0.01
	XGBoost	0.69±0.01	0.49±0.02	0.82±0.03	0.68±0.01	0.80±0.01	**0.67±0.03**	0.87±0.04	0.79±0.01

3.3 Ecological significance – variable importance

In addition to the different variable set selected, each tree-based ensemble approach rendered different variable ranking (Fig. 6). The continuous variables selected for XGBoost (i.e. depth and velocity) presented higher importance compared to these of cover. A similar pattern was found in RFs and GBMs. Conversely, oblique RFs, as well as the conditional RFs, indicated higher importance for the variables related with cover. Based on the selected variables, cover was more important for bleak than it was for pumpkinseed; especially reeds, aquatic vegetation and rocks, which were selected in the five approaches. Pumpkinseed selected microhabitats with reeds and shade. Substrate composition was considered of minor importance and only conditional and oblique RFs selected it for bleak.

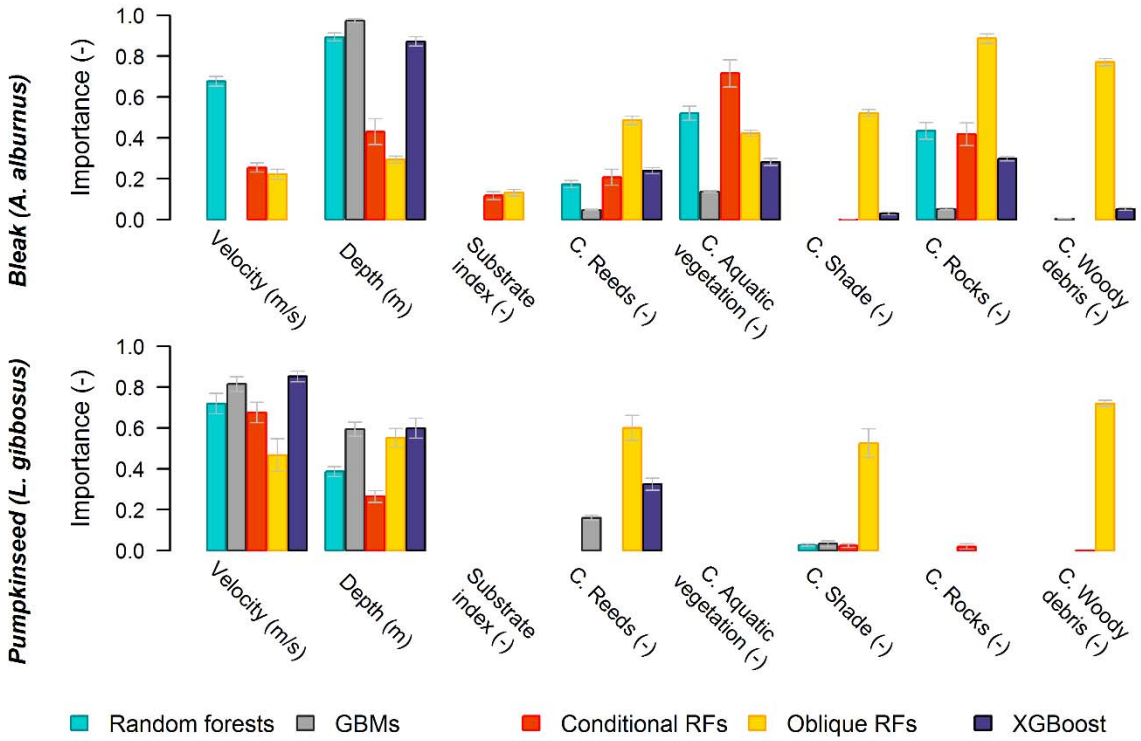
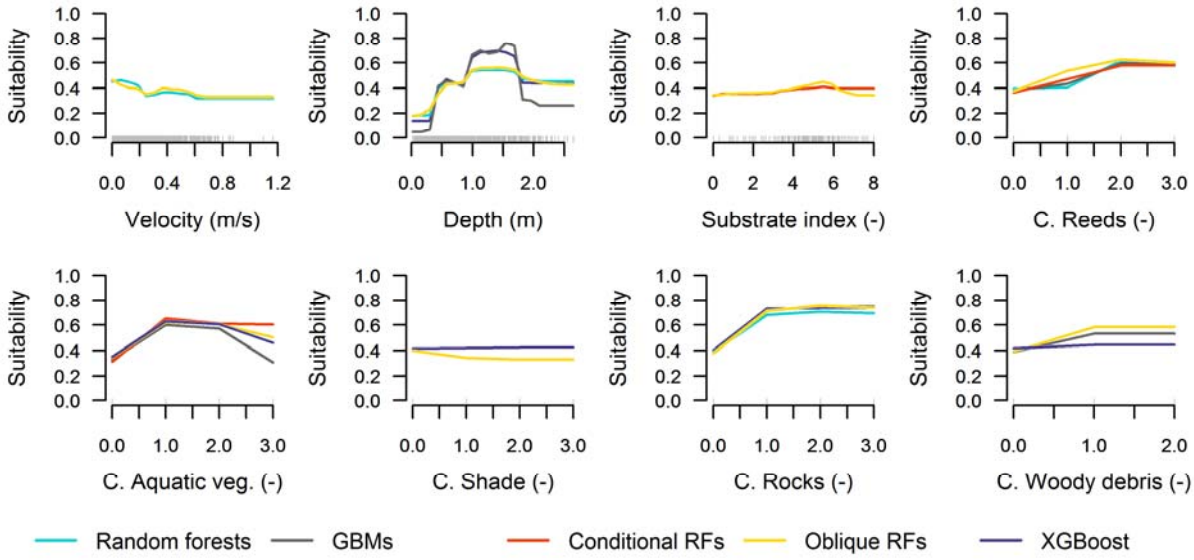


Fig. 6. Mean variable importance computed with the nine tree-based model ensembles (3x3 cross validation) per technique. Error bars show the 0.99 confidence interval.

