A comparison of modeling techniques to predict juvenile 0+ fish species occurrences in a large river system

J. Leclere a,⁎, T. Oberdorff a, J. Belliard b, F. Leprieur c,⁎

a UMR CNRS 5178-IRD 207-MNHN-UPMC « Biologie des Organismes et Écosystèmes Aquatiques », DMPA, Muséum National d’Histoire Naturelle, 43 rue Cuvier 75005 Paris, France
b UMR « Hydrostéomatiques et bioprocédures », Cemagref, 1 rue Pierre-Gilles de Gennes, CS 10030, F-92 761 Antony, France
c UMR 5119 ECOSYM « Ecologie des Systèmes Marins Côtières » (CNRS-IRD-IFREMER-UM1-UM2), Université Montpellier 2, Place Eugène Bataillon 34095 Montpellier Cedex 5, France

A R T I C L E   I N F O

Article history:
Received 1 February 2011
Received in revised form 13 May 2011
Accepted 17 May 2011
Available online xxxx

Keywords:
Machine learning
Young-of-the-year fishes (YOY)
Predictive models
Habitat variables
Large rivers
France
Restoration tools

A B S T R A C T

Even if European river management and restoration are largely supported by the use of reliable tools, these tools are most often “generalist” and provide only initial leads of alteration sources. Acknowledging that young-of-the-year (YOY) fish assemblages are highly dependent on riverine habitat conditions, the development of a YOY-based tool might be very useful or even essential in the design and implementation of conservation or restoration plan of large rivers, in measuring more straightforward the losses and gains of hydro-ecological functionalities. In the past 20 years, new modeling techniques have emerged from a growing sophistication of statistical model applied to ecology. “Machine learning methods” (ML) are now recognized as holding great promise for the advancement of understanding and prediction of ecological phenomena. The aim of this work was to select the appropriate statistical technique to model YOY assemblages according to different meso-scale habitat variables that are meaningful to planners. To do this, two “Machine Learning” methods, Classification and Regression Trees (CART) and Boosted Regression Trees (BRT), were compared to Generalized Linear Models (GLM). We modeled the occurrence of 9 species from the Seine River basin (France) in order to compare models abilities to accurately predict the presence and absence of each species. BRT appeared to be the best technique for modeling 0+ fish occurrences in our dataset.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

In Europe, river management and restoration are largely supported by the use of a panel of reliable bioindicators in order to (1) assess ecosystems condition, (2) identify the origin of observed alterations and (3) give valuable information on possible restoration measures (see the Water Framework directive WFD, 2000). At this time, a variety of biological tools from many assemblages (i.e., diatoms, macrophytes, benthic invertebrate or fish) have been developed to respond to these stakes. However they are most often “generalist” tools assessing the ecological status of rivers over large spatial scales and providing only initial leads of alteration sources. Consequently, there are still high expectations from decision makers and water managers, to have local enforcement tools measuring more straightforward the losses and gains of hydro-ecological functionalities. Using young-of-the-year (YOY) fish assemblages could be interesting in this context. Indeed, YOY fish assemblages are highly dependent on riverine habitat conditions (Juradja et al., 2010). For example, the success of fish reproduction and recruitment in bank habitats are strongly influenced by factors such as climate, river flow, temperature, water quality, habitat structure, and food availability (Copp et al., 1991). Since most European YOY fish occupy similar habitats to those where adult spawned (Copp, 1992b), YOY assemblages are potentially good descriptors for assessing and predicting the features of a river in terms of reproduction and recruitment functionalities (Copp et al., 1991). Consequently, the development of a YOY-based tool might be thus very useful or even essential in the design and implementation of conservation or restoration plan of large rivers (Copp, 1992b).

