ISEIS

Journal of Environmental Informatics 32(2) 112-124 (2018)

Journal of Environmental Informatics

www.iseis.org/jei

Tree-Based Methods: Concepts, Uses and Limitations under the Framework of Resource Selection Models

J. Carvalho[1,2,*], J. P. V. Santos[1,3], R. T. Torres[1], F. Santarém[4], and C. Fonseca[1]

[1] *Department of Biology, Centre for Environmental and Marine Studies, University of Aveiro, Aveiro 3810-193, Portugal*
[2] *Servei d'Ecopatologia de Fauna Salvatge, Departament de Medicina i Cirurgia Animals, Universitat Autònoma de Barcelona, Bellaterra 08193, Spain*
[3] *Sanidad y Biotecnología, Instituto de Investigación en Recursos Cinegéticos, Consejo Superior de Investigaciones Científicas – Universidad de Castilla-La Mancha - Junta de Comunidades de Castilla-La Macha, Ciudad Real 13071, Spain*
[4] *Research Centre in Biodiversity and Genetic Resources, University of Porto, Vairão 4485-661, Portugal*

Received 9 August 2015; revised 18 April 2016; accepted 3 May 2016; published online 4 May 2017

ABSTRACT. The use of empirical models to predict species distribution is recognized as an important tool in wildlife management. Tree-based methods gained considerable attention in the last years mostly due to their flexibility and robustness. Here, we provide an overview of tree-based methods by addressing some of their concepts, uses and limitations. For illustrative purposes, we modelled the distribution of a red deer (*Cervus elaphus*) population using fine-scale predictors while applying four modelling methods: three tree-based methods (classification trees, random forests and boosted trees) and the generalized linear model by stepwise regression. In order to explore alternative trees and achieve the best model performance, a series of classifiers were run with different tuning parameters. The random forests and boosted trees models were the most accurate classifiers followed by classification trees and generalized linear model by stepwise regression. Despite differences in the predictive accuracy, the results of the four models were consistent with the species ecological requirements. Red deer occurred further away from disturbed areas (e.g. villages and other human settlements), agricultural fields and near shrubs and forest patches. Furthermore, the species often occurred in areas with gentle slopes, preferentially with a southern exposure. We observed that classification trees are easy to interpret but may produce unstable decision trees and unwieldy results in the presence of sharp discontinuities. We state that ensemble methods such as random forests and boosted trees are valuable tools in predicting species distributions. This study provides the necessary background for the understanding of tree-based methods, which will be of great help in further studies in ecological modelling, as it will shed light in the most appropriate technique to be used.

Keywords: boosted trees, classification trees, ecological modelling, fine-scale predictors, random forests, red deer

1. Introduction

Modelling environmental scenarios became a key tool in distinct research fields in order to help manage real-world problems. Over the last years, the use of species distribution models (SDMs, hereafter) has been increasing due to the advances in computing capacity, the increased number of bioinformatics, the accessibility to species occurrences and the availability of environmental data (see Guisan and Zimmermann, 2000; Guisan and Thuiller, 2005; Elith et al., 2006; Araújo and New, 2007; Elith and Leathwick, 2009 for reviews). This prominent tool has been successfully used to address a wide variety of ecological issues such as the management of threatened species and biological invasions, the prediction of species distribution under current and future environmental scenarios, as well as the determination of phylogeographic patterns (Guillera-Arroita et al., 2015). Amongst the classes of SDMs, the habitat or resource selection models (RSMs) have been applied to predict the species occurrence and habitat selection at finer-scales (Hegel et al., 2010). Essentially, both classes of models describe the interactions between species distribution and environmental predictors. The different types of species data (*e.g.* presence-only, presence-absence, presence-pseudo-absence or presence-background, abundance and opportunistic records), the characteristics of environmental predictors (*e.g.* high dimensionality, multicollinearity) and the relationships between the response and explanatory variables (*e.g.* non-linear relationships, heteroscedasticity) pose different methodological and statistical challenges. Currently, a plethora of modelling techniques is available. Even though regression-like techniques remain as the backbone of modelling approaches, tree-based methods, a branch of machine learning, constitute an emerging set of quantitative tools which have been fairly used (Araújo et al., 2011; Engler et al., 2011; Broennimann et al., 2012; Ewijk et al., 2014). These include classification and regression trees (CART, Brei-

[*] Corresponding author. Tel: +351 234370350; fax: +351 234372587.
E-mail address: jlcarvalho@ua.pt (J. Carvalho).

ISSN: 1726-2135 print/1684-8799 online
doi: 10.3808/jei.201600352

man et al., 1984; Clark and Pregibon, 1992), random forests (RF, Breiman, 2001) or boosted trees (BT, Freund and Schapire, 1996). Their main advantage lays on its high predictive power and the flexibility to handle interactions and nonlinearities (De'ath and Fabricius, 2000; Prasad et al., 2006; De'ath, 2007). Furthermore, they are non-parametric, they normally do not require previous variable selection and are able to deal with missing values, outliers and unbalanced data (Vayssiéres et al., 2000). Due to their characteristics, tree-based methods are suitable for a range of ecological applications (Debeljak and Džeroski, 2011). Despite all these advantages, tree-based methods also exhibit some drawbacks, since they are time consuming and their high flexibility can lead to a major pitfall called overfitting. Overfitting is an undesirable model situation which means that flexible models fit the noise rather than the general data behavior (James et al., 2013). Overfitting mainly occurs in complex models with high variance and low bias and could lead to poor model performance on predicting new data (Warren et al., 2014).

Several studies were performed to assess the performance of different modelling methods (Thuiller et al., 2003; Segurado and Araújo, 2004; Tsoar et al., 2007; García-Callejas and Araújo, 2015). Although these studies may guide methodological and statistical choices, the results showed that models predictive performance is highly variable, making it difficult to select a statistical method by first-time users. Despite several reviews and comparative studies under the framework of SDMs, comprehensive interpretations are still scarce (see Elith et al., 2010 and Merow et al., 2013). The aim of this work is to offer an introductory and non-exhaustive description of the key tree-based concepts using an illustrative case study. First, we provide a literature review focused on the conceptual and methodological aspects of tree-based methods. Then, based on a dataset of a red deer (*Cervus elaphus*) population we assess (i) the current species distribution considering three different tree-based methods: classification trees, random forests, and boosted trees; (ii) the relative importance of the environmental variables; and (iii) the performance of the above mentioned techniques in comparison with a widely used regression-like method, the generalized linear model by stepwise regression (GLM, McCullagh and Nelder, 1989). Considering the results of previous studies, we hypothesized that the predictive performance of ensemble methods (RF and BT) were equivalent (Hypothesis 1) and outperformed single classification trees and generalized linear models (Hypothesis 2). Our study provides further insights for the understanding of tree-based methods, which will be of great help in further studies in ecological modelling, as it will shed light in the most appropriate technique to be used.

2. Tree-Based Methods – An Overview

Tree-based methods are a set of supervised approaches (Table 1), which are successfully applied in different research fields. Their popularity lies on their flexibility to handle multifaceted data. The methodological basis involves the segmentation of a predictor space into a particular number of simple subsets (Hastie et al., 2009).

In the last years, the development of powerful graphical

Table 1. Terms Used in the Context of Tree-based Methods

Concept	Definition
Bagging	Also known as bootstrap aggregation, is the most simple and one of the most common ways to generate an ensemble of weak learners.
Bag fraction	Proportion of the training dataset randomly selected for model fitting.
Boosting	Sequential ensemble process in which a highly accurate predictor is created through the combination of several weak and inaccurate classifiers.
Branch	Path taken by individual records until the next node. Branch width reflects the proportion of instances that follows a determined path.
Child node	Subset of observations resulting from a parent node split.
Complexity parameter	Define the tree size through its control over the pruning procedure. Pivotal parameter to reduce model overfitting.
Ensemble	Set of classifiers which, presumably, produce more stable and accurate results than a single model. Particularly useful when the classifiers exhibit an erratic and unstable behavior.
Error rate	The probability that a model incorrectly classifies an instance. Measures the effectiveness of a classifier.
Learning rate	Also known as the shrinkage parameter, determines to what extent the addition of a new tree contributes to improve the final model, *i.e.* controls the impact of subsequent fitted learners in the final model. Higher predictive performances are often associated to smaller values of shrinkage (Natekin and Knoll, 2013).
Leaves	Terminal node representing class labels. It is not partitioned if the node reached the minimum number of observations defined, a threshold of splits was achieved or the observations of a child node are homogeneous, *i.e.* belongs to the same class.
Loss function	Commonly used to weight differently the type of errors once some loss functions are more robust to noisy and unbalanced data (see Hastie et al., 2009).

