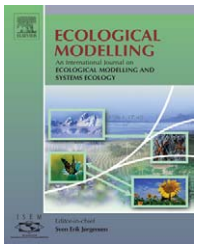
available at www.sciencedirect.com

SCIENCE DIRECT®

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



Predicting habitat suitability with machine learning models: The potential area of *Pinus sylvestris* L. in the Iberian Peninsula

Marta Benito Garzón[a,*], *Radim Blazek*[b], *Markus Neteler*[b], *Rut Sánchez de Dios*[a], *Helios Sainz Ollero*[a], *Cesare Furlanello*[b]

[a] *Biology Department, Botany Unit, Autónoma University, Carretera de Colmenar, km 15, 28049 Madrid, Spain*
[b] *Predictive Models for Biological and Environmental Data Analysis, ITC-irst. Via Sommarive 18, I-38050 Povo (Trento), Italy*

ARTICLE INFO

Article history:
Received 13 June 2005
Received in revised form 26 February 2006
Accepted 14 March 2006
Published on line 18 April 2006

Keywords:
Machine learning
Random forest
Neural networks
Classification and regression trees
AUC
Kappa
Iberian Peninsula
Pinus sylvestris L.
Habitat suitability

ABSTRACT

We present a modelling framework for predicting forest areas. The framework is obtained by integrating a machine learning software suite within the GRASS Geographical Information System (GIS) and by providing additional methods for predictive habitat modelling. Three machine learning techniques (Tree-Based Classification, Neural Networks and Random Forest) are available in parallel for modelling from climatic and topographic variables. Model evaluation and parameter selection are measured by sensitivity-specificity ROC analysis, while the final presence and absence maps are obtained through maximisation of the kappa statistic. The modelling framework is applied at a resolution of 1 km with Iberian subpopulations of *Pinus sylvestris* L. forests. For this data set, the most accurate algorithm is Breiman's random forest, an ensemble method which provides automatic combination of tree-classifiers trained on bootstrapped subsamples and randomised variable sets. All models show a potential area of *P. sylvestris* for the Iberian Peninsula which is larger than the present one, a result corroborated by regional pollen analyses.

1. Introduction

The study of the potential distribution areas of species is a discipline of great interest to many researchers, due to the difficulty involved in establishing these areas in highly modified environments like Europe. Modelling of species distributions has become necessary in many aspects of biology, ecology and biogeography. Habitat suitability models could constitute a good tool for decision-making within the framework of applied biology. They have mainly been used in strategies for conservation, planning and forest management. In addition, habitat suitability models have recently aroused greater interest on being used for predicting the movement of species in the alternative impact scenarios that might be caused by the climate change predicted by the IPCC (Bakkenes et al., 2002; Pearson et al., 2002; Thuiller, 2003). They are also of evident scientific interest with regard to gleaning more in-depth knowledge about the differences existing between actual and potential species distribution areas.

* *Corresponding author.*
E-mail address: marta.benito@uam.es (M.B. Garzón).
0304-3800/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2006.03.015

As a result of the efforts made in modelling habitat suitability, predictive techniques have become more numerous and have been improved in recent years (Guisan and Zimmerman, 2000), with a direct effect upon the quality and credibility of the models. Many different models are currently available. Among these, classical statistical models such as linear regression (Augustin et al., 2001), generalized linear models (Guisan et al., 1999), generalized additive models (Seoane et al., 2004; Luoto et al., 2005) and GRASP (generalized regression analysis and spatial prediction) (Lehmann et al., 2003), have been widely used. Also other approaches based on the delimitation of an hyperspace or envelope based on ecogeographical variables have been used for predicting habitat suitability. These models are known as environmental envelope models. Some of the most popular are BIOCLIM (Busby, 1991; Beamount et al., 2005), HABITAT (Walker and Cocks, 1991), DOMAIN (Carpenter et al., 1993) and ENFA (Hirtzel et al., 2002). Another kind of models are based on Bayesian inference (Fleishman et al., 2001; Ellison, 2004) for predicting species or communities distributions. Original approaches have also been used for specific problems such as the lack of absence data (Robertson et al., 2003; Hirtzel et al., 2002; Ottaviani et al., 2004; Phillips et al., 2006), or the use of phytosociological data as an input for prediction (Duckworth et al., 2000). In recent years, greater use has been made of machine learning methods, which comprise a series of non-parametric techniques capable of synthesising regression or classification functions based on available data. Machine learning methods present some advantages with respect to statistical methods: they are able to deal with complex relationships between predictors that can arise within large amounts of data, are able to process non-linear relationships between predictors and are able to process complex and noise data (Recknagel, 2001). The first techniques used for prediction of species distribution within machine learning methodology were classification and regression trees (Vayssières et al., 2000; Debeljak et al., 2001; Miller and Franklin, 2002; Džeroski and Drumm, 2003; Seoane et al., 2005), based on variants of the recursive partitioning CART model (Breiman et al., 1984). Later on, artificial neural networks were also utilized for building habitat suitability models (Lek and Guegan, 1999; Pearson et al., 2002; Dedecker et al., 2004), obtaining models that are complex superpositions of sigmoidal functions (Bishop, 1995). Recently, genetic algorithms (Peterson et al., 2002; Anderson et al., 2003; Dudik et al., 2004) have been used, based upon genetic and evolutionary models (Holland, 1975).