Even if YOY fishes have received considerable attention since early 70s, owing to the recognition of their important contribution to the success of the species as a whole (Garner, 1996), investigations of this particular stage are still much scarcer than for adults. YOY species richness and/or abundance in riverine systems have been linked to physical habitat (Langler and Smith, 2001; Stakenas, 2002), water chemical conditions (Copp, 2003; Sandstrom and Karas, 2002; Stakenas, 1999), hydroclimatic variables (Cattaneo, 2005; Cattaneo et al., 2001; Grenouillet and Pont, 2001; Juradja et al., 2004; Nunn et al., 2003; Tales, 2008) and biotic interactions (Bischoff and Freyhof, 1999; Grenouillet et al., 2000; Grenouillet et al., 2002; Grenouillet and Pont, 2001; Stakenas 1999). Other studies have focused, using multivariate analyses, on developing empirical models for describing (Baras et al., 1995; Copp, 1989, 1992a; Copp et al., 1991; Juradja, 1995; Valova et al., 2006) and predicting the microhabitat of YOY fish
species (Baras et al., 1995; Copp, 1992b; Gozlan et al., 1998; Grifò et al., 2003; Juradja, 1999; Keckeis et al., 1997; Nicolas and Pont, 1995; Penaz et al., 1995).

In the past 20 years, new modeling techniques have emerged from a growing sophistication in the types of statistical model applied to ecology, with impetus from substantial advances in both statistics and computing (Elith et al., 2008; Lek and Guégan, 2000; Olden et al., 2008). For example, “machine learning methods” (ML) are now recognized as holding great promises for the advancement of understanding and prediction of ecological phenomena. This is mainly due to their ability to model complex, non-linear relationships in ecological data without having to satisfy the restrictive assumptions required by conventional, parametric approaches like Generalized Linear Models (Olden et al., 2008). Numerous recent papers confirm this interest for modeling fish–habitat relationships and fish distribution (Cappo et al., 2005; Elith et al., 2008; Knudby et al., 2010; Leathwick et al., 2006; Olden and Jackson, 2001; Olden and Jackson, 2002; Pittman et al., 2009), whereas these methods are still rarely used for YOY fish data (but see Brosse et al., 1999; Berrebi Dit Thomas, 1999).

This study was conducted to select an appropriate statistical technique to predict 0+ fish species occurrences in large river systems. Specifically, we compared two “Machine Learning” methods, i.e. Classification and Regression Trees (CART) and Boosted Regression Trees (BRT), with Generalized Linear Models (GLM), which can be considered as a conventional statistical method against which other models can be compared. To do this, we modeled the occurrence of 9 common European riverine species according to different meso-scale habitat variables. The main goal of mesohabitat models is to provide information at a scale that is both meaningful to environmental managers and decision makers, and relevant to model fish–habitat relationships (Eisner et al., 2005). We first compared model abilities to accurately predict the presence and absence of each 0+ fish species. We then discussed and compared, for each modeling approach, the relative influence of the selected variables according to their nature: quantitative, semi-quantitative and qualitative. We finally looked at the ecological soundness of environmental variables selected by each model. This study is a first step for the development of predictive models assessing the ecological integrity of large rivers through reproductive and recruitment success.

2. Materials and methods

2.1. Data sets

The data set used in this study consisted of 130 habitats spread over 18 sites, all located in two large rivers of the Seine River basin: the Seine river and the Marne river (Fig. 1). These sites were selected for their good water physicochemical quality assessed from 11 parameters during years 2004 to 2007 (i.e. temperature, pH, dissolved oxygen, oxygen saturation, biological oxygen demand, ammonium, nitrite, nitrate, phosphate, suspended solids and chlorophyll a) (data available on request or on AESN website: www.eau-seine-normandie.fr). Localized on potentially navigable waterways (≥ 25 m width), the whole sample area is generally regarded as sub-natural, because no strong intervention was conducted on the banks except for some localized boulders ripp-rapp.

0+ juvenile fishes were collected from boat by electrofishing using a Point Abundance Sampling strategy (Nelva et al., 1979) modified for young fishes (Persat and Copp, 1989). Sampling plan was based on a partial and stratified random sampling within the main habitats characterizing the site (e.g. beaches with different kind of substratum, macrophytes beds, and wooded banks with submerged roots...). Each site, measuring 1000 m on average, was explored through a minimum of five habitats; each of them was subject to 10 sampling points. When possible, YOYS have been identified in the field and immediately released. Unidentified fishes were fixed in formaldehyde 5% for later determination at the laboratory using Spillmann (1961) and Koblickaja (1981) determination keys. Data were mostly collected from mid July to mid August 2009. The summer period was chosen because during this period (1) most individuals have reach the juvenile stage, facilitating their determination, and (2) YOY populations are quite stable (Copp, Penaz et al., 1995).