Out-of-bag error	The subset of training observations used to grow a decision tree is called "bag". The instances left out called as "out-of-bag" are used to estimate the error rate.
Overfitting	Commonly occurs when a model is excessively complex. Overfitting could lead to poor predictive performance on new and unseen datasets (Warren et al., 2014). Overfitting can be avoided through the use of several techniques such as the complexity parameter, cross-validation, and regularization, to name a few.
Recursive partitioning	Iterative top-down process where nodes are split sequentially in order to increase the homogeneity of child nodes with respect to the response variable.
Root	The first node in the traditional structure of a decision tree. Encompasses all the instances that form the dataset.
Split	Data partition in two datasets based on a particular question. The proper choice of the splitting criteria is necessary to increase the homogeneity of outputs and information gain.
Supervised learning	Machine learning branch that entails learning the relationship between known predictors and response variable, seeking predictive models to forecast the response to unseen data.
Tree complexity	Number of nodes in a tree. Tuning parameter of boosted trees that, together with the learning rate, rules the number of trees added to the ensemble.

user interfaces (GUIs), routines and packages for statistical software made these techniques accessible to the majority of ecologists. Commercial software companies such as Salford Systems (www.salford-systems.com) provide the Windows-based programs CART® and RandomForests®. DTREG (www.dtreg.com) offers a broad set of predictive modelling methods including Decision Trees®, TreeBoost® and Decision Tree Forests®. Routines and packages for *R* statistical software include 'BIOMOD' (Thuiller, 2003), 'ModelMap' (Freeman, 2009) and 'Rattle' library (Williams, 2009). Other options encompass the R packages 'rpart' (Therneau et al., 2013) to fit classification and regression tree models, 'random Forest' (Liaw and Wiener, 2012) to develop random forests models and 'gbm' (Ridgeway, 2013) for boosting regression trees. These recent advances led to a growing use of tree-based methods in ecological modelling.

2.1. Classification and Regression Trees – The Classic Tree Algorithm

Classification (categorical variable) and regression (numerical variable) trees (CART) consist in the binary recursive partition (*i.e.* successive segmentation) of the data into simpler and more homogeneous subsets (Breiman et al., 1984; De'ath and Fabricius, 2000; Vayssières et al., 2000; Hastie et al., 2009; James et al., 2013). A traditional structure of a decision tree includes a single root node composed by all instances or cases, which is then split into two branches resulting in two child nodes (Figure 1). The resulting variance from the data partition is as homogeneous as possible considering the dependent variable.

The CART models have two main challenges: i) as CART considers all possible splits regarding all variables, it is defiant to find good splits and ii) to avoid data overfitting.

The determination of information gain and/or node impurity measures (entropy, Gini index of diversity or misclassification error) allows overcoming the former challenge (Breiman et al., 1984; Hastie et al., 2009; Therneau et al., 2015). For illustrative purposes, we provide an example of splitting criteria for binary data using the information gain (Figure S1). The information gain is an entropy-based concept whose values range between 0 (no entropy; target variable only comprises observations with the same values, which means that no further information is required to classify the observations) and 1 (maximum entropy; corresponds to higher amounts of disorder, meaning that the values of the target variable are equally distributed across the records).

Nonetheless, pruning the tree reduces model overfitting. A pruned tree allows a simple and systematic representation of the data, increasing the accuracy of predictions of unobserved data. The pruning process can be performed by specifying the number of instances per terminal node, by using the minimum split argument, or a more refined method, the complexity parameter (*cp*) (Williams, 2011; Therneau et al., 2015).

The development of a CART model is performed in four main steps: i) partition of the training dataset; ii) fitting a model to the data considering previous data partitions; iii) stop when the residuals of the model are approximately zero or the number of remaining observations is low; and iv) pruning the tree to avoid overfitting.

2.2. Random Forests – More than A Bagging Approach

The high variance of environmental data leads to unwieldy classification and decision trees (Hastie et al., 2009). To overcome this shortcoming, several samples from the training dataset can be taken through a procedure referred as bootstrap aggregation. Also known as bagging, this averaging model works by reducing the variance through the construction of *n* decision trees using *n* bootstrapped training sets and then averaging a set of predictions (James et al., 2013).

The RF is a modified bagging-based algorithm where

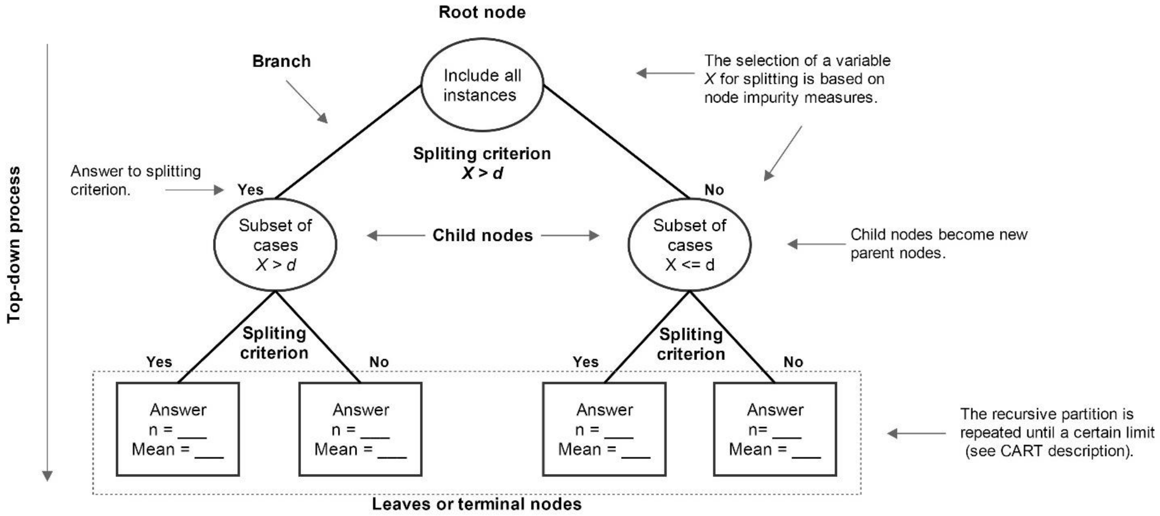
Root node
Include all instances
Branch
The selection of a variable X for splitting is based on node impurity measures.
Spliting criterion X > d
Answer to splitting criterion.
Yes
No
Subset of cases X > d
Child nodes
Subset of cases X <= d
Child nodes become new parent nodes.
Top-down process
Spliting criterion
Yes
No
Spliting criterion
Yes
No
Answer n = ___ Mean = ___
Answer n = ___ Mean = ___
Answer n = ___ Mean = ___
Answer n= ___ Mean = ___
The recursive partition is repeated until a certain limit (see CART description).
Leaves or terminal nodes

Figure 1. The structure of classification or regression trees (adapted from Olden et al., 2008).

each tree is independently constructed using several bootstrap samples and the new trees are independent from the previous ones (Breiman, 2001). The RF algorithm fits many decision trees to a dataset, and then combines the predictions from all the trees (Figure S2). This algorithm begins with the selection of several bootstrap samples from the initial dataset, *i.e.* each classifier has access to a different data subset. A decision tree is fitted to each bootstrap sample, but at each node only a smaller number of randomly selected variables (*e.g.* the square root of the number of variables) are available for the binary partitioning. This clear degree of randomness constitutes a safeguard against overfitting. The final model is built by aggregating a set of predictions from the individual trees (Cutler et al., 2007; Hastie et al., 2009). The RF algorithm retains the variables that provide more information in the discrimination of item classes (Evans et al., 2011). The RF is able to measure the contribution of each predictor, even when its effects are covered by multicollinearity issues (Strobl et al., 2009).

The method's performance depends on two fundamental parameters: i) the overall number of trees (*nt*), which should increase as the number of variables/instances increases, and ii) the number of variables considered for each split, whose value restriction ensures that correlation among fitted trees is small.

2.3. Boosted Trees – A Collection of Weak Learners

The BT models are built using an algorithm that combines decision trees and boosting (Friedman et al., 2000; De'ath, 2007; Elith et al., 2008). AdaBoost was the original boosting algorithm (Freund and Schapire, 1996). Here, sequences of models are assembled and successive trees change the observation weights giving lower or higher relevance to those cases correctly or incorrectly classified, respectively. Each observation has an initial weight calculated as $w_i = 1/n$, where w_i is the weight of each observation and n is the number of observations (Figure S3). This process starts with the fit of the first classifier of the weighted data. A detailed statistical description of AdaBoost rationale is presented in Friedman et al. (2000) and Hastie et al. (2009).