In practice, the alternative predictive techniques do not produce the same distribution areas, with differences also depending on the species under study (Robertson et al., 2003; Thuiller, 2003; Segurado and Araujo, 2004). The modelling task therefore involves testing of several predictive techniques: if the study involves many species and a high spatial resolution, developing and comparing models may easily become complex and computationally challenging.

Apart from the availability of a predictive technique adjusted to one's specific needs, other factors that might help to improve the results obtained by the models should also be taken into consideration, such as the spatial resolution of the input data. This resolution depends very much on the geographic area being covered by the model. Studies for the whole of Europe generally regard a 50 km resolution (Bakkenes et al., 2002; Thuiller, 2003). Some regional studies have been developed at higher resolution, for example, for Portugal, at 10 km (Segurado and Araujo, 2004); for the United Kingdom, studies exist at resolutions of 5 and 1 km (Pearson et al., 2004). For the Iberian Peninsula, there are regional vegetation models for the North-East of Spain (Catalonia) with 1 km grids (Rouget et al., 2001; Thuiller et al., 2003).

The Mediterranean basin is one of the areas with the highest level of plant diversity in Europe, partly due to the fact that it comprises a transition area to North African flora. Within the Mediterranean basin, some peninsulas are of particular interest, presenting a certain geographic isolation. In this study, we consider the Iberian Peninsula, one of the large-scale European hot-spots (Gómez-Campo and Malato-Béliz, 1985). The importance of this geographic region lies in the fact that it also served as a refuge to the migration of numerous European taxa during the glaciations (Hewitt, 1999). Furthermore, at present no potential vegetation model exists for the Iberian Peninsula, except for several intuitive approaches based upon the phytosociological interpretation of the vegetation series (Folch i Guillén, 1981; Rivas Martínez, 1987; Loidi and Bascones, 1995).

This study focuses upon the design of habitat suitability models at detailed scale for the whole Iberian Peninsula, with the aim of establishing the potential distribution areas of forests by comparing the predictive maps of species distribution generated by alternative methods. In order to reach this objective, we implemented a general modelling framework and used the *Pinus sylvestris* L. forests of Iberia as an example. Machine learning methods were used to support flexible modelling strategies, capable of detecting and making use, for prediction, of more complex relationships among the variables without assuming fixed hypotheses, such as a linear dependence on the predictor variables.

The Scots pine is a Northern European conifer that ranges from Eastern Siberia to Scotland, and from the Arctic in Scandinavia to its southernmost limit in Spain. In the Northern zone, its area is relatively continuous, whereas in the South it is fragmented and limited to mountain ranges (Farjon, 1984). Iberian populations or subpopulations differ morphologically and genetically from the remaining European populations (Ruby, 1967; Prus-Glowacki and Stephan, 1994; Prus-Glowacki et al., 2003), probably as a result of the Iberian Peninsula's role as a refuge during the Holocene.

With the Scots pine distribution in the Iberian Peninsula as a specific example, we designed a modelling framework for the prediction of the habitat of forest species, introducing for the first time the random forest (RF) algorithm (Breiman, 2001) for predicting species distribution areas. Within this modelling framework, we obtained presence/absence maps of the species, comparing maps obtained by three different predictive techniques. The modelling was obtained connecting two open source software systems: GRASS-GIS (Neteler and Mitasova, 2004) and R (R Development Core Team, 2004), by means of the GRASS/R interface (Bivand, 2000).

In order to improve the biological significance of the models, we propose, wherever possible, to validate the results with available biological data. This biological validation can be grounded by the use of historic data on the presence of species

in the past. For the Scots pine on the Iberian Peninsula, these data were available in the bibliography.