Please cite this article as: Leclere, J., et al., A comparison of modeling techniques to predict juvenile 0+ fish species occurrences in a large river system, Ecological Informatics (2011), doi:10.1016/j.ecoinf.2011.05.001

Fig. 1. (A) Location of the Seine river basin in France and (B) location of the sample sites in the Marne and Seine rivers.
For each model, a cross-validation procedure was employed. The global dataset was randomly divided into 2 subsets: a training set (100 habitats) and a validation set (30 habitats) (Van Houwelingen and Le Cessie, 1990). The size of the two subsets was arbitrarily chosen to allow a sufficient number of habitats for models calibration and to provide a satisfactory validation of the models. The mean depth and the mean width were log-transformed to approach normality. All calculations were performed under the statistical software “R” and its contributed packages (R Core and Development Team, 2010).

**Generalized Linear Model (GLM)**

GLMs are commonly used for modeling species presence/absence because of their strong statistical foundation and ability to realistically model ecological relationships (Austin, 2002). For these reasons, this statistical method appears to be a good benchmark to assess quality of the two other modeling techniques. GLM involves a logit link and binomial error distribution to model the presence–absence of each species. Models were fitted with a maximum likelihood method (McCullagh and Nelder, 1989). Selection of variables was performed using a backward stepwise technique and Akaike Information Criterion (AIC). Only variables with both a significant regression slope and a significant contribution to the explained deviance were kept in the final models. GLM were developed under the MASS package (Venables and Ripley, 2011).

**Classification and Regression Trees (CART)**

CARTs are non-parametric classification techniques that are implemented using a binary recursive partitioning algorithm (Breiman et al., 1984). Tree building is based on an algorithm that repeatedly partitions the data set into a nested series of mutually groups, each of them as homogeneous as possible with respect to the response variable. The procedure begins with the entire data set, also called roots node, and formulates split-defining conditions for each possible value of the

---

**Table 1**

0+ fish species and occurrences found in the reference stretches of the Marne and Seine rivers, with their reproductive, ecological and trophic guilds.

<table>
<thead>
<tr>
<th>Code</th>
<th>Common name</th>
<th>Scientific name</th>
<th>Reproductive guild (1)</th>
<th>Ecological guild (2)</th>
<th>Trophic guild (3)</th>
<th>Prevalence (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aa</td>
<td>Bleak</td>
<td>Alburnus alburnus</td>
<td>Phytolithophil</td>
<td>Eurytop</td>
<td>Zooplanktophageous</td>
<td>60.8</td>
</tr>
<tr>
<td>Bb</td>
<td>Barbel</td>
<td>Barbus barbus</td>
<td>Lithophil</td>
<td>Rheophil</td>
<td>Polyplageous</td>
<td>86.9</td>
</tr>
<tr>
<td>Bj</td>
<td>Silher bream</td>
<td>Blicca Bjoerka</td>
<td>Phytophil</td>
<td>Eurytop</td>
<td>Zooplanktophageous</td>
<td>10.0</td>
</tr>
<tr>
<td>Cn</td>
<td>Nase</td>
<td>Chondrostoma nasus</td>
<td>Lithophil</td>
<td>Rheophil</td>
<td>Periplageous</td>
<td>58.5</td>
</tr>
<tr>
<td>Gg</td>
<td>Gudgeon</td>
<td>Gobio gobio</td>
<td>Psammophil</td>
<td>Rhoophil</td>
<td>Zoobenthophageous</td>
<td>56.9</td>
</tr>
<tr>
<td>Li</td>
<td>Dace</td>
<td>Leucticus leuciscus</td>
<td>Phytolithophil</td>
<td>Rhoophil</td>
<td>Zoobenthophageous</td>
<td>26.9</td>
</tr>
<tr>
<td>Rr</td>
<td>Roach</td>
<td>Rutilus rutilus</td>
<td>Phytolithophil</td>
<td>Eurytop</td>
<td>Zooplanktophageous</td>
<td>58.5</td>
</tr>
<tr>
<td>Rs</td>
<td>Bitterling</td>
<td>Rhoeus sericus</td>
<td>Ostracophil</td>
<td>Limnophil</td>
<td>Phytoplageous</td>
<td>16.2</td>
</tr>
</tbody>
</table>