Friedman (2001) developed a new approach called gradient boosting. The performance of such method is improved and the overfitting is reduced by the introduction of randomness and by stochastic gradient boosting, where each decision tree is constructed by taking a random subsample of the training dataset (Friedman, 2002). The aim of the gradient BT is to improve the model performance by combining a large number of simple trees, *i.e.* the final outcome is a collection of weak learners. The model fit over different trees is improved by considering the previous learners and by emphasizing those observations incorrectly classified. For regression issues, each new tree added to the ensemble is fitted to the residuals of the previous tree (Elith et al., 2008).

The BT procedure is optimized by two main parameters: i) the learning rate (*lc*), also known as the shrinkage parameter (λ), which determines how quickly the algorithm adapts to a training dataset and ii) the tree complexity (*tc*), which controls the tree size, the ensemble complexity and whether an interaction is fit. The *lc* and the *tc* control the number of trees added to the ensemble (Elith et al., 2008).

3. Methods

3.1. Case Study

Presence data of red deer (n = 539 observations) were systematically collected in *Lombada* (east part of Montesinho Natural Park - Portugal) and *Sierra de la Culebra* (Spain) hunting areas (LSCHA), in Northwestern Iberian Peninsula (41º43′ ~ 42º03′N; 6º43′ ~ 6º27′W, Figure 2). The LSCHA covers an area of 48,740 ha and is characterized by a hetero-

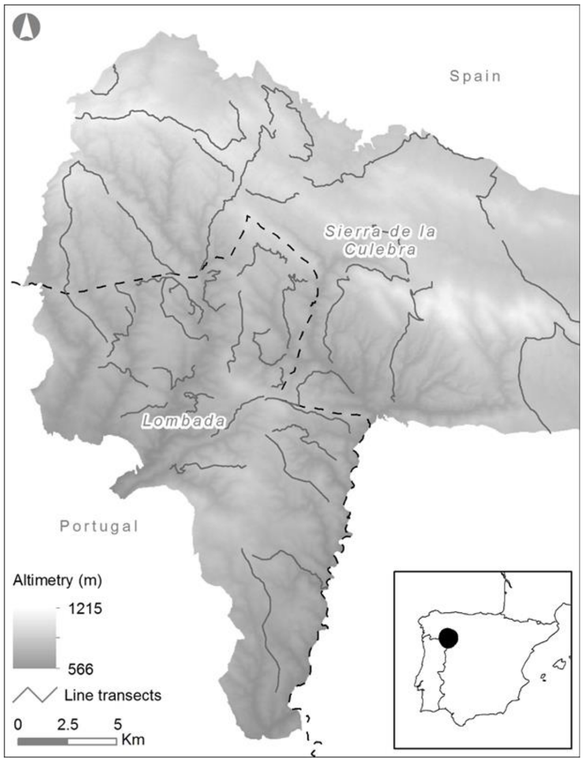

Figure 2. Detail of the study area showing the location of the line transects covered during red deer population surveys.

geneous orography with elevation ranges from 566 to 1215 m.a.s.l. The LSCHA experiences a Mediterranean climate with well-marked seasons (Kottek et al., 2006). The vegetation is varied and mainly dominated by scrublands and forest stands interspersed by semi-natural pastures and meadows. Scattered and small-cultivated fields can also be found along LSCHA.

The study area comprises a rich biodiversity. Red deer populations experienced a remarkable recovery during the last decades and are now common and well-established in several regions of the Iberian Peninsula, namely in the LSCHA. The rural exodus and occasional reintroduction projects are some of the main causes underlying the increase of numbers and distribution range of red deer populations (Vingada et al., 2010). In LSCHA, red deer is considered a game species and is also one of the main preys of the endangered Iberian wolf (*Canis lupus signatus*). Given the high ecological and socio-economic relevance of red deer, and their recent expansion, and by taking advantage of precise data on red deer monitored in the LSCHA, we defined a methodological approach to describe the environmental determinants of red deer presence at a fine geographical scale.

3.2. Presence Data and Environmental Predictors

Line transects surveys were conducted during the rut period (September ~ October 2012). Transects were chosen to provide an equal coverage of the most representative habitats in the study area and, thus, reduce the bias associated with the systematic prospection of areas with different deer densities. Forty-eight transects with an average length of 4.6 km (0.43 ~ 12.7 km) were surveyed (Figure 2). Whenever an animal/group was detected, the distance from the observation point to the animal/group was recorded. Through GPS location and trigonometric operations, the exact position of the observed animals was determined. For modelling purposes, the 539 records of red deer presence were divided into training (70%) and test (30%) datasets. Presence data was used to establish the favourable conditions for the species occurrence, while background data (*i.e.* random set of points within the study area, Phillips et al., 2009; Phillips and Elith, 2011) characterized the environmental domain where the survey was carried out. The ratio of presences/background data was set at 1:1, as recommended by Barbet-Massin et al. (2012) for classification techniques. The LSCHA was divided into a hexagonal grid composed by individual units (side length = 250 m; area = 16.24 ha) that retained the environmental characteristics of the corresponding section at a finer-scale. Hexagonal units were considered more suitable for a range of ecological applications than the commonly used rectangular grids as they provided a better representation of the spatial heterogeneity (Clausnitzer et al., 2009). Then, considering the red deer ecological requirements, their applicability to the study area and potential predictive significance, 14 fine-scale variables (Table 2) associated with habitat structure (3), human disturbance (2), land use (3), vegetation productivity (1), topography (4) and water availability (1) were selected.

3.3. Geographical Background Delimitation

A variation in the geographical background (GB) extent leads to differences in the discriminatory power of SDMs (Acevedo et al., 2012). GB can underestimate the role of course-scale factors (*e.g.* climatic drivers), if it is too restrict (Sánchez-Fernández et al., 2011). However, if it is too large, GB can limit the model's predictive power when the aim is to determine the influence of fine-scale conditions in geographic patterns of species distribution (Lobo et al., 2010). Here, we adopted the criterion proposed by Acevedo et al. (2012) to delimitate the geographical background. By applying a trend surface analysis, we improved the model performance and decreased the extent effect on the final outcome. Trend surface analysis fits a polynomial surface by least-squares regression of geographical coordinates. This method is used to find general data trends and modulation of curvilinear structures is performed through the addition of polynomial terms to the explanatory data (Legendre and Legendre, 1998). Therefore, trend surfaces are normally formulated as n^{th} polynomials, creating gradually varying surfaces that describe the physical or geographical processes. We fitted several surfaces by increasing the polynomial order and, consequently, their complexity. The root mean square error of interpolation was used to determine the best value to use for the polynomial order.

Table 2. Variables Used to Develop the Models of Red Deer Occurrence*

Factor	Variable	Code	Average (Min ~ Max)
Habitat structure	Agricultural area (ha)	AAgr	2.96 (0.00 ~ 16.24)
	Forest area (ha)	AFor	4.02 (0.00 ~ 16.24)
	Shrub area (ha)	ASch	7.16 (0.00 ~ 16.24)
Human disturbance	Distance to road network (m)	DRoad	950 (15 ~ 4785)
	Distance to villages (m)	DUrb	1754 (0 ~ 5078)
Land use	Distance to agricultural fields (m)	DAgr	465 (0 ~ 3257)
	Distance to forest (m)	DFor	357 (0 ~ 2650)
	Distance to shrubs (m)	DSch	136 (0 ~ 1511)
Vegetation productivity	Normalized Difference Vegetation Index	NDVI	0.30 (-0.15 ~ 0.51)
Topography	Slope (degrees)	Slp	10 (0 ~ 27)
	Northness	NNESS	-0.08 (-0.92 ~ 0.96)
	Eastness	ENESS	-0.01 (-0.90 ~ 0.93)
	Terrain roughness	RNESS	155 (4 ~ 326)
Water availability	Distance to main water lines (m)	DRiv	1327 (55 ~ 5035)

*The average and range values were obtained from individual hexagonal units.

3.4. Performance Evaluation

We used the area under (AUC) the receiver operating characteristic curve (ROC) to measure the discrimination power. The AUC provides a way to compare classifiers by testing the model accuracy and allows their validation independent of distortions and potential bias (Fielding and Bell, 1997). The output values range between 0.5, *i.e.* the scores of the two groups do not differ, and 1, *i.e.* the scores of the two groups do not overlap. A set of threshold-dependent measures based on the outputs of the confusion matrix was also used to assess the model accuracy. These measures included the overall accuracy (OA), the sensitivity (Se, true positive rate), the specificity (Sp, true negative rate), the true skill statistics (TSS) and the Cohen's kappa. Considering that spatial autocorrelation between "training" and "testing" datasets may inflate the AUC values (Veloz, 2009), we used a truly independent dataset from another red deer population located in *Lousã* mountain, centre of Portugal, to assess the model discriminative performance. Furthermore, we evaluated the model's accordance considering the selection of the most important variables for species occurrence. For this purpose, Spearman's rank tests were used to assess the correlation between variables importance, ranked according to their order of selection by each model.