2. Methods

2.1. Study area

The study area comprises the Iberian Peninsula (Spain and Portugal) and the Balearic Islands. The resolution chosen for the study was 1 km grid, for a total area of 585,700 km^2.

2.2. Environmental variables

The variables used as predictors were both climatic and topographic. The topographic variables slope and aspect were added due to the detailed resolution of the model and were derived from the digital elevation model SRTM V1 DEM (Shuttle Radar Topographic Mission, at 3″ resolution) by applying the GRASS *r.slope.aspect* module. The topographic variables were: slope and aspect. The climatic variables used were: seasonal average temperature, seasonal precipitation, annual precipitation, annual average temperature, minimum average temperature of the coldest month and maximum average temperature of the warmest month. In short, a total of 14 environmental variables were considered for modelling. The climatic variables were interpolated by means of trend surfaces (Mitasova and Mitas, 1993) at 3″ resolution by the *v.surf.rst* GRASS module (Mitasova and Mitas, 1993) based on applied to a dataset derived from the Agronomic Characterisation of Spanish provinces (Sánchez Palomares et al., 1999), covering a period from 1974 to 1990 with 2605 weather stations.

2.3. Forest distribution

The presence of *P. sylvestris* L. was taken from the most recent Spanish forest map (Ruiz de la Torre, 2001). The map, at the original scale of 1:200,000, was rasterised to 1 km for this study, covering a total of 8255 grid cells indicating the presence of forests of this tree, and with a total number of cells of 585,700 for the whole Iberian Peninsula.

2.4. The modelling framework

We created a modelling framework including all the process steps to be followed in the design of predictive vegetation maps (Fig. 1). This framework was specifically created to train, select and validate models based on the predictive machine learning techniques from the available data. The final result is presented as a potential species presence/absence map. This modelling suite can be used on any data set of environmental variables, for different geographic areas and resolutions. Within the machine learning paradigm, we chose three predictive models, ranging from the simplest and most intuitive classification and regression trees to more complex methods, including Breiman's random forest algorithm, used for the first time for predicting species distribution. In this study, we wanted to develop the possible models permitted by the modelling framework in order to quantitatively and qualitatively compare the different final maps.

Within a target region, the modelling framework builds and selects the models on a randomly selected subset of the available cells. The models are then tested on a separated dataset from the remaining geographical area. In this study, the total data set available for modelling and testing comprised 16,510 cells to ensure the prevalence (defined as the frequency of species occurrence) of the model-building data of 50%.

The main phases of this modelling procedure are: model selection, training, prediction and final map selection (Fig. 1). The implementation of all the processes was mainly obtained with the use of two free software environments for data analysis and scientific computation. The geographic analysis was performed within the GRASS GIS, and the modelling analysis in the R system for statistical computing. They were connected by the GRASS-R interface (Bivand, 2000, 2004; Bivand and Neteler, 2000), also using scripts and programs of the Linux operating system.

Within the framework, the modelling process follows the order defined by different steps, as sketched in Fig. 1. The process is developed by training on data samples constituted by the environmental variables used as predictors and the presence of the species as label to be predicted. In order to evaluate the models, the original database is randomly divided in two datasets. The first one (1/3 of the original dataset) is the evaluation set, and it is used to evaluate the models in the model selection phase. The second one (2/3 of the original dataset) is the training set, and it is used to train the data in the model development phase. We will now describe the steps defining the different processes.

2.4.1. Model selection

The first step of our modelling strategy regards the selection of the most appropriate model. In this study, the evaluation set is used to develop alternative models and choose one in terms of an indicator of predictive accuracy. Different predictive models are available in our framework and tested in this study: classification and regression trees, random forest and neural networks. For each model, parameters may be tuned for optimal accuracy on new data (predictive accuracy).

The following predictive models were used within this modelling framework:

2.4.1.1. Classification and regression trees (CART). This method was applied by using the *rpart* library (Therneau and Atkinson, 1997), which provides the CART methodology (Breiman et al., 1984) also within the R statistical computing environment. CART models are developed by recursively partitioning the data set: the model is defined by a tree structure, whose nodes are associated to splits of the data along one variable (Venables and Ripley, 2002). There are two basic steps in the construction of the model: the first one involves growing a maximal tree model with the training dataset. The maximal tree is usually overfitted, i.e. the algorithm extracts complete descriptive information from the data, including noise information. The second step is focused on constraining this overfitting by pruning the tree at its best generalisation size. There are several pruning methods; in this study, a cost-complexity criterion was used (Therneau and Atkinson, 1997). This criterion is defined by one tuning parameter, cp, which sets the optimum tree as a trade-off