**Table 2**

Environmental variables (measured from map or during field sampling) used to predict species occurrences in the 130 habitats.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Code</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shallow waters (when depth is &lt;50 cm at 1 m from the bank)</td>
<td>Hfd</td>
<td>A (absent)</td>
</tr>
<tr>
<td>Shelters (Anfractuity, Roots and Helophytes)</td>
<td>ANF, ROR, HEL</td>
<td>A (absent)</td>
</tr>
<tr>
<td>Semi-quantitative variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bottom substrat (Boulders, Stones, Pebbles, Gravels, Sand, Silt, Clay)</td>
<td>Bld, Str, Pbl, Grav, Sd, Silt, Clay</td>
<td>A (absent)</td>
</tr>
<tr>
<td>Current velocity</td>
<td>Vc</td>
<td>A (absent)</td>
</tr>
<tr>
<td>Shade</td>
<td>Shade</td>
<td>A (absent)</td>
</tr>
<tr>
<td>Overall macrophytes complexity structures</td>
<td>MacP</td>
<td>1 (absent)</td>
</tr>
<tr>
<td>Quantitative variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depth at the sampling point (cm) Log transformed</td>
<td>LogD</td>
<td>Mean and range</td>
</tr>
<tr>
<td>Stream width (m) Log transformed</td>
<td>LogW</td>
<td>1.710 (1.301–2.204)</td>
</tr>
<tr>
<td>Substrat diversity (Shannon index)</td>
<td>Shannon</td>
<td>1.630 (1.380–1.689)</td>
</tr>
<tr>
<td>Weighted average of proportional cover of bed sediment</td>
<td>LocsED</td>
<td>0.052 (0.177)</td>
</tr>
</tbody>
</table>

---

Please cite this article as: Leclere, J., et al., A comparison of modeling techniques to predict juvenile 0+ fish species occurrences in a large river system, Ecological Informatics (2011), doi:10.1016/j.ecoinf.2011.05.001
explanatory variables to create candidate splits (Olden and Jackson, 2002). The algorithm selects the candidate splits that minimize the average “impurity” of the two child nodes. Evaluation of “impurity” is done through the Gini index. The algorithm continues recursively with each of the new children nodes until tree building is stopped (Olden et al., 2008). We determined tree size by trial and error.

CARTs were developed with rpart package developed by Therneau and Atkinson (2010).

Boosted Regression Trees (BRT)

BRT is a relatively new method combining statistical and machine learning techniques by using two algorithms: decision trees from CART group of models (see above), and boosting technique that builds and combines a collection of models (Elith et al., 2008). Utilization of boosting technique permits to combine large numbers of relatively simple tree models to optimize predictive performance (Elith et al., 2006). This method is based on the idea that it is easier to find and average many rough rules of thumb than to find a single, highly accurate prediction rule (Schapire, 2003). This recent technique is still sparsely used in ecology despite its apparent high predictive potential (Møsen et al., 2006).

The approach used in this study develops multiple classification trees by iteratively fitting new trees to the residual errors of the existing tree assemblage (De'ath, 2007). It can be understood as a forward stage wise method in which the fit of the model to data is progressively improved through the addition of individual trees. These additions are based on the gradient of log-likelihood, which is progressively increased (Leathwick et al., 2006). Existing trees are not changed through iterations, and the final model is a linear combination of all the tree assemblages (Elith et al., 2008).

BRT were developed with gbm package (Ridgeway, 2006) extension developed by Elith et al., 2008. We used a BRT model assuming a binomial error distribution. Models fitting was carried out with the gbm.step function. We also simplified the model to reduce the number of explanatory variables with the gbm.simplify function. Hyper-parameters of the boosting algorithm, number of trees and the shrinkage variable are tuned automatically with internal cross-fitting was carried out with the rpart function. We also simplified the model to reduce the number of explanatory variables with the gbm.simplify function. Hyper-parameters of the boosting algorithm, number of trees and the shrinkage variable are tuned automatically with internal cross-validation while learning rate, tree complexity and bag fraction were determined by trial and error.