3.5. Model Specifications and Software

Distinct tuning parameters were tested to assess their effects on model performance. In CT, a series of 10-fold cross-validations was run and the most frequent occurring tree size was chosen using the standard error (1-SE) rule (De'Ath and Fabricius, 2000). We tested two splitting functions (information gain and Gini index) and controlled tree size through a complexity parameter corresponding to the minimum cross-validation error. Classification trees were fitted using the 'rpart' library (Therneau et al., 2013) and plotted using the 'rpart.plot' library (Milborrow, 2012) for R software (version 2.15.3, R Development Core Team, 2013). The RF models were fitted using the square root of the number of variables and a series of 10 ensemble sizes (5, 10, 20, 50, 100, 200, 500, 1000, 2000 and 5000). The 'randomForest' library (Liaw and Wiener, 2012) was used for model development. The model was fitted with 10-fold cross-validations. The BT based on stochastic gradient boosting was built using the 'gbm' library (Ridgeway, 2013) for R software. For this model, different values of *tc* (1 to 5), *lr* (0.05, 0.01 and 0.005) and bag fraction (*bf*; 0.3, 0.5 and 0.7) were combined. As the response variable is categorical, we used a Bernoulli loss function. Finally, a GLM was performed using a stepwise variable selection (family = 'binomial'; link function = 'logit'; criterion-based procedure = 'Akaike Information Criterion (AIC; Akaike, 1974)'), which is a common approach implemented in distribution modelling (stepwise GAM, Araújo et al., 2005; stepwise GLM, Barbosa et al., 2008).

4. Results

A root mean square error of 0.43 was obtained for the trend surface analysis through the application of a third-order polynomial. The computed area represents the geographical delimitation where all the models were fitted and validated (41º38′ ~ 42º06′N; 6º49′ ~ 6º17′W, Figure S4).

4.1. Tree-Based Outcomes

For classification trees, 106 out of the 377 instances were misclassified, giving an overall error of 28%. The first tree, constructed to its maximum depth, was pruned using the complexity parameter (*cp* = 0.021). The cross-validation (*cv*) reached a minimum value of 0.61 for the largest tree of size 17 (Figure 3a). There was a relative reduction in the error as the size of tree increased and the complexity parameter decreased. Based on the *cv* error and *cp*, the optimal tree size comprised six data splits and retained four variables: *DUrb*, *DSch*, *NNESS* and *NDVI* (Figure 3b). Among the CT models developed, these four variables were most frequently selected due to the amount of explained deviance. The relatively high number of splits showed that the estimated response did not depend only on the main effects. An increasing distance to urban areas, a decreasing distance to shrubs and a negative exposure to north predicted red deer presence.

Regarding the RF model, the results showed that the addition of trees to the ensemble had three main effects: i) the model discriminative performance increased, ii) the out-of-bag (OOB) error decreased (Figure S5), and iii) the cartographic projections became more stable (Figure 4). Considering the model that fits 1000 trees (highest AUC and one of lowest out-of-bag errors), the variable with the strongest effects was *DUrb*, followed by *DAgr*, *AAgr*, *DRiv*, *DSch* and *ASch*. The variables *DRoad*, *Slp*, *NDVI*, *DFor* and *AFor* were associated with moderate effects, while variables related with terrain roughness (*RNESS*) and exposure (*NNESS* and *ENESS*) showed weak effects in the prediction of red deer occurrence (Figure 5).

The best BT models were achieved using the following parameters: *lr* of 0.005, and a *tc* of 5 reserving a *bf* of 0.5. The variable importance plot showed strong effects for *DUrb*, *DRiv*, *DSch*, *DRoad* and *DAgr*, which are partially in agreement with the previous models tested. Moderate effects were observed for *RNESS*, *AAgr*, *Slp* and NNESS. Finally, weak effects were observed for *DFor*, *NDVI*, *ENESS* and *AFor* (Figure 5). The partial dependence plots showed: i) a predominantly linear positive trend for *DUrb*, ii) a downward trend for *DSch*, *Slp* and *DFor*, iii) weak effects for *RNESS*, *NNESS* and *ENESS*, and iv) modal, *i.e*. difficulty in defining a pattern or a trend, effects for the remaining variables (Figure 6). The response of red deer to the 12 most influential variables indicates that the species occurs in areas further away from human settlements, near water lines, shrubs, forest patches and

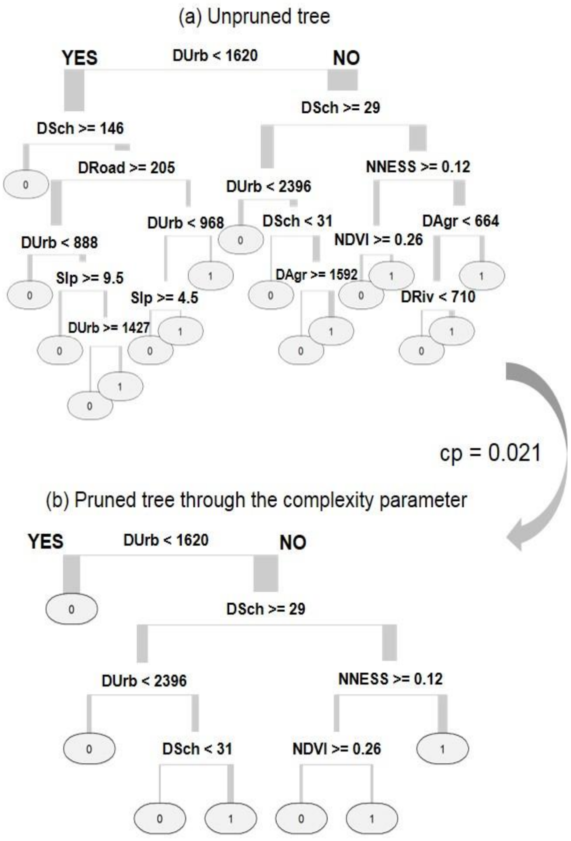

Figure 3. Classification tree of the red deer occurrence. (a) Unpruned tree; (b) Pruned tree through the complexity parameter. Branch width provides a visual proportion of instances that were clustered on each side. Predictor variable abbreviations are shown in Table 2.

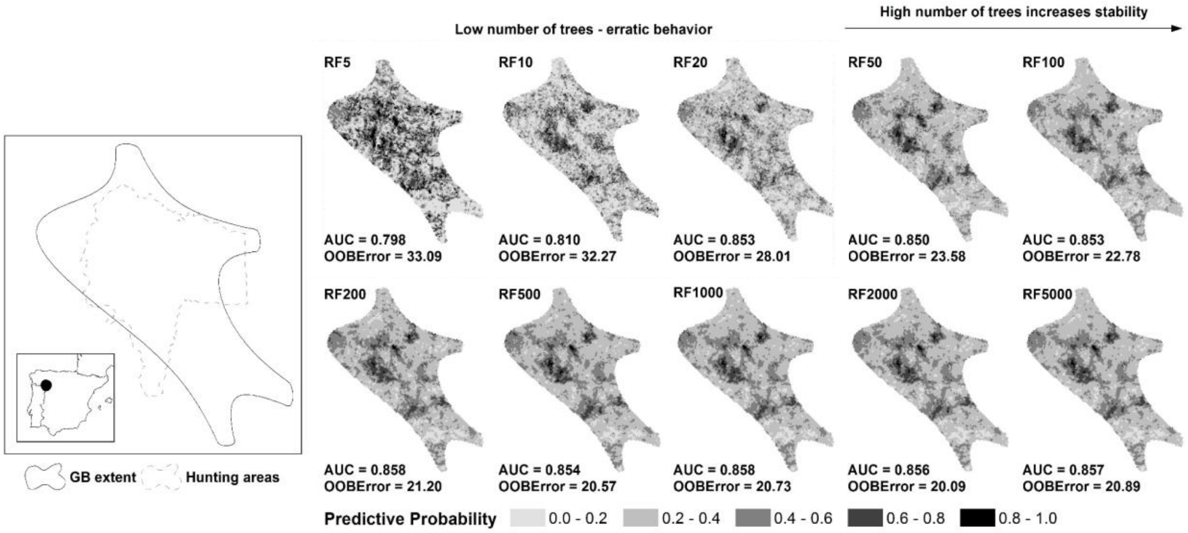

Figure 4. Cartographic representation showing an increased stability of random forest models with an increase in the number of trees in an ensemble. (AUC) - area under the curve and (OOBError) - out-of-bag error.

in areas with smooth terrain slopes. The exposure effects represented by *NNESS* and *ENESS* had little effect, although a tendency for deer occurrence in hillsides exposed to south was observed.