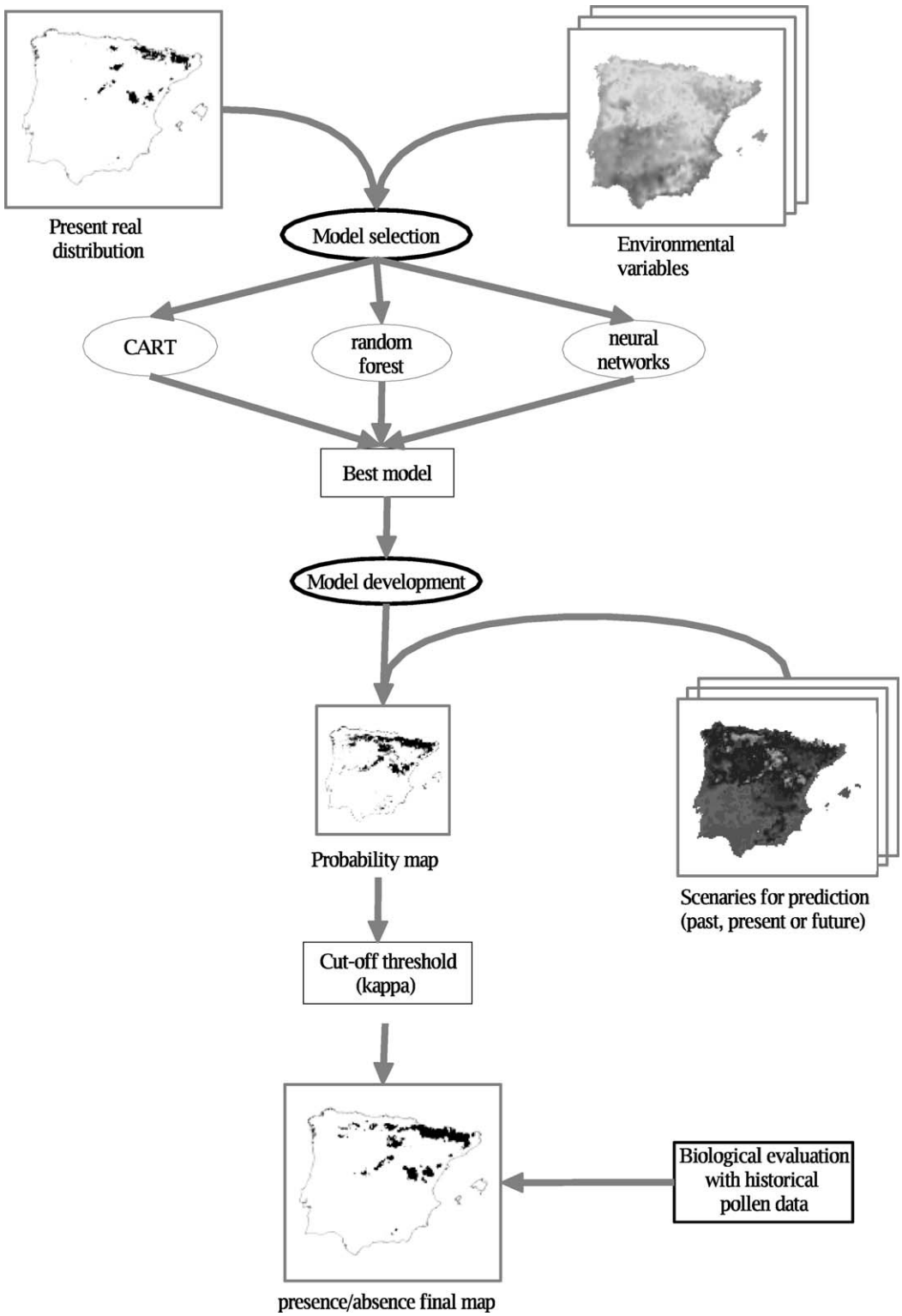


Fig. 1 – A flowchart of the main processes for the predictive mapping of species distribution.

between goodness of fit on training data and size of the tree.