2.3. Predictive models comparison

Assessment of predictive performances

We quantified the accuracy of the predictive models for each of the species using 6 measures (Allouche et al., 2006; Manel et al. 2001): the kappa index, the true statistical skills (TSS), the sensitivity (SN), the specificity (SP), the overall prediction success (CC) and the area under the receiving operation characteristic curve (AUC) (see details in Table 3). Performances were assessed with independent data under the receiving operation characteristic curve (AUC) (see details in Table 3). Performances were assessed with independent data under the receiving operation characteristic curve (AUC) (see details in Table 3). Performances were assessed with independent data under the receiving operation characteristic curve (AUC) (see details in Table 3). Performances were assessed with independent data under the receiving operation characteristic curve (AUC) (see details in Table 3).

<table>
<thead>
<tr>
<th>Species</th>
<th>GLM</th>
<th>CART</th>
<th>BRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aa</td>
<td>0.856</td>
<td>0.605</td>
<td>0.662</td>
</tr>
<tr>
<td>Bb</td>
<td>0.966</td>
<td>0.671</td>
<td>0.571</td>
</tr>
<tr>
<td>Bj</td>
<td>0.684</td>
<td>0.232</td>
<td>0.223</td>
</tr>
<tr>
<td>Cn</td>
<td>0.952</td>
<td>0.714</td>
<td>0.694</td>
</tr>
<tr>
<td>Gg</td>
<td>0.794</td>
<td>0.605</td>
<td>0.639</td>
</tr>
<tr>
<td>Li</td>
<td>0.843</td>
<td>0.596</td>
<td>0.683</td>
</tr>
<tr>
<td>Pf</td>
<td>0.172</td>
<td>0.122</td>
<td>0.321</td>
</tr>
<tr>
<td>Rr</td>
<td>0.754</td>
<td>0.508</td>
<td>0.517</td>
</tr>
<tr>
<td>Rs</td>
<td>0.506</td>
<td>0.014</td>
<td>0.019</td>
</tr>
</tbody>
</table>

That might have occurred by chance alone, comparing the number of correct forecast, minus those attributable to random guessing, to that of a hypothetical set of perfect forecasts. It has already been demonstrated that TSS is not prevalence-dependant (Allouche et al., 2006).

AUC quantifies the ability of a model to discriminate between sites where a species is present versus those where it is absent (Hanley and Mc Neil, 1982). It was shown to be independent of prevalence, and to be a threshold-independent measure for model performance. It can be interpreted as indicating the probability that, when a presence or an absence are drawn at random from the population, the first one will have a higher predicted value than the second one (Elith et al., 2006).

Variables contribution in predictive models

Variables importance assessment of each model was set up by using the relative explain deviance contribution.

For GLM, a measure of relative influence was calculated for each of the predictor terms in the model to facilitate comparisons of term-wise contributions in deviance reduction. Although this measure may be biased because the predictors are not completely orthogonal (Araújo and Guisan, 2006), we believe that deviance remains in our case an appropriate measure of the contribution of variables in this model.

To calculate the relative importance of each predictor, CART quantifies the improvement measure attributable to each variable in its role as a surrogate to the primary split. The values of these improvements are summed over each node and scaled relative to the best performing variable, i.e., expressed as a relative importance (Breiman et al., 1984).

For BRT, the relative influence of each environmental variable in the models was determined using scripts from the gbm library. Estimation of the relative influence are based on the number of times a variable is selected for splitting, weighted by the squared improvement as a result of each split, and averaged over all trees (Friedman and Meulman, 2003). The relative contribution of each variable is scaled so that the sum adds to 100 (Elith et al., 2008).

Spearman rank-correlation coefficient between relative importance contributions has been used to compare patterns in the
importance of the variables for predicting the occurrence of each species (Olden and Jackson, 2002).