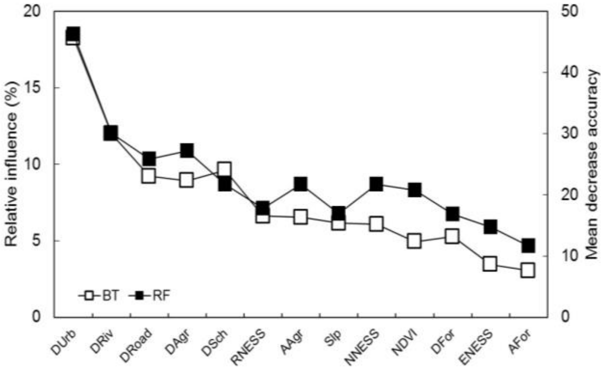

Figure 5. Variable importance plot providing a summary of the relative contributions (%) of environmental predictors in boosted trees models (open squares) and random forests (black squares).

4.2. GLM by Stepwise Regression

The most parsimonious model retained 10 variables. From the initial set of 14 environmental variables, the following factors were dropped: *DAgr*, *ASch*, *AFor*, and *ENESS*. The model outcome was partially in agreement with the tree-based methods. The *DUrb* had a positive effect in the occurrence probability of red deer. Additionally, the variables *DSch*, *Slp*, *DRoad*, *AAgr*, *DRiv*, *NNESS*, *NDVI*, *DFor* and *RNESS* had negative effects. An increase in the measured units of these variables decreased the species occurrence probability.

4.3. Evaluation of Models Performance

The results for model discrimination and threshold-dependent measures are summarized in Table 3. The predicted distributions for all the algorithms are presented in the Figure

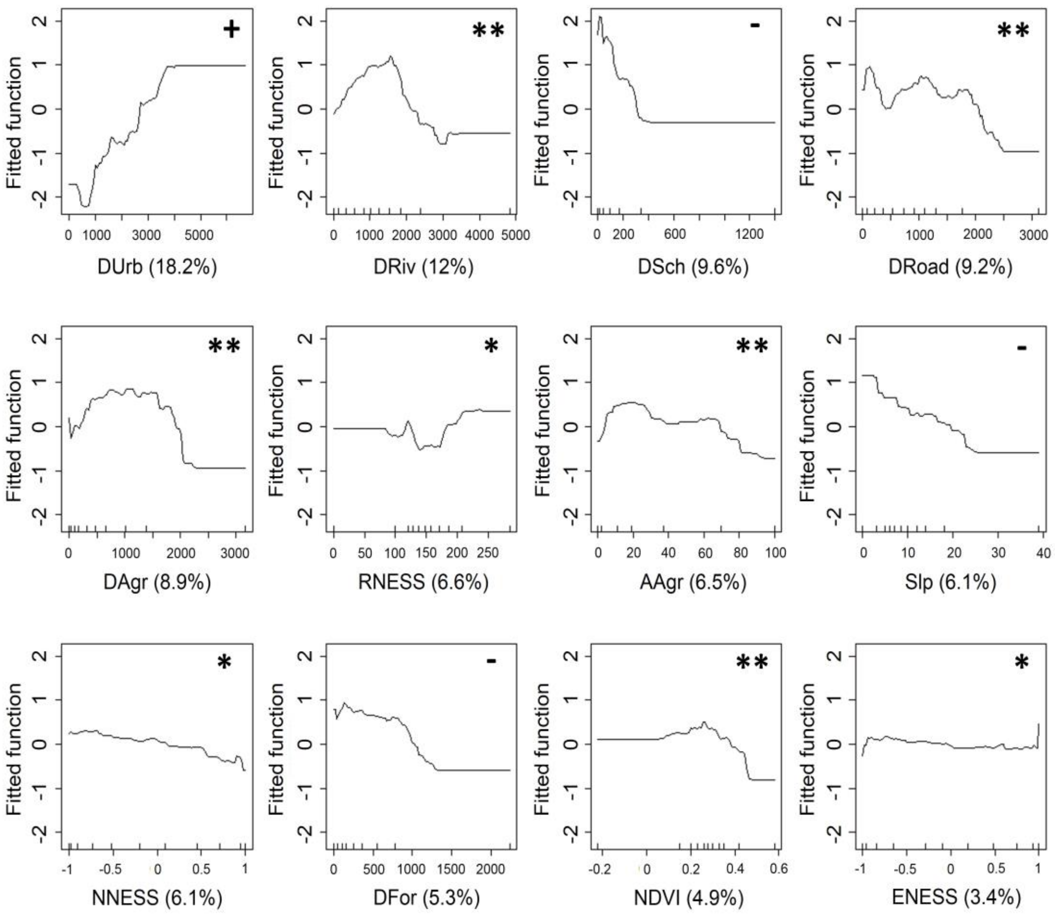

Figure 6. Partial dependence plots for the twelve most relevant predictor variables identified in boosted trees (BT) model. Rug marks at the bottom of plots show the distribution of records across the variable range, in deciles. Symbols in the upper right corner represent, (+) positive trend, (-) downward trend, (*) weak effects, and (**) modal effects.

Table 3. Threshold-Dependent and -Independent Measures Assessed by Models*

Method	CT	RF	BT	GLM
OA	0.72	0.83	0.81	0.69
Se	0.72	0.82	0.80	0.68
Sp	0.70	0.84	0.82	0.71
TSS	0.42	0.66	0.62	0.39
Cohen's kappa	0.42	0.64	0.62	0.38
AUC	0.76 (0.70)	0.86 (0.85)	0.85 (0.84)	0.74 (0.77)

*CT – Classification trees; RF – Random forests; BT – Boosted trees; GLM – Generalized linear model by stepwise regression. Performance measures evaluated were: (OA) – Overall accuracy; (Se) – Sensitivity; (Sp) – Specificity; (TSS) – True skill statistics; Cohen's kappa and (AUC) – Area under the curve. Values in brackets represent the AUC values gathered from a truly independent dataset.

7. Models discrimination and their respective spatial projections, using a truly independent dataset, are also reported (Table 3, Figure S6). The results showed positive correlations and spatial agreement between the models, indicating a partial concordance of variable importance among the methods applied (Table 4). Considering the results from all the models, the variables were ranked as follows: *Durb*, *DSch*, *DRoad*, *DRiv*, *DAgr*, *Slp*, *AAgr*, *NNESS*, *NDVI*, *RNESS*, *DFor*, *ASch*, *AFor* and *ENESS*.

5. Discussion

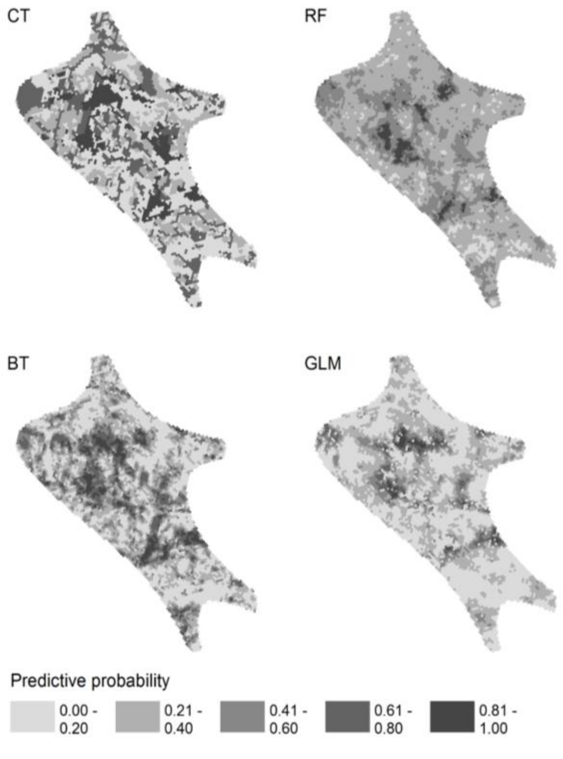

Figure 7. Predictive probability of red deer presence in the study area considering the four modelling techniques.