3.4 Ecological significance – partial dependence plots

Unlike the variable importance rankings, the partial dependence plots were largely consistent among the five approaches (Fig. 7 and Fig. 8). The highest suitability for bleak was obtained for deep (> 1 m) microhabitats with low flow velocity (flow velocity below 0.4 m/s), reeds, aquatic vegetation and, especially, rocks (Fig. 7). Shade and woody debris presented very small or no effect. For GBMs and oblique RFs, the plot for woody debris indicated positive effects, unlike for XGBoost. Finally, conditional and oblique RFs indicated higher suitability over coarse substrates (from fine gravel to boulders).



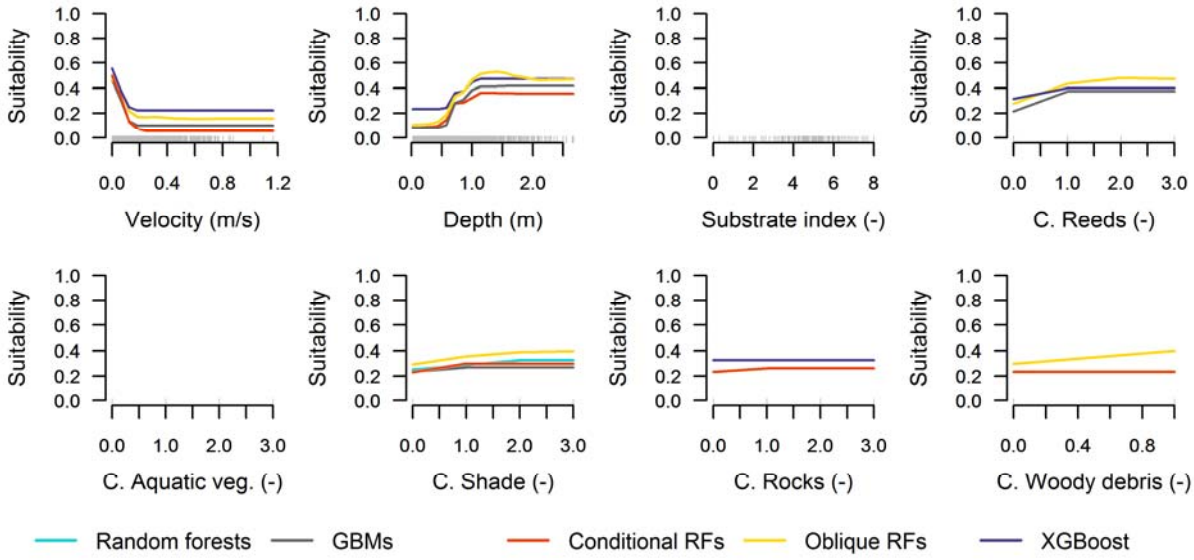
516

517 Fig. 7. Partial dependence plots for bleak (*Alburnus alburnus* L.), obtained with the five tree-
518 based ensemble approaches, depicting the marginal relationship between the suitability
519 (i.e. probability of class Presence) and the selected microhabitat variables.

520

521 The highest suitability for pumpkinseed was obtained for stagnated (i.e. null flow velocity)
522 and deep (> 1m) microhabitats with presence of reeds and/or shade (Fig. 8). Rocks and
523 woody debris presented very small or no effect. Substrate and aquatic vegetation were not
524 selected by any tree-based ensemble approach.

525



526

527 Fig. 8. Partial dependence plots for pumpkinseed (*Lepomis gibbosus* L.), obtained with the
528 five tree-based ensemble approaches, depicting the marginal relationship between the
529 suitability (i.e. probability of class Presence) and the selected microhabitat variables.

530

531 # 4 Discussion

532 ## 4.1 Tree-based ensembles comparison

533 This study demonstrated that eXtreme Gradient Boosting machines (XGBoost) can
534 outperform other approaches, particularly the conditional and oblique Random Forests
535 (RFs). However, the corrected non-parametric Friedman aligned ranks test – comparing the
536 values of TSS obtained during the cross-validation – indicated no statistical difference with
537 RFs or Gradient Boosting Machines (GBMs). These results, contrast with a former
538 comparison (Xiao et al., 2017) concluding that XGBoost significantly outperforms any other
539 tree-based ensemble approach. Nevertheless, in that study none of the hyper-parameters
540 were optimised, which may have hindered RFs or GBMs to find a competent solution.

Conversely, for XGBoost a sequential scheme to determine both the best hyper-parameters and most relevant variable subset was followed in that study (Xiao et al., 2017).

The conditional RFs and oblique RFs presented lower performance and the non-parametric Friedman aligned ranks test indicated no statistical difference between them. In both cases, in addition to the variable set, only the parameter controlling the number of variables randomly sampled as candidates at each split (i.e. *mtry*) was optimised, which led to smaller searching spaces. In accordance, the parameters settings of the Genetic Algorithm (GA) should not be considered the cause of the lower performance because, compared to previous studies (e.g. Fukuda et al., 2013), with these settings the GA was able to find proficient solutions for the other three approaches with a proportionally lower searching intensity per parameter.

The reasons for such under-performance may be diverse. Conditional RFs is not intended to be particularly accurate because it was conceived to render statistically-grounded variable importance rankings, which may lead to less accurate models. In accordance, our results were to some extent expected and in line with some experiences carried out by the conceivers of conditional RFs (A. Zeileis 2018 – personal communication). On the contrary, oblique random forests were conceived to adjust better to non-axis-parallel discriminant surfaces. In accordance, it has been claimed to outperform RFs on nearly all datasets but those with discrete features (Menze et al., 2011), which can be one of the main causes of the results obtained because the cover variables were discrete. Regarding the training datasets, the highest performance was achieved, in every criterion, with oblique RFs (Table

562 3). However, tuning the parameter *mtry* and selecting the input variables proved
563 insufficient to achieve a generalization comparable with the other approaches; i.e. the TSS
564 values were lower for the validation datasets.