3. Results

3.1. Predictive performance

For each of the six measures of predictive performance, predictability of species occurrences was not concordant among the nine species (Table 3). According to Hanley and McNeil (1982) AUC interpretation, GLM had “worse accuracy than random” for one species (AUC < 0.5) and a “good” accuracy for four species (AUC > 0.8). CART had also “worse accuracy than random” for one species and a “good” accuracy for three species. BRT models obtained the best results with AUC ranging from 0.500 to 0.925, no accuracy was “worse than random”, and accuracies were “good” for 5 species. In accordance with the Kappa index, and following Fielding and Bell (1997), accuracy of the GLM and CART models was “poor” for three species (K < 0.4) and “good” for six species (0.4 < K < 0.75) whereas BRT models was “poor” for only one species, “good” for seven species and “excellent” for one species (K > 0.75). True Statistical Skills (TSS) confirmed these previous findings. Accuracy of the GLM models was “poor” for three species and “good” for six species. CART models were a somewhat worse with “poor” accuracy for four species and “good” performances for five species. BRT models were “poor” for only two species, “good” for six species and “excellent” for one species. Ratios of correct classification still confirm that BRT performs better. Best correct classification rate were obtained for seven species versus four species with GLM and two with CART. It is worth noting that for some species, the best accuracy was found in different models. Specificity and sensitivity reacted in the same way. BRT models were best performing for six species, CART for three species and GLM for two species.

3.2. Variable importance

Correlation analysis of ranked variables revealed similarities between the 3 modeling approaches in the importance of variables used to predict the occurrence of species. Significant pair-wise similarities between approaches highlight which methods converge in their ranked importance of habitat variables for each species. Models exhibited marked similarities in the importance of the habitat for predicting occurrence of silver bream (Blicca Bjoerkna) and nase (Chondrostoma nasus), whereas they exhibited strong differences for roach. Of the 15 significant pair-wise similarities, 5 were found for each pairs of modeling techniques (i.e. 5 between CART and BRT, 5 between GLM and CART methods, and 5 between GLM and BRT methods) (Fig. 2).

Based on all species, GLM and CART had the closest agreement in the mean ranked importance of variables. Inversely, GLM and BRT exhibit the strongest differences. These last differences are especially marked for quantitative variables (Fig. 3).

Analyzing separately the mean rank importance of the 3 kinds of variables (qualitative, semi-quantitative and quantitative) highlighted marked differences depending on the models. The 3 kinds of variable are fairly distributed in GLM models, while importance of qualitative variables is reduced compared to quantitative variables in CART and BRT models. This trend comes with a reduction in importance of semi-quantitative variables in BRT (Fig. 3).

On the general trend, the number of selected variables in GLM was consistently lower. Compared with other models, BRT models included generally more predictive variables. The number of predictive variables entering the models is ranging from 1 in GLM to 10 for BRT depending on the species. The selected variables for each model are summarized in Table 4.
Table 4

<table>
<thead>
<tr>
<th>Species</th>
<th>Qualitative</th>
<th>Semi-quantitative</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aa</td>
<td>HFd</td>
<td>ANFROT</td>
<td></td>
</tr>
<tr>
<td>Bb</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bj</td>
<td>HEL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cn</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gg</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ll</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pf</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rr</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rs</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Selected variables (X) for Generalized linear models (GLM), regression trees (CART) and boosted regression trees (BRT). Species codes are defined in Table 1 and variable codes are defined in Table 2.
3.3. Species–habitat relationships

The application of different techniques generally results in models differing in their response, as they are based on different algorithms. Some predictor variables were systematically selected by all techniques for a same species, highlighting their interest in modelling 0+ fish occurrences. The most revealing example is LogD, which was selected by the three techniques for barbel (Barbus barbus), nase (C. nasus) and gudgeon (Gobio gobio) models, with a general increased of their probability of occurrence with decreasing depth. Similarly, MacP was selected by the three techniques for bleak (Alburnus alburnus) and silver bream (B. Bjoerkna) models, leading to an improvement in the probability of occurrence with increasing complexity of the structure of macrophytes. It is worth noting that Shade and LogW were also selected by the three modeling techniques, respectively for nase (C. nasus) and dace (Leuciscus leuciscus) (Table 4). The mean rank importance of variables, based on relative contributions across the 3 modeling approaches, confirmed that these predictors are of primary importance for predicting YOYs presence/absence (Fig. 3).

To compare the ecological rationality of the different predictive models, we selected response of species for which accuracy was “good” (Aa, Cn, Gg) and for which the most important variables (i.e. MacP, Shade, and LogD) were selected together in the 3 models.

For the MacP variable, trends in responses of bleak (A. alburnus) were studied (Fig. 4). GLM model indicates that a more complex macrophyte structure improves the probability of occurrence of this species. With a two step function, BRT also indicates that a more complex macrophyte structure improves the probability of occurrence with almost the same probabilities for the two most complex structures (classes 3 and 4). CART model exhibits a different trend. For structural complexity upper than 2, bleak (A. alburnus) is always predicted as present (without distinction between the three other classes), whereas, for structural complexity lesser than 2 (which corresponds to an absence of macrophyte), its predicted presence depends on several other variables (Shade, LocSED, Vc and LogW).