Table 4. Spearman's Correlation Matrix Comparing the Rank Order of Variable Selection by the Four Modelling Techniques*

	CT	BT	RF	GLM
CT	1.00	0.73	0.50	0.69
BT	-	1.00	0.62	0.65
RF	-	-	1.00	0.46
GLM	-	-	-	1.00

*CT – Classification Trees; BT – Boosted Trees; RF – Random Forests; GLM – Generalized linear model by stepwise regression.

The clear identification of environmental processes that shape species distribution and resources selection is of paramount importance for wildlife management and conservation (Guisan et al., 2013). Here, we demonstrated the usefulness and accuracy of tree-based methods to explain and predict the patterns of species' distributions at population scale. As expected, our results suggested that the predictive accuracy and spatial projections of tree-based methods vary as a function of the model parameterization. We found that the discriminative ability of RF and BT models were equivalent, which is in agreement with our first hypotheses (Figure 5; Table 3). This result corroborates a recent study where RF and BT showed similar performances (García-Callejas and Araújo, 2015). The predictive capabilities of RF while assessing species range shifts under climate change (Prasad et al., 2006), predicting invasive species (Cutler et al., 2007) and managing important economic species (Vincenzi et al., 2011) were already demonstrated. Likewise, BT proved its usefulness in modelling species richness (De'ath, 2007) and distribution even when handle sporadically sampling (Elith et al., 2008). Notwithstanding their predictive capabilities, some authors suggest to avoid RF because it may be more computationally taxing than BT (García-Callejas and Araújo, 2015). We showed that CT and GLM by stepwise regression perform substantially worse than RF and BT methods, which corroborates our second hypothesis. The high performance of RF and BT has already been documented with and without temporal transferability (Prasad et al. 2006; García-Callejas and Araújo, 2015) and may be associated with a smoother response surface in which predictive probability gradually increases without skipping classes.

Further, through a forward stagewise fitting, *i.e.* the fitted trees are kept unchanged while the number of trees added to the ensemble increased, and model averaging, BT algorithm reduces the bias and the variance of the final model. The RF algorithm cannot achieve bias reduction because the classification trees that form the ensemble are fitted in the same way, however, due to averaging, the model variance decreases. Regarding CT, one of the main model weaknesses is the fact that the final classifier probably would not be the optimal tree, which promotes erratic model behavior (Hastie et al., 2009). Non-linear relationships, strong correlations among variables and the fact that all predictors are on a continuous scale are possible reasons for the outperformance of ensemble approaches in relation to single trees. Additionally, the presence of modal effects and even a slightly change in the value of a particular variable may lead to a large variation in the predictions (Hastie et al., 2009). Nevertheless, the hierarchical structure, *i.e.* the response to one variable in top splits influences the response in the splits below, of decision trees (even single trees), makes them potentially more resilient to multicollinearity than conventional statistical approaches such as regression-like methods (*e.g.* GLM). However, the hierarchical structure of recursive partitioning techniques can be also an issue, once classification errors recorded in top splits are reflected on final data partitions. Yet, the model outcomes depend on each particular situation and involve many factors such as the spatial and environmental distribution of species (Segurado and Araújo, 2004), the structural characteristics of the data (Foody et al., 2011), the geographical background (Acevedo et al., 2012), the selection of optimal model settings (Elith et al., 2008), among others. For instance, Dettmers et al. (2002) concluded that CT and GLM showed equivalent results. Nonetheless, Franklin (1998) demonstrated that the CT outperforms GLM, but Thuiller et al. (2003) achieved opposite results. Even though some previous studies showed that ensemble methods perform consistently better than other approaches (*e.g.* GLM), a recent research reported that GLM showed higher predictive performances than RF in modelling the species richness of vascular plants (Lopatin et al., 2016). These inputs corroborate the idea that model selection is context-dependent. Some authors state that this variability in model projections endangers their applicability in real-world problems and suggest the use of multiple models in combination, *i.e.* ensemble forecasting, instead of single-model forecasts (Araújo and New, 2007).

5.1. Models Accuracy, Variable Importance and Ecological Meaning

Notwithstanding the differences in predictive accuracy, the main results gathered from the four models coincided with species ecological requirements and the predictors deemed important by the models were partially in agreement (Table 4). All the models identified the distance to urban areas and the distance to shrubs as the most important variables in predicting red deer occurrence, thus confirming previous findings (Carvalho et al., 2012; Torres et al., 2012; Torres et al., 2014). Red deer occurred further away from disturbed areas (*e.g.* villages and other human settlements), agricultural fields and near shrubs and forest patches. Additionally, the species often occurred in areas with gentle slopes, preferentially with southern exposure. The remaining variables exhibited modal effects, which makes it difficult to define patterns of selection. Some studies on the diet of red deer found a preference for shrub species like *Pterospartum tridentatum*, *Cistus ladanifer*, *Halimium lasianthum*, *Rubus ulmifolius* and *Erica* sp. (Alvarez and Ramos, 1991; Ferreira, 1998). Transition areas (ecotones) hold a great importance for red deer to cope with seasonal changes in resource availability being also important regarding the use of open feeding sites and the proximity of refuge areas against adverse weather and predators (Putman and Flueck, 2011). We showed that red deer occurred far from agricultural fields, which is in disagreement with Mysterud et al. (2002), however similar results were already reported in our study population (Torres et al., 2014). This fact can be an adaptation to contrasting seasonality, human disturbance or a hiding behavior to reduce predation risk.

5.2. Tree-Based Methods – Pros and Cons

The CT models are intuitive and easy to interpret. Nevertheless, when handling with sharp discontinuities of variables distributions, the algorithm may produce unstable decision trees and unwieldy results visualization (Hastie et al., 2009). In RF models each single tree is developed with a random subset of instances and variables, which reduces the variance. However, the method also shows some limitations once the aggregation by average does not allow bias reduction. Furthermore, the ability of RF to handle with non-symmetric error distributions is questionable (Lopatin et al., 2016). Regarding BT, one advantage is that the use of appropriate loss functions (*e.g.* Bernoulli, Poisson) allows the analysis of different response variable distributions (*e.g.* binomial, count; Natekin and Knoll, 2013). Moreover, the BT algorithm is able to reduce both bias and variance of the final outcome. While in regression-like modelling the AIC is commonly used to identify the most parsimonious model by penalizing the addition of predictor variables, in tree models a problem arises when deciding the number of splits and the tree size (Zuur et al., 2007). Large tree sizes result in a lot of information, which is then difficult to interpret. Contrarily, small trees may result in a poor fit, which may hamper the description of occurrence-environment relationships. Besides, knowing that complex models (*e.g.* trees with several splits or ensembles with hundred trees) are more likely to match training data and lose performance when applied to new unseen instances, the selection of the optimal settings represents a challenging task. Understanding how model complexity affects model predictions is beyond the scope of our paper, however it was discussed in detail by Merow et al. (2014). In tree-models, the model complexity can be controlled by several parameters addressed in introductory sections (*e.g.* minimum number of observations *per* terminal node, complexity parameter, number of trees and learning rate/shrinkage parameter).

6. Conclusions

Machine learning methods, namely those based on decision trees, are highly customizable and became increasingly prominent in the modelling arena. By offering an illustrative explanation of tree-based methods key concepts, we showed its applicability and usefulness in predicting species distribution. As reported in previous studies there is no “million dollar” model, however we corroborate some previous findings in showing that ensemble techniques perform consistently better than other approaches. Although ensemble algorithms are extremely useful tools to identify the main interactions, describe the dominant patterns and quantify the importance of predictor variables, some factors (*e.g.* attributes of the data, geographical background, variables and error distributions, multicollinearity, among others) should be considered in the selection of modelling approaches and model parameters once they may influence the model performance and the optimal settings.

Acknowledgements. We would like to thank all directors, game managers and rangers from *Instituto da Conservação da Natureza e das Florestas* (ICNF) and *Reserva Regional de Caza Sierra de la Culebra* (RRC Sierra de la Culebra) for their field support. We would like to thank J. Elith and P. Acevedo for all valuable insights on an early version of the article. We thank Maria Beatriz for reviewing the language and structure of the final version of the article. We are deeply grateful to three anonymous reviewers who provide insightful comments on this manuscript. J. Carvalho and J.P.V. Santos are supported by a PhD Grant (JC – SFRH/BD/98387/2013, JPVS – SFRH/BD/65880/2009) from *Fundação para a Ciência e a Tecnologia* (FCT). We would like to thank University of Aveiro (Department of Biology) and FCT/MEC for the financial support to CESAM RU (UID/AMB/50017) through national funds and, where applicable, co-financed by the FEDER, within the PT2020 Partnership Agreement.