565 In theory, testing other multivariate models for the binary splits by modifying the
566 parameter *training_method* (*e.g.* to support vector machines – *svm*), could improve the
567 performance of the ultimate model. This option was considered interesting. However, it
568 would impede the direct calculation of the variable importance. In accordance, we
569 disregarded this option because it was considered that potential users are likely to prefer
570 functions already implemented in the software packages.

571 In terms of performance, the results obtained for the five tree-based ensemble techniques
572 were not surprising because usually those techniques with greater flexibility and allowing
573 regularization (e.g. XGBoost), and linked to a proficient parameter optimisation approach
574 are able to find better solutions. This is the case of generalized additive models compared
575 to generalized linear models (Fukuda et al., 2013) or heteroscedastic probabilistic neural
576 networks compared to the homoscedastic variant (Muñoz-Mas et al., 2018). In this regard,
577 the approaches to optimise the parameters of XGBoost are only incipient. Therefore,
578 compared with the results obtained in several competition challenges (Chen and Guestrin,
579 2016), the inexperience may have favoured discrepant results where XGBoost under-
580 performed compared to support vector machines (Fan et al., 2018). Regarding the approach
581 used to optimise the hyper-parameters, the results obtained with Bayesian optimization
582 indicated that it is a satisfactory approach, although, in terms of accuracy, either the results

583 are not comparable with ours (Xia et al., 2017) or they were statistically-similar to those
584 obtained employing GAs (Martinez-de-Pison et al., 2017).

585 Concerning the parameters of the optimisation performed with the GA (*e.g.* population size
586 – $N_{population}$), they have been related to the characteristics of the problem at hand (e.g.
587 Gibbs et al., 2011). In accordance, in the optimisation with the GA, we were tempted to
588 simply increase the searching intensity by enlarging $N_{population}$ and/or $N_{generations}$.
589 However, compared to the time spent for the standard RFs, the optimisation lapse spent for
590 XGBoost shift from minutes (RFs) to hours (XGBoost), which dissuaded us from increasing
591 these two parameters. The reasons for that increment in time were two. The formula used
592 to determine the GA-parameters led, in the case of XGBoost, to a larger population, higher
593 maximum number of iterations and higher number of iterations without change before
594 early stopping, because the number of parameters is larger in XGBoost. On the other hand,
595 the value of *nrounds* ranged between 10 and 5000, which led to a computation time
596 manifold higher than that of the standard RFs, for which only 250 trees were trained to
597 inspect the performance of every potential solution or chromosome.

598 Furthermore, previous experiences indicated that class overlapping and prevalence are
599 prominent causes of the moderate performance obtained during microhabitat suitability
600 modelling (Muñoz-Mas et al., 2016d, 2016b). The eighth variables sampled only encompass
601 part of the drivers of the microhabitat selection by fish, which we model in terms of fish
602 presence. In such a situation, over-prediction (specificity ≤ sensitivity) has been affirmed to
603 be more reliable – from an ecological viewpoint – than under-prediction (Mouton et al.,

604 2010) because it is assumed that there are not enough fish to occupy every suitable
605 microhabitat due to, for instance, low reproduction success or predation. Proper habitat
606 assessment has to evaluate a large percentage of the unoccupied microhabitats positively
607 (i.e. largely as suitable) because the reasons of the absence are not related to the quality
608 (i.e. hydraulics) of the microhabitat. This assumption may lead to a situation where the
609 eight variables do not allow better discrimination without incurring under-prediction
610 (specificity > sensitivity). There could simply be no room for improvement of an ecologically-
611 reliable data discrimination. Overall, the difference in computational burden and the usual
612 characteristics of microhabitat datasets (low prevalence and overlapping between
613 categories) led us to conclude that, in the short term, XGBoost cannot be assumed to
614 replace the other tree-based techniques in studies involving microhabitat suitability
615 modelling, particularly standard RFs or GBMs,. Still, XGBoost should stand out over
616 problems involving larger datasets (samples and variables) and/or smaller overlapping
617 between classes.

618 The comparison of the variable importance rankings did not render similar patterns.
619 Standard RFs and GBMs presented the higher importance for the two selected continuous
620 variables (i.e. velocity and depth), which is most likely reflecting the bias of these
621 approaches towards variables of this nature (Strobl et al., 2007), and this ranking resembles
622 that obtained with XGBoost. To the best of our knowledge, there are no specific studies on
623 the potential variable importance bias of XGBoost, although some users raised concerns in
624 several fora, which indicate that it should be the subject of specific research. Oblique

625 random forests rendered the opposite pattern because the variables of cover presented
626 higher importance for both species. In light of these results, and in accordance with other
627 authors (Giam and Olden, 2015), we consider that variable importance rankings obtained
628 with conditional RFs may be the most credible as statistically-grounded (Strobl et al., 2007).
629 Nevertheless, based on the discrepant results obtained here, it is indubitable that care
630 should be taken when using the variable importance ranking obtained with the other tree-
631 based ensemble approaches either to develop tree-based ensemble models using the most
632 important variables (e.g. with VSURF - Genuer et al., 2015) or to infer the most relevant
633 factors conditioning species presence.