For the Shade variable, we selected the nase (C. nasus) for examining the models response (Fig. 5). The general response is similar for the 3 models, showing a general negative trend. BRT and CART models discriminate total absence of shadow (class A) from the three other classes, while GLM model indicates a more gradual relation with a lowering of the probability of occurrence proportional to the intensity of shading, represented by a negative exponential function.

For the last variable, LogD, responses of gudgeon (G. gobio) was studied (Fig. 6). CART model only indicates a strong probability of occurrence below a depth of around 65 cm. BRT model describes a first plateau from 0 to 20 cm and then a second plateau between 40 and 60 cm where the probability of occurrence is constant. Then, BRT describes a regular decrease of probability up to 80 cm. After this depth the probability of occurrence stays uniform. Instead, GLM model shows a regular decrease of the probability of occurrence with increasing depth.

4. Discussion

4.1. Predictive performance

Overall, no single modeling technique was optimal for modelling occurrence of all species. However, results suggest that BRT is the most accurate modeling method for several reasons. First, BRT...
produced the greatest number of best models under each of the six metrics. Second, BRT was the only approach that gave no prediction “worse than random” with AUC. Third, considering Kappa index and TSS, BRT technique produced the smallest number of “poor” models, and was the only approach that gave “excellent” prediction. Finally, best overall correct classification, sensitivity and specificity were also obtained more often with BRT. Indeed, BRT are able to select relevant variables, fit accurate functions and automatically identify and model interactions. This gives a potential predictive advantage since large suites of candidate variables is handled better than in GLM developed under stepwise selection (Elith et al., 2008). As explained by Olden and Jackson (2002), the ability of non-linear approach to perform as well as linear methods when the data show linear relationship, and to capture and model better results when the data relationships are non-linear provides another advantage. A primary point given that non-linearity is considered as common in individual fish–habitat relationships (Pittman et al., 2009). Some studies confirm these results, Moisen et al. (2006) found that boosted trees outperformed individual regression trees for predictions of tree basal area. More recently, Aersten et al. (2010) found that boosted trees outperformed both individual regression trees and linear models to model site index of homogenous stands of three important tree species, while Knudby et al. (2010) found that for predictions of species richness and diversity of reef fishes the tree-based models (with BRT) are generally superior.

Despite the superior performance of BRT models, this modeling technique, like the two others, was unable to accurately predict the occurrence of the bitterling (Rodeus sericeus). We explain these poor predictive performances by the low prevalence of this species in the dataset (16.2%) and the relatively small sample size. This phenomenon was already described in modeling literature as influencing the reliability of the predictions (Jiménez-Valverde et al., 2009). Secondly, it could be due to its belonging to a particular ecological guild (see Table 1). The fact that bitterling is an ostracophil species means that it needs bivalves for the accomplishment of its reproduction. This particularity makes that its occurrence difficult to model because this host presence–absence was not integrated in our predictive variables.

4.2. Variables importance

Concerning the ranking of habitat variables by the models, we found only weak convergence. This indicates that the different approaches established different quantitative relationships between the habitat variables and species occurrence, even when these approaches correctly predicted occurrence. This result is consistent with those obtained by Olden and Jackson (2002). The majority of the observed differences were found between BRT and GLM. This finding is not surprising given that these two models display inherent differences in their statistical foundations (see Materials and methods). Differences in mean ranked importance of the variables between GLM and BRT could be partially explained by the nature of the variables (i.e. qualitative, semi-qualitative and quantitative). In GLM, mean ranked importance of the three kinds of variables were almost equal whereas in tree-based models, importance of continuous variables was greatly superior, especially in BRT. In the same way, the comparison between response variables and model types reveals marked differences in the number of variables that each model identifies as important. The GLM identifies the fewest number of important variables whereas CART and BRT models identify a greater number of variables. This greater number of variables in tree-based techniques passed either through the substitution of categorical variables by continuous ones, or directly by the addition of semi-quantitative and continuous variables. Continuous variables were preferred in tree-based models, both quantitatively and qualitatively, because these allow a large number of possible splits in the regression trees, as opposed to few possible splits for quantitative or semi-quantitative variables. Overall, this large number of split makes easier the building of non-linear relationships. Variable importance is thus dependent on the set of variables included in the data set, their nature (i.e. qualitative, semi-qualitative and quantitative) and the type of model (Knudby et al., 2010).