References

Azhar, S., Carlton, W. A., Olsen, D., and Ahmad, I. (2011). Building information modeling for sustainable design and LEED® rating analysis. *Autom. Constr.*, 20(2), 217-224. http://dx.doi.org/10.1016/j.autcon.2010.09.019

Acevedo, P., Jiménez-Valverde, A., Lobo, J.M. and Real, R. (2012). Delimiting the geographical background in species distribution modelling. *J. Biogeogr.*, 39(8), 1383-1390. http://dx.doi.org/10.1111/j.1365-2699.2012.02713.x

Akaike, H. (1974). A new look at the statistical model identification. *IEEE T. Automat. Contr.*, 19(6), 716-723. http://dx.doi.org/10.1109/TAC.1974.1100705

Alvarez, G. and Ramos, J. (1991). Estrategias alimentarias del ciervo (Cervus elaphus L.) en Montes de Toledo, *Donana Acta Vertebrata*, 181, 63-99.

Araújo, M.B., Thuiller, W., Williams, P.H. and Reginster, I. (2005). Downscaling European species atlas distributions to a finer resolution: implications for conservation planning. *Global Ecol. Biogeogr.*, 14(1), 17-30. http://dx.doi.org/10.1111/j.1466-822X.2004.00128.x

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

Araújo, M.B., Alagador, D., Cabeza, M., Nogués-Bravo, D. and Thuiller, W. (2011). Climate change threatens European conservation areas. *Ecol. Lett.*, 14, 484-492. http://dx.doi.org/10.1111/j.1461-0248.2011.01610.x

Barbet-Massin, M., Jiguet, F., Albert, C.H. and Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods Ecol. Evol.*, 3, 327-338. http://dx.doi.org/10.1111/j.2041-210X.2011.00172.x

Barbosa, A.M., Real, R. and Vargas, J.M. (2008). Transferability of environmental favourability models in geographic space: The case of the Iberian desman (Galemys pyrenaicus) in Portugal and Spain. *Ecol. Model.*, 220(5), 747-754. http://dx.doi.org/10.1016/j.ecolmodel.2008.12.004

Birant, D. (2011). Comparison of decision tree algorithms for predicting potential air pollutant emissions with data mining models. *J. Environ. Inf.*, 17(1), 46-53. http://dx.doi.org/10.3808/jei.201100186

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*, Wadsworth International Group, Belmont, California.

Breiman, L. (2001). Statistical modelling: the two cultures. *Stat. Sci.*, 16(3), 199-231. http://dx.doi.org/10.1214/ss/1009213726

Broennimann, O., Fitzpatrick, M.C., Pearman, P.B., Petitpierre, B., Pellissier, L., Yoccoz, N.G., Thuiller, W., Fortin, M.J., Randin, C., Zimmermann, N.E., Graham, C.H. and Guisan, A. (2012). Measuring ecological niche overlap from occurrence and spatial environmental data. *Global Ecol. Biogeogr.*, 21(4), 481-497. http://dx.doi.org/10.1111/j.1466-8238.2011.00698.x

Carvalho, J., Martins, L., Silva, J.P., Santos, J., Torres, R.T. and Fonseca, C. (2012). Habitat suitability model for red deer (Cervus elaphus Linnaeus, 1758): spatial multi-criteria analysis with GIS application. *Galemys*, 24, 47-56. http://dx.doi.org/10.7325/Galemys.2012.A05

Clark, L.A. and Pregibon, D. (1992). Tree-based models. In: Chambers, J.M., and Hastie, T.J. (Eds.), *Statistical Models in S*, Wadsworth & Brooks/Cole, pp. 377-420.

Clausnitzer, V. et al. (2009). Odonata enter the biodiversity crisis debate: The first global assessment of an insect group. *Biol. Conserv.*, 142(8), 1864-1869. http://dx.doi.org/10.1016/j.biocon.2009.03.028

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

De'ath, G. and Fabricius, K.E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology,* 81(11), 3178-3192. http://dx.doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2

De'ath, G. (2007). Boosted trees for ecological modelling and prediction. *Ecology,* 88(1), 243-251. http://dx.doi.org/10.1890/00 12-9658(2007)88[243:BTFEMA]2.0.CO;2

Debeljak, M. and Džeroski, S. (2011). Decision trees in ecological modelling, in Jopp, F., Reuter, H., and Breckling, B. (Eds.), *Modelling Complex Ecological Dynamics*, Springer Verlag, pp. 197-209. http://dx.doi.org/10.1007/978-3-642-05029-9_14

Dettmers, R., Buehler, D.A. and Bartlett, J.G. (2002). A test and comparison of wildlife-habitat modelling techniques for predicting bird occurrence at a regional scale, in Scott, J.M., Morrison, M.L. and Heglund, P.J. (Eds.), *Predicting Species Occurrences: Issues of Accuracy and Scale*, Island Press, pp. 607-615.

Elith, J. et al. (2006). Novel methods improve predictions of species' distributions from occurrence data. *Ecography*, 29(2), 129-151. http://dx.doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., Leathwick, J.R. and Hastie, T. (2008). A working guide to boosted regression trees. *J. Anim. Ecol.*, 77(4), 802-813. http://dx.doi.org/10.1111/j.1365-2656.2008.01390.x

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

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. and Yates, C.J. (2010). A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.,* 17(1), 43-57. http://dx.doi.org/10.1111/j.1472-4642.2010.00725.x

Engler, R. et al. (2011). 21st century climate change threatens mountain flora unequally across Europe. *Global Change Biol.,* 17(7), 2330-2341. http://dx.doi.org/10.1111/j.1365-2486.2010.02393.x

Evans, S.E., Murphy, M.A., Holden, Z.A. and Cushman, S.A. (2011). Modelling species distribution and change using Random Forest, in Drew, C.A., Wiersma, Y.F. and Huettmann, F. (Eds.), *Predictive Species and Habitat Modelling in Landscape Ecology – Concepts and Applications*, Springer Verlag, pp. 139-160. http://dx.doi.org/10.1007/978-1-4419-7390-0_8

Ewijk, K.Y.V., Randin, C.F., Treitz, P.M. and Scott, N.A. (2014). Predicting fine-scale tree species abundance patterns using biotic variables derived from LiDAR and high spatial resolution imagery. *Remote Sens. Environ.,* 150, 120-131. http://dx.doi.org/10.1016/j.rse.2014.04.026

Ferreira, S. (1998). *Estudo da Dieta de Duas Populações de Veado (Cervus elaphus Linnaeus, 1758) em Portugal*, First degree thesis in Biology, University of Coimbra, Portugal.

Fielding, A.H. and Bell, J.F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.,* 24(1), 38-49. http://dx.doi.org/10.1017/S0376892997000088

Foody, G.M. (2011). Impacts of imperfect reference data on the apparent accuracy of species presence – absence models and their predictions. *Global Ecol. Biogeogr.,* 20(3), 498-508, http://dx.doi.org/10.1111/j.1466-8238.2010.00605.x

Franklin, J. (1998). Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *J. Veg. Sci.,* 9(5), 733-748. http://dx.doi.org/10.2307/3237291

Freund, Y. and Schapire, R.E. (1996). Experiments with a new boosting algorithm. *Proc. of the Thirteenth International Conference on Machine Learning*, pp. 148-156.

Freeman, E. (2009). ModelMap: An R Package for Modeling and Map production using Random Forest and Stochastic Gradient Boosting. USDA Forest Service.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *Ann. Stat.,* 28, 337-407. http://dx.doi.org/10.1214/aos/1016120463

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

Friedman, J.H. (2002). Stochastic gradient boosting. *Comput. Stat. Data An.,* 38(4), 367-378. http://dx.doi.org/10.1016/S0167-9473(01)00065-2

García-Callejas, D. and Araújo, M. (2015). The effects of model and data complexity on predictions from species distributions models. *Ecol. Model., 326, 4-12.* http://dx.doi.org/10.1016/j.ecolmodel.2015.06.002

Guillera-Arroita, G., Lahoz-Monfort, J.J., Elith, J., Gordon, A., Kujala, H., Lentini, P.E., McCarthy, M.A., Tingley, R. and Wintle, B.A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecol. Biogeogr.,* 24(3), 276-292. http://dx.doi.org/10.1111/geb.12268

Guisan, A. and Zimmermann, N.E. (2000). Predictive habitat distribution models in ecology. *Ecol. Model.,* 135(2-3), 147-186. http://dx.doi.org/10.1016/S0304-3800(00)00354-9

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

Guisan, A. et al. (2013). Predicting species distributions for conservation decisions. *Ecol. Lett.,* 16(12), 1424-1435. http://dx.doi.org/10.1111/ele.12189

Hastie, T., Tibshirani, R. and Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Verlag.