634

635 **4.2 Ecological significance and management implications**

636 Compared to former studies performed at the mesohabitat scale, the variable importance
637 and partial dependence plots rendered new insights at the micro-scale on the habitat
638 suitability for these invasive species. These studies classified bleak as an eurytopic species
639 (Fladung et al., 2003; Muñoz-Mas et al., 2016d) whereas others, specifically addressed to
640 environmental flow assessment, considered it as a limnophilic species (Harby et al., 2007).
641 At the microhabitat scale bleak partial dependence plots highlighted the acknowledge
642 preference for open waters of lakes and medium-to-large rivers observed in its native range
643 (Amat-Trigo et al., 2019; Muñoz-Mas et al., 2016a). However, flow velocity depicted higher

644 suitability up to 0.4 m/s, which corroborates the great versatility of bleak (Latorre et al.,
645 2018).

646 Abundance has been neglected in this study in favour of presence-absence, because it tends
647 to render better accuracy (Fukuda et al., 2011) and scale the habitat suitability between
648 zero and one, which is easy to interpret, and fits well the requirements of physical habitat
649 simulation studies (Muñoz-Mas et al., 2016a). Nevertheless, the partial dependence plots
650 render hints about the species abundance in the sampled microhabitats. In the surveyed
651 river segment the deeper areas affected by artificial impoundment hosted large schools
652 with hundreds of individuals, and the models indicated larger suitability, whereas, in the
653 microhabitats sampled in run-type segments, they occurred in tens, and our models
654 indicated inferior but non-null suitability. In the latter case, bleak was observed in lower
655 number in microhabitats located in fast-flow river segments provided that structural cover
656 (i.e. rocks) was present, which is corroborated by the relevance of this type of cover
657 depicted in the partial dependence plots. Consequently, reservoirs and impounded river
658 segment can be considered the bridgehead of their invasions thought the river segments,
659 which may be assisted by their ability to stand relatively high flow velocities (Muñoz-Mas et
660 al., 2016d). In the meanwhile, reservoirs would be used as the bases for the establishment
661 and rearing of this invasive species (Almeida et al., 2014).

662 The partial dependence plots for pumpkinseed did not pose any doubt about the
663 limnophilic preferences of the species. Therefore, although its invasion success outside their
664 native range is often explained by its ecological plasticity (Ribeiro and Collares-Pereira,

665 2010; Vila-Gispert et al., 2005; Vila-Gispert et al., 2007), its habitat preferences observed in
666 our study can be considered similar to that indicated in numerous studies (e.g. Top et al.,
667 2016; Vilizzi et al., 2012). In addition to the cover provided by plants and macrophytes, in
668 those studies pumpkinseed presented the highest suitability in near-bank microhabitats
669 with low flow velocity. However, although the maximum sampled depth was not reported,
670 they indicated that pumpkinseed did not select the largest depth. Conversely, our partial
671 dependence plots indicated higher suitability above 1 m deep. We hypothesize that it could
672 be caused by an interaction between flow velocity, depth and plant auto-ecology. Low flow
673 velocity and intermediate depth can favour aquatic vegetation and reeds proliferation
674 (Strayer and Findlay, 2010). In accordance, in the river sampled in the aforementioned
675 studies, the flow velocity could have simply disfavoured the establishment of aquatic plants
676 and reeds in the deeper part of the sampled habitats causing the discrepancy with our
677 results.

678 Mediterranean flow variability have been considered to be a leading factor controlling fish
679 invasions, and the loss of its natural variability a facilitating element for their the
680 establishment (Clavero et al., 2013; Ribeiro and Collares-Pereira, 2010). There are no
681 studies on the microhabitat preferences of bleak and pumpkinseed in rivers with natural
682 flow regime, and very little about the impact of altered flow regimes in their populations
683 (e.g. Lamouroux et al., 2006). Nevertheless, our results have implications in the
684 development of alternative environmental flows addressed to counteract the presence of
685 these unwanted invasive species. Some experiences demonstrated that re-naturalizing the

686 flow regime displaces alien species in favour of that native (Kiernan et al., 2012), although,
687 in the light of our results, it may be very difficult or impossible to completely eliminate our
688 target species, particularly bleak. In accordance, we conclude that most probably once
689 these invasive species colonise a river segment, alternative water management protocols
690 could be inefficient to eliminate them.

691

692 # 5 Conclusions

693 According to the mean values obtained for the True Skill Statistic (TSS), eXtreme Gradient
694 Boosting machines (XGBoost) outperformed the other approaches, particularly conditional
695 and oblique random forests (RFs), although there were no statistical differences with
696 standard RFs and Gradient Boosting Machines (GBMs). Therefore, based on the difference
697 in the computational burden and, especially, the characteristics of the datasets on
698 microhabitat use, it has been conclude that, in the short term, XGBoost is not destined to
699 replace properly optimised RFs or GBMs for microhabitat suitability modelling.
700 Furthermore, the differences in the best hyper-parameter settings obtained for each
701 technique and species indicated that default values may be suboptimal. The variable
702 importance rankings differed significantly among techniques. In accordance, we consider
703 that the variable importance ranking obtained with conditional RFs, which is statistically-
704 grounded, may be a better option to induce parsimonious models. Nevertheless, the partial
705 dependence plots for each species were very consistent, reflecting the lacustrine origins of

706 pumpkinseed and the preference for lentic habitats of bleak. The latter species depicted a
707 larger tolerance for fast-flow microhabitats found in run-type river segments, which is likely
708 to hinder the development of counteracting environmental flow regimes. We expect that
709 the inferred habitat suitability may help ecosystem managers to develop management plans
710 addressed to impede the proliferation of these two broadly spread invasive species.

711

712 6 Acknowledgments

713 This project had the support of Fundación Biodiversidad, of Spanish Ministry for Ecological
714 Transition. We want to thank the volunteering students of the Universitat Politècnica de
715 València, Marina de Miguel, Carlos A. Puig-Mengual, Cristina Barea, Rares Hugianu, and Pau
716 Rodríguez. R. Muñoz-Mas benefitted from a postdoctoral Juan de la Cierva fellowship from
717 the Spanish Ministry of Science, Innovation and Universities (ref. FJCI-2016-30829). This
718 research was supported by the Government of Catalonia (ref. 2017 SGR 548).