4.3. Species–habitat relationships

Whether it is by their mean rank importance across species and models, or by the number of times that the variables are selected, some predictive variables already appear as important in predicting 0+ fishes species occurrence. These “key” variables, MacP, Shade and LogD, are commonly highlighted in 0+ literature to explain and predict density or species presence/absence. Nevertheless, when prediction success appears to be satisfactory, it doesn’t necessary mean that shape is ecologically relevant. This is why analyzing ecological rationality is needed to choose the appropriate modeling technique in a specific situation. MacP variable has been considered important because of its simultaneous selection in the three techniques for two species, but also because of its mean rank importance. Its importance for 0+ fishes is not surprising, since macrophytes are known to provide both shelters against predation and a significant food resource (Dionne and Folt, 1991). Considering bleak 0+ (A. alburnus), Juradja (1999) found a significant effect of emergent vegetation (helophytes) and Copp (1992a) shows a general preference of bleak 0+ for macrophytes (Copp, 1989). Shade variable, has been considered as important because of its average rank importance and because it was also selected simultaneously by the

Fig. 6. Relationships between LogD variable and gudgeon (Gobio gobio) occurrence predicted from (A) GLM, with points in bold showing the probability of occurrence of the species for each values of the predictor variable and normal points inside of the graph showing distribution of samples along that variable, (B) CART, with for nodes, the splits value for each predictor variable and for leaves, the mean probability of occurrence and the number of habitat concerned and (C) BRT modeling techniques, with partial dependence plot of the predictor variable for predicting the occurrence of gudgeon. Y axis is on the logit scale. Rugplot at inside top of the graph shows distribution of samples along that variable.
three techniques for the nase (C. nasus) model. In all models, reduction of the luminosity leads to a decrease in occurrences probabilities. These results are consistent with those of Copp (1992a, 1993), indicating a preference for habitat directly illuminated for larvae or juveniles 0+. LogD variable obtained the highest mean rank importance and was simultaneously retained by the three models for three species and in more than half of all species models. This variable represents the shoreline profile, since it corresponds precisely to the depth of the river embankment. Considering bank depth and gudgeon 0+ (C. gobio) relationship, some studies highlighted a particular attraction of this species for shallow waters and especially for bank depth (Copp, 1992a, 1992b; Gozlan et al., 1998). These results confirm that there is no major inconsistency between the literature and our findings, even if it is important to note that literature is not always concordant. Finally, only CART models, by the nature of their responses, seems somewhat simplistic compared to the two other techniques in modelling species habitat relationships.

5. Conclusion
Three modeling techniques were compared and evaluated for predicting 0+ fish occurrence for nine common species in the Seine River basin. Based on six measures of accuracy, BRT appeared to be the best technique for modeling 0+ fish occurrences in our dataset. BRT selected a greater number of variables, and proportionally more continuous variables than the two other techniques. Continuous variables, that provide more precise information than categorical ones have been preferred by BRT probably because better fitting non-linear relationships. According to the comparison between literature and the response curves of the different modeling techniques, BRT also provides better ecological interpretability and rationality in their response curves.

Acknowledgements
This work was supported by the “National Association for Research and Technology” (ANRT), the “Hydrosphere Society” (CIFRE contract no. 383/2008), and “Voies Navigables de France” (VNF). We thank Mr. Souben J. for his very helpful participation during the 2009 fishing session and all Hydrosphere Society’s employees, especially Mr. Michel P. (Hydrosphere director), the project designer.

References
Eissner, A., Young, C., Scheiner, M., Kopecki, I., 2005. MesoCASIMIR – new mapping method and comparison with other current approaches. COST 626, European Aquatic Modelling, Network, 05-74.


Please cite this article as: Leclere, J., et al., A comparison of modeling techniques to predict juvenile 0+ fish species occurrences in a large river system, Ecological Informatics (2011), doi:10.1016/j.ecoinf.2011.05.001