Hegel, T.M., Cushman, S.A., Evans, J. and Huettmann, F. (2010). Current state of the art for statistical modelling of species distributions, in Cushman, S.A. and Huettmann, F. (Eds.), *Spatial Complexity, Informatics, and Wildlife Conservation*, Springer Verlag, pp. 273-311. http://dx.doi.org/10.1007/978-4-431-87771-4_16

Hernandez, P.A., Graham, C.H., Master, L.L. and Albert D.L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography,* 29(5), 773-785. http://dx.doi.org/10.1111/j.0906-7590.2006.04700.x

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*, Springer Verlag.

Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rubel, F. (2006). World Map of the Köppen-Geiger climate classification updated. *Meteorol. Z.,* 15(3), 259-263. http://dx.doi.org/10.1127/0941-2948/2006/0130

Legendre, B.P. and Legendre, L. (1998). *Numerical Ecology(2nd ed.),* Elsevier Science.

Liaw, A. and Wiener, M. (2012). Package randomForest: Breiman and Cutler's random forests for classification and regression (accessed March 12, 2014). http://cran.r-project.org/web/packages/randomForest/index.html

Lobo, J.M., Jiménez-Valverde, A. and Hortal, J. (2010). The uncertain nature of absences and their importance in species distribution modelling. *Ecography,* 33(1), 103-114. http://dx.doi.org/10.1111/j.1600-0587.2009.06039.x

Lopatin, J., Dolos, K., Hernández, H.J., Galleguillos, M. and Fassnacht, F.E. (2016). Comparing Generalized Linear Models and random forest to model vascular plant species richness using LiDAR data in a natural forest in central Chile. *Remote Sens. Environ.,* 173, 200-210. http://dx.doi.org/10.1016/j.rse.2015.11.029

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models (2nd ed.),* Chapman and Hall.

Merow, C., Smith, M.J. and Jr, J.A.S. (2013). A practical guide to Maxent for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography,* 36(10), 1058-1069. http://dx.doi.org/10.1111/j.1600-0587.2013.07872.x

Merow, C., Smith, M.J., Edwards, T.C., Guisan, A., McMahon, S.M., Normand, S., Thuiller, W., Wuest, R.O., Zimmermann, N.E. and Elith, J. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography,* 37(12), 1267-1281. http://dx.doi.org/10.1111/ecog.00845

Milborrow, S. (2012). Package rpart.plot: Plot rpart Models. An Enhanced Version of plot.rpart (accessed February 21, 2014). http://cran.r-project.org/web/packages/rpart.plot/index.html

Mysterud, A., Langvatn, R., Yoccoz, N.G. and Stenseth, N.C. (2002). Large-scale habitat variability, delayed density effects and red deer populations in Norway. *J. Anim. Ecol.,* 71(4), 569-580. http://dx.doi.org/10.1046/j.1365-2656.2002.00622.x

Natekin, A. and Knoll, A. (2013). Gradient boosting machines, a tutorial. Front. Neurorobot., 7, 1-21. http://dx.doi.org/10.3389/fnbot.2013.00021

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

Phillips, S.J., Dudı́k, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. and Ferrier, S. (2009). Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.*, 19, 181-197. http://dx.doi.org/10.1890/07-2153.1

Phillips, S.J. and Elith, J. (2011). Logistic methods for resource selection functions and presence-only species distribution models. *Proc. of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pp. 1384-1389.

Prasad, A.M., Iverson, L.R. and Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9(2), 181-199. http://dx.doi.org/10.1007/s10021-005-0054-1

Putman, R. and Flueck, W.T. (2011). Intraspecific variation in biology and ecology of deer: magnitude and causation. *Anim. Prod. Sci.*, 51(4), 277-291. http://dx.doi.org/10.1071/AN10168

R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Ridgeway, G. (2013). Package gbm: Generalized Boosted Regression Models (accessed February 24, 2014). http://cran.r-project.org/web/packages/gbm/index.html

Sánchez-Fernández, D., Lobo, J.M. and Hernández-Manrique, O.L. (2011). Species distribution models that do not incorporate global data misrepresent potential distributions: a case study using Iberian diving beetles. *Divers. Distrib.*, 17(1), 163-171. http://dx.doi.org/10.1111/j.1472-4642.2010.00716.x

Segurado, P. and Araújo, M.B. (2004). An evaluation of methods for modelling species distributions. *J. Biogeogr.*, 31(10), 1555-1568. http://dx.doi.org/10.1111/j.1365-2699.2004.01076.x

Strobl, C., Malley, J. and Tutz, G. (2009). An introduction to recursive partitioning: rational, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods*, 14(4), 323-348. http://dx.doi.org/10.1037/a0016973

Therneau, T., Atkinson, B. and Ripley, B. (2013). Package rpart: Recursive partitioning and regression trees (accessed March 10, 2014). http://cran.r-project.org/web/packages/rpart/index.html

Therneau, T.M. and Atkinson, E.J. (2015). An Introduction to Recursive Partitioning Using the RPART Routines (accessed May 7, 2015). http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Thuiller, W. (2003). BIOMOD: Optimising predictions of species distributions and projecting potential future shifts under global change. *Global Change Biol.*, 9(10), 1353-1362. http://dx.doi.org/10.1046/j.1365-2486.2003.00666.x

Thuiller, W., Araújo, M.B. and Lavorel, S. (2003). Generalised models vs. classification tree analysis: a comparative study for predicting spatial distributions of plant species at different spatial scales. *J. Veg. Sci.*, 14(5), 669-680. http://dx.doi.org/10.1111/j.1654-1103.2003.tb02199.x

Torres, R.T., Virgós, E., Santos, J., Linnell, J.D.C. and Fonseca, C. (2012). Habitat use by smpatric red and roe deer in a Mediterranean ecosystem. *Anim. Biol.*, 62(3), 351-366. http://dx.doi.org/10.1163/157075612X631213

Torres, R.T., Santos, J. and Fonseca, C. (2014). Factors influencing red deer occurrence at the southern edge of their range: A Mediterranean ecosystem. *Mamm. Biol.*, 79(1), 52-57. http://dx.doi.org/10.1016/j.mambio.2013.09.002

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D. and Kadmon, R. (2007). A comparative evaluation of presence-only methods of model distribution. *Divers. Distrib.*, 13, 397-405. http://dx.doi.org/10.1111/j.1472-4642.2007.00346.x

Vayssiéres, M.P., Plant, R.E. and Allen-Diaz, B.H. (2000). Classification Trees: An Alternative Non-Parametric Approach for Predicting Species Distributions. *J. Veg. Sci.*, 11(5), 679-694. http://dx.doi.org/10.2307/3236575

Veloz, S.D. (2009). Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. *J. Biogeogr.*, 36(12), 2290-2299. http://dx.doi.org/10.1111/j.1365-2699.2009.02174.x

Vincenzi, S., Zucchetta, M., Franzoi, P., Pellizzato, M., Pranovi, F., De Leo, G.A. and Torricelli, P. (2011). Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, *Italy. Ecol. Model.*, 222(8), 1471-1478. http://dx.doi.org/10.1016/j.ecolmodel.2011.02.007

Vingada, J., Fonseca, C., Cancela, J., Ferreira, J. and Eira, C. (2010). Ungulates and their management in Portugal, in Apollonio, M., Andersen, R., and Putman, R. (Eds.), *European Ungulates and their Management in the 21st Century*, Cambridge University Press, pp. 392-418.

Warren, D.L., Wright, A.N., Seifert, S.N., and Shaffer, H.B. (2014). Incorporating model complexity and spatial sampling bias into ecological niche models of climate change risks faced by 90 California vertebrate species of concern. *Divers. Distrib.*, 20(3), 334-343. http://dx.doi.org/10.1111/ddi.12160

Williams, G. (2009). rattle: A graphical user interface for data mining in R(accessed May 6, 2015). http://cran.r-project.org/web/packages/rattle/index.html

Williams, G. (2011). *Data Mining with Rattle and R. The Art of Excavating Data for Knowledge Discovery*, Springer Verlag. http://dx.doi.org/10.1007/978-1-4419-9890-3

Zuur, A.F., Ieno, E.N. and Smith, G.M. (2007). *Analysing Ecological Data*, Springer Verlag. http://dx.doi.org/10.1007/978-0Zhang, H. (2008). Multi-objective simulation-optimization for earth-moving operations. *Autom. Constr.*, 18(1), 79-86. http://dx.doi.org/10.1016/j.autcon.2008.05.002