719

720 7 References

721 Acreman, M., Arthington, A.H., Colloff, M.J., Couch, C., Crossman, N.D., Dyer, F., Overton, I.,
722 Pollino, C.A., Stewardson, M.J., Young, W., 2014. Environmental flows for natural,
723 hybrid, and novel riverine ecosystems in a changing world. Front. Ecol. Environ. 12,
724 466–473. doi:10.1890/130134

725 Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution
726 models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43, 1223–1232.
727 doi:10.1111/j.1365-2664.2006.01214.x

728 Almeida, D., Stefanoudis, P. V, Fletcher, D.H., Rangel, C., Da Silva, E., 2014. Population traits
729 of invasive bleak *Alburnus alburnus* between different habitats in Iberian fresh waters.
730 Limnologica 46, 70–76. doi:10.1016/j.limno.2013.12.003

731 Amat-Trigo, F., Torralva, M., Ruiz-Navarro, A., Oliva-Paterna, F.J., 2019. Colonization and
732 plasticity in population traits of the invasive bleak Alburnus alburnus along a
733 longitudinal river gradient in a Mediterranean river basin. Aquat. Invasions 14, 310–
734 331. doi:10.3391/ai.2019.14.2.10

735 Bergmann, B., Hommel, G., 1988. Improvements of General Multiple Test Procedures for
736 Redundant Systems of Hypotheses, in: Bauer, P., Hommel, G., Sonnemann, E. (Eds.),
737 Multiple Hypothesenprüfung / Multiple Hypotheses Testing. Springer, Berlin,
738 Heidelberg (Germany), pp. 100–115.

739 Bourel, M., Crisci, C., Martínez, A., 2017. Consensus methods based on machine learning
740 techniques for marine phytoplankton presence–absence prediction. Ecol. Inform. 42,
741 46–54. doi:10.1016/j.ecoinf.2017.09.004

742 Bovee, K.D., 1986. Development and evaluation of habitat suitability criteria for use in the
743 instream flow incremental methodology, Instream Flow Information Paper 21.
744 Washington, D.C. (USA).

745 Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. doi:10.1023/A:1010933404324

746 Breiman, L., 1996. Bagging predictors, in: Machine Learning. pp. 123–140.
747 doi:10.1023/A:1018054314350

748 Breiman, L., Friedman, J.H., Stone, C.J., Olshen, R.A., Stone, C.J., Olshen, R.A., 1984.
749 Classification And Regression Trees, CRC Texts in statistical science. Chapman &
750 Hall/CRC Texts in Statistical Science, New York, NY (USA).

751 Brown, G., Wyatt, J., Harris, R., Yao, X., 2005. Diversity creation methods: A survey and
752 categorisation. Inf. Fusion 6, 5–20. doi:10.1016/j.inffus.2004.04.004

753 Calvo, B., Santafé, G., 2016. scmamp: Statistical comparison of multiple algorithms in
754 multiple problems. R J. 8, 248–256.

755 Carnell, R., 2016. lhs: Latin Hypercube Samples.

756 Chawla, N. V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority
757 over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

758 Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system, in: Proceedings of
759 the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data
760 Mining - KDD '16. ACM Press, New York, New York, USA, pp. 785–794.
761 doi:10.1145/2939672.2939785

762 Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., 2017. xgboost: Extreme gradient
763 boosting.

764 Clavero, M., Hermoso, V., Aparicio, E., Godinho, F.N., 2013. Biodiversity in heavily modified
765 waterbodies: Native and introduced fish in Iberian reservoirs. Freshw. Biol. 58, 1190–
766 1201. doi:10.1111/fwb.12120

767 Clavero, M., Ninyerola, M., Hermoso, V., Filipe, A.F., Pla, M., Villero, D., Brotons, L., Delibes,
768 M., 2017. Historical citizen science to understand and predict climate-driven trout
769 decline. Proc. R. Soc. London B Biol. Sci. 284. doi:10.1098/rspb.2016.1979

770 Clusa, L., Miralles, L., Fernández, S., García-Vázquez, E., Dopico, E., 2018. Public knowledge
771 of alien species: A case study on aquatic biodiversity in North Iberian rivers. J. Nat.
772 Conserv. 42, 53–61. doi:10.1016/j.jnc.2018.01.001

773 Dietterich, T.G., 2000. Ensemble methods in machine learning, in: Multiple Classifier
774 Systems. Springer, Berlin, Heidelberg (Germany), pp. 1–15.

775 Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J.
776 Anim. Ecol. 77, 802–813. doi:10.1111/j.1365-2656.2008.01390.x

777 Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., Lu, X., Xiang, Y., 2018. Comparison of
778 Support Vector Machine and Extreme Gradient Boosting for predicting daily global
779 solar radiation using temperature and precipitation in humid subtropical climates: A
780 case study in China. Energy Convers. Manag. 164, 102–111.
781 doi:10.1016/j.enconman.2018.02.087

782 Fladung, E., Scholten, M., Thiel, R., 2003. Modelling the habitat preferences of preadult and

783 adult fishes on the shoreline of the large, lowland Elbe River. J. Appl. Ichthyol. 19, 303–
784 314.

785 Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and
786 an application to boosting. J. Comput. Syst. Sci. 55, 119–139.

787 Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378.
788 doi:10.1016/S0167-9473(01)00065-2

789 Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann.
790 Stat. 29, 1189–1232. doi:10.1214/aos/1013203451

791 Friedman, M., 1940. A comparison of alternative tests of significance for the problem of *m*
792 rankings. Ann. Math. Stat. 11, 86–92.

793 Fukuda, S., De Baets, B., 2016. Data prevalence matters when assessing species' responses
794 using data-driven species distribution models. Ecol. Inform. 32, 69–78.
795 doi:10.1016/j.ecoinf.2016.01.005

796 Fukuda, S., De Baets, B., Waegeman, W., Verwaeren, J., Mouton, A.M., 2013. Habitat
797 prediction and knowledge extraction for spawning European grayling (*Thymallus*
798 *thymallus* L.) using a broad range of species distribution models. Environ. Model. Softw.
799 47, 1–6. doi:10.1016/j.envsoft.2013.04.005

800 Fukuda, S., Mouton, A.M., De Baets, B., 2011. Abundance versus presence/absence data for
801 modelling fish habitat preference with a genetic Takagi-Sugeno fuzzy system. Environ.

802 Monit. Assess. 184, 6159–6171. doi:10.1007/s10661-011-2410-2

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

807 García, S., Fernández, A., Luengo, J., Herrera, F., 2010. Advanced nonparametric tests for
808 multiple comparisons in the design of experiments in computational intelligence and
809 data mining: Experimental analysis of power. Inf. Sci. (Ny). 180, 2044–2064.
810 doi:10.1016/j.ins.2009.12.010

811 García, S., Herrera, F., 2008. An extension on "statistical comparisons of classifiers over
812 multiple data sets" for all pairwise comparisons. J. Mach. Learn. Res. 9, 2677–2694.

813 Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2015. VSURF: An R Package for Variable Selection
814 Using Random Forests. R J. 7, 19–33.

815 Giam, X., Olden, J.D., 2015. A new R^2-based metric to shed greater insight on variable
816 importance in artificial neural networks. Ecol. Modell. 313, 307–313.
817 doi:10.1016/j.ecolmodel.2015.06.034

818 Gibbs, M.S., Maier, H.R., Dandy, G.C., 2011. Relationship between problem characteristics
819 and the optimal number of genetic algorithm generations. Eng. Optim. 43, 349–376.
820 doi:10.1080/0305215X.2010.491547

821 Gobeyn, S., Volk, M., Dominguez-Granda, L., Goethals, P.L.M., 2017. Input variable selection
822 with a simple genetic algorithm for conceptual species distribution models: A case
823 study of river pollution in Ecuador. Environ. Model. Softw. 92, 269–316.
824 doi:10.1016/j.envsoft.2017.02.012

825 Gómez-Ríos, A., Luengo, J., Herrera, F., 2017. A study on the noise label influence in
826 boosting algorithms: AdaBoost, GBM and XGBoost, in: Martínez de Pisón, F.J., Urraca,
827 R., Quintián, H., Corchado, E. (Eds.), Hybrid Artificial Intelligent Systems - HAIS 2017.
828 Springer International Publishing, Cham (Switzerland), pp. 268–280. doi:10.1007/978-3-
829 319-59650-1_23

830 Greenwell, B., Boehmke, B., Cunningham, J., Developers, G.B.M., 2019. gbm: Generalized
831 Boosted Regression Models.

832 Harby, A., Olivier, J.M., Merigoux, S., Malet, E., 2007. A mesohabitat method used to assess
833 minimum flow changes and impacts on the invertebrate and fish fauna in the Rhône
834 River, France. River Res. Appl. 23, 525–543. doi:10.1002/rra.997

835 Holland, J.H., 1992. Genetic algorithms. Sci. Am. 267, 66–72.

836 Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A., Van Der Laan, M.J., 2005. Survival
837 ensembles. Biostatistics 7, 355–373. doi:10.1093/biostatistics/kxj011

838 Hothorn, T., Hornik, K., Strobl, C., Zeileis, A., Hothorn, M.T., 2015. Package 'party.' Packag.
839 Ref. Man. Party Version 0.9-998 16, 37.

840 Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: A conditional
841 inference framework. J. Comput. Graph. Stat. 15, 651–674.
842 doi:10.1198/106186006X133933

843 Hothorn, T., Hornik, K., Zeileis, A., Strobl, C., Zeileis, A., 2010. Party: A laboratory for
844 recursive partytioning, R package version 0.9-0, URL http://CRAN. R-project. org.

845 Ilhéu, M., Matono, P., Bernardo, J.M., 2014. Invasibility of mediterranean-climate rivers by
846 non-native fish: The importance of environmental drivers and human pressures. PLoS
847 One 9. doi:10.1371/journal.pone.0109694

848 Kiernan, J.D., Moyle, P.B., Crain, P.K., 2012. Restoring native fish assemblages to a regulated
849 California stream using the natural flow regime concept. Ecol. Appl. 22, 1472–1482.
850 doi:10.1890/11-0480.1

851 Knowles, J., Hughes, E.J., 2005. Multiobjective optimization on a budget of 250 evaluations,
852 in: Coello Coello, C.A., Hernández Aguirre, A., Zitzler, E. (Eds.), Evolutionary Multi-
853 Criterion Optimization. Springer, Berlin, Heidelberg (Germany), pp. 176–190.

854 Kottelat, M., Freyhof, J., 2007. Handbook of European freshwater fishes. Kottelat & Freyhof
855 Publishing, Cornol (Switzerland) & Berlin (Germany).

856 Lamouroux, N., Olivier, J.-M., Capra, H., Zylberblat, M., Chandesris, A., Roger, P., 2006. Fish
857 community changes after minimum flow increase: testing quantitative predictions in
858 the Rhone River at Pierre-Benite, France. Freshw. Biol. 51, 1730–1743.

doi:10.1111/j.1365-2427.2006.01602.x

Latorre, D., Masó, G., Hinckley, A., Verdiell-Cubedo, D., Tarkan, A.S., Vila-Gispert, A., Copp, G.H., Cucherousset, J., Silva, E. da, Fernández-Delgado, C., García-Berthou, E., Miranda, R., Oliva-Paterna, F.J., Ruiz-Navarro, A., Serrano, J.M., Almeida, D., 2018. Inter-population variability in growth and reproduction of invasive bleak *Alburnus alburnus* (Linnaeus, 1758) across the Iberian Peninsula. Mar. Freshw. Res. 69, 1–7. doi:10.1071/MF17092

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

López, V., Fernández, A., García, S., Palade, V., Herrera, F., 2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Inf. Sci. (Ny). 250, 113–141. doi:10.1016/J.INS.2013.07.007

Marmion, M., Hjort, J., Thuiller, W., Luoto, M., 2008. A comparison of predictive methods in modelling the distribution of periglacial landforms in Finnish Lapland. Earth Surf. Process. Landforms 33, 2241–2254. doi:10.1002/esp.1695

Martinez-de-Pison, F.J., Gonzalez-Sendino, R., Aldama, A., Ferreiro, J., Fraile, E., 2017. Hybrid methodology based on bayesian optimization and GA-PARSIMONY for searching parsimony models by combining hyperparameter optimization and feature selection, in: Martínez de Pisón, F.J., Urraca, R., Quintián, H., Corchado, E. (Eds.), Hybrid Artificial Intelligent Systems - HAIS 2017. Springer International Publishing, Cham (Switzerland), pp. 52–62. doi:10.1007/978-3-319-59650-1_5

879 Masó, G., Latorre, D., Tarkan, A.S., Vila-Gispert, A., Almeida, D., 2016. Inter-population
880 plasticity in growth and reproduction of invasive bleak, *Alburnus alburnus* (Cyprinidae,
881 Actinopterygii), in northeastern Iberian Peninsula. Folia Zool. 65, 10–14.
882 doi:10.25225/fozo.v65.i1.a3.2016

883 Mebane Jr, W.R., Sekhon, J.S., 2011. Genetic optimization using derivatives: The rgenoud
884 package for R. J. Stat. Softw. 42, 1–26.

885 Menze, B., Splitthoff, N., Splitthoff, M.D.N., 2012. obliqueRF: Oblique random forests from
886 recursive linear model splits.

887 Menze, B.H., Kelm, B.M., Splitthoff, D.N., Koethe, U., Hamprecht, F.A., 2011. On oblique
888 random forests, in: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (Eds.),
889 European Conference on Machine Learning and Principles and Practice of Knowledge
890 Discovery in Databases, ECML PKDD 2011. Springer, Athens (Greece), pp. 453–469.
891 doi:10.1007/978-3-642-23783-6_29

892 Mouton, A.M., Alcaraz-Hernández, J.D., De Baets, B., Goethals, P.L.M., Martínez-Capel, F.,
893 2011. Data-driven fuzzy habitat suitability models for brown trout in Spanish
894 Mediterranean rivers. Environ. Model. Softw. 26, 615–622.
895 doi:10.1016/j.envsoft.2010.12.001

896 Mouton, A.M., De Baets, B., Goethals, P.L.M., 2010. Ecological relevance of performance
897 criteria for species distribution models. Ecol. Modell. 221, 1995–2002.
898 doi:10.1016/j.ecolmodel.2010.04.017

899 Muñoz-Mas, R., Fukuda, S., Pórtoles, J., Martínez-Capel, F., 2018. Revisiting probabilistic
900 neural networks: a comparative study with support vector machines and the
901 microhabitat suitability for the Eastern Iberian chub (*Squalius valentinus*). Ecol. Inform.
902 43, 24–37. doi:10.1016/J.ECOINF.2017.10.008

903 Muñoz-Mas, R., Fukuda, S., Vezza, P., Martínez-Capel, F., 2016a. Comparing four methods
904 for decision-tree induction: A case study on the invasive Iberian gudgeon (*Gobio*
905 *lozanoi*; Doadrio and Madeira, 2004). Ecol. Inform. 34, 22–34.
906 doi:10.1016/j.ecoinf.2016.04.011

907 Muñoz-Mas, R., Garófano-Gómez, V., Andrés-Doménech, I., Corenblit, D., Egger, G., Francés,
908 F., Ferreira, M.. T., García-Arias, A., Politti, E., Rivaes, R., Rodríguez-González, P.M.,
909 Steiger, J., Vallés-Morán, F.. J., Martínez-Capel, F., 2017. Exploring the key drivers of
910 riparian woodland successional pathways across three European river reaches.
911 Ecohydrology 10, e1888–e1888. doi:10.1002/eco.1888

912 Muñoz-Mas, R., Lopez-Nicolas, A., Martínez-Capel, F., Pulido-Velazquez, M., 2016b. Shifts in
913 the suitable habitat available for brown trout (*Salmo trutta* L.) under short-term
914 climate change scenarios. Sci. Total Environ. 544, 686–700.
915 doi:10.1016/j.scitotenv.2015.11.147

916 Muñoz-Mas, R., Martínez-Capel, F., Schneider, M., Mouton, A.M., 2012. Assessment of
917 brown trout habitat suitability in the Jucar River Basin (Spain): Comparison of data-
918 driven approaches with fuzzy-logic models and univariate suitability curves. Sci. Total

919 Environ. 440, 123–131. doi:10.1016/j.scitotenv.2012.07.074

920 Muñoz-Mas, R., Papadaki, C., Martínez-Capel, F., Zogaris, S., Ntoanidis, L., Dimitriou, E.,
921 2016c. Generalized additive and fuzzy models in environmental flow assessment: A
922 comparison employing the West Balkan trout (*Salmo farioides*; Karaman, 1938). Ecol.
923 Eng. 91, 365–377. doi:10.1016/j.ecoleng.2016.03.009

924 Muñoz-Mas, R., Vezza, P., Alcaraz-Hernández, J.D., Martínez-Capel, F., 2016d. Risk of
925 invasion predicted with support vector machines: A case study on northern pike (*Esox*
926 *Lucius*, L.) and bleak (*Alburnus alburnus*, L.). Ecol. Modell. 342, 123–134.
927 doi:10.1016/j.ecolmodel.2016.10.006

928 Nguyen, H.T., Everaert, G., Boets, P., Forio, A.M., Bennetsen, E., Volk, M., Hoang, H.T.,
929 Goethals, L.P., 2018. Modelling tools to analyze and assess the ecological impact of
930 hydropower dams. Water. doi:10.3390/w10030259

931 Piotrowski, A.P., 2017. Review of Differential Evolution population size. Swarm Evol.
932 Comput. 32, 1–24. doi:10.1016/j.swevo.2016.05.003

933 Poff, N.L., Allan, J.D., Bain, M.B., Karr, J.R., Prestegaard, K.L., Richter, B.D., Sparks, R.E.,
934 Stromberg, J.C., 1997. The natural flow regime: A paradigm for river conservation and
935 restoration. Bioscience 47, 769–784. doi:10.2307/1313099

936 Ren, Y., Zhang, L., Suganthan, P.N., 2016. Ensemble classification and regression-recent
937 developments, applications and future directions. IEEE Comput. Intell. Mag. 11, 41–53.

938 doi:10.1109/MCI.2015.2471235

939 Ribeiro, F., Collares-Pereira, M.J., 2010. Life-history variability of non-native centrarchids in
940 regulated river systems of the lower River Guadiana drainage (south-west Iberian
941 Peninsula). J. Fish Biol. 76, 522–537. doi:10.1111/j.1095-8649.2009.02506.x

942 Ridgeway, G., 2007. Generalized boosted models: A guide to the gbm package. Compute 1,
943 1–12.

944 Somodi, I., Lepesi, N., Botta-Dukát, Z., 2017. Prevalence dependence in model goodness
945 measures with special emphasis on true skill statistics. Ecol. Evol. 7, 863–872.
946 doi:10.1002/ece3.2654

947 Stoffels, R.J., Bond, N.R., Nicol, S., 2018. Science to support the management of riverine
948 flows. Freshw. Biol. n/a-n/a. doi:10.1111/fwb.13061

949 Strasser, H., Weber, C., 1999. On the asymptotic theory of permutation statistics. Math.
950 Methods Stat. 8, 220–250.

951 Strayer, D.L., Findlay, S.E.G., 2010. Ecology of freshwater shore zones. Aquat. Sci. 72.
952 doi:10.1007/s00027-010-0128-9

953 Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable
954 importance measures: illustrations, sources and a solution. BMC Bioinformatics 8, 25.
955 doi:10.1186/1471-2105-8-25

956 Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable

importance for random forests. BMC Bioinformatics 9. doi:10.1186/1471-2105-9-307

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14, 323–348. doi:10.1037/a0016973

Thomsen, M., Wernberg, T., Olden, J., Byers, J.E., Bruno, J., Silliman, B., Schiel, D., 2014. Forty years of experiments on aquatic invasive species: are study biases limiting our understanding of impacts? NeoBiota 22, 1–22. doi:10.3897/neobiota.22.6224

Thuiller, W., Lafourcade, B., Engler, R., Araújo, M.B., 2009. BIOMOD – a platform for ensemble forecasting of species distributions. Ecography (Cop.). 32, 369–373. doi:10.1111/j.1600-0587.2008.05742.x

Top, N., Tarkan, A.S., Vilizzi, L., Karakuş, U., 2016. Microhabitat interactions of non-native pumpkinseed Lepomis gibbosus in a Mediterranean-type stream suggest no evidence for impact on endemic fishes. Knowl. Manag. Aquat. Ecosyst. doi:10.1051/kmae/2016023

Torgo, L., 2010. Data Mining with R, learning with case studies. Chapman and Hall/CRC.

Vila-Gispert, A., Alcaraz, C., García-Berthou, E., 2005. Life-history traits of invasive fish in small Mediterranean streams. Biol. Invasions 7, 107–116. doi:10.1007/s10530-004-9640-y

Vila-Gispert, A., Fox, M.G., Zamora, L., Moreno-Amich, R., 2007. Morphological variation in

976 pumpkinseed *Lepomis gibbosus* introduced into Iberian lakes and reservoirs;
977 adaptations to habitat type and diet? J. Fish Biol. 71, 163–181. doi:10.1111/j.1095-
978 8649.2007.01483.x

979 Vilizzi, L., Stakenas, S., Copp, G.H., 2012. Use of constrained additive and quadratic
980 ordination in fish habitat studies: An application to introduced pumpkinseed (*Lepomis*
981 *gibbosus*) and native brown trout (*Salmo trutta*) in an English stream. Fundam. Appl.
982 Limnol. 180, 69–75. doi:10.1127/1863-9135/2012/0277

983 Waters, B.F., 1976. A methodology for evaluating the effects of different streamflows on
984 salmonid habitat, in: Proceedings of the Symposium and Specialty Conference on
985 Instream Flow Needs. American Fisheries Society, Bethesda, MD (USA), p. 13.

986 Xia, Y., Liu, C., Li, Y., Liu, N., 2017. A boosted decision tree approach using Bayesian hyper-
987 parameter optimization for credit scoring. Expert Syst. Appl. 78, 225–241.
988 doi:10.1016/j.eswa.2017.02.017

989 Xiao, Z., Wang, Y., Fu, K., Wu, F., 2017. Identifying different transportation modes from
990 trajectory data using tree-based ensemble classifiers. ISPRS Int. J. Geo-Information 6.
991 doi:10.3390/ijgi6020057

992