2.4.1.2. *Random forest (RF)*. The *randomforest* library (Liaw and Wiener, 2002) was used within the R environment. The RF algorithm (Breiman, 2001) implements the automatic combination of tree predictors. As in bagging (Breiman, 1996), the model is obtained by combining base models trained on different bootstrap replicate samples of the data. In addition, only a random subset of the available variables is used for the candidate splitting variables at each node: this feature alleviates the problem of correlated variables because they may be extracted in turn, thus contributing to the aggregated tree model. On a battery of 20 machine learning datasets, RF gave better predictive

accuracy of single tree models (Breiman, 2001). The graphical visualization provided by CART, which has been questioned for unstability and for poorly dealing with correlated variables, is recovered by several diagnostic functions in the RF framework. In particular, the RF algorithm also provides a measure of variable importance in the modelling, both for classification as well as for regression. Importance is derived from the contribution of each variable accumulated along all nodes and all trees where it is used (Breiman, 2002). The algorithm also includes the computation of the OOB ("out of bag") error estimate, which is computed for each tree over the data remaining out of the corresponding bootstrap sample, and then averaged (Breiman, 2002). In regression, the average predicted error of RF is proven to be always lower than the predicted error of a single tree by a factor which is the correlation between residual errors of single trees (Breiman, 2001; Liaw and Wiener, 2002). The RF has been used in numerous applicative contexts: here we expand the integration of RF and GIS demonstrated in (Furlanello et al., 2003). In this study, we also used a test set in order to compare and optimise the random forest model with neural networks and regression trees. Random forest may control variance and overfitting, and it mainly requires only one tuning hyper-parameter: the number of variables randomly used at each split (*mtry*). For regression, the recommended value for *mtry* is the number of predictors divided by three (Liaw and Wiener, 2002), but it is often convenient to optimise the model by selecting an optimal value for *mtry*.

2.4.1.3. *Neural networks (NN)*. The *nnet* library (Venables and Ripley, 2002) is available in the R system and it provides a neural networks predictor. In this study, a feed-forward multilayer perceptron (MLP) was used. This NN has three types of layers of units: input, hidden and output layers. In our study, one single hidden layer architecture was used, the number of neurons in the hidden serving as a tuning hyper-parameter of the whole model. The activation function of the hidden layer units is a logistic function, and the output a linear function, an architecture generally providing good approximation capabilities (Venables and Ripley, 1999). The coefficients of the MLP are trained by minimization of an error function ($E = 1/2\Sigma(y_k - t_k)^2$); in this study the backpropagation algorithm was used to minimize the loss (Bishop, 1995). To avoid overfitting in NN, a cross-validation methodology was implemented, stopping the training network before overfit occurs (Bishop, 1995).

2.4.1.4. *Model selection*. After building these three models with the evaluation set, the selection of the best optimal one was performed. It is important to use different predictive models when working with environmental data in the prediction of habitat suitability, as it is known that strongly different responses may be obtained for different species with different predictive models (Thuiller, 2003). The selection of the best model is obtained by considering the Receiver Operating Characteristics Curve (ROC) in terms of the underlying area (AUC), a threshold independent index widely used in ecology. ROC and AUC are based on the concept of class-dependent accuracy, which may tabulated through a confusion matrix (further reading: Fielding and Bell, 1997; Manel et al., 2001; Anderson et al., 2003; McPherson et al., 2004) indicating the true positive (TP), false positive (FP), false negative (FN), and true negative (TN) predictions (Table 1). Given a model M(h) and a hyper-parameter h, the points on the ROC curve are defined, at different values of h, by the sensitivity, or true positive rate (TP/(TP + FN)), obtained as a function of the 1-specificity indicator, or false positive rate (FP/(FP + TN)). The AUC is a measure of the area under the ROC, ranging from 0.5 (random accuracy) to a maximum value of 1, which represents the most accurate model theoretically achievable.

Table 1 – Definition of the confusion matrix

Predicted	Real	
	+	–
+	TP	FP
–	FN	TN

TP: True positive, FN: False negative, FP: False positive, TN: True positive.

2.4.2. *Model development*

Once we have established the most suitable predictive method for the species, a model is developed on the training dataset (Fig. 1), including parameter tuning. Thereafter, the modelling framework will therefore only work with the most accurate model for the species. In this study, however, all the processes have been continued for the three predictive models, in order to compare the resulting maps.

2.4.3. *Predictive maps*

The next step is the application of the model over the whole region (prediction: step 3, Fig. 1). The result of this process is a probability map of the presence of the study species. In this paper, the predictive maps were developed for the present: after calibration of the model, the procedure may be applied using environmental data from simulations of the future or the past.

To facilitate the interpretation of the results, a presence/absence map is derived from the probability map. Several statistical methods derived from a confusion table have been used to get presence/absence map from probability map (Fielding and Bell, 1997; Manel et al., 2001; Liu et al., 2005): sensitivity, specificity, odds ratio, kappa, overall prediction success, normalised mutual information statistics, etc. We have generated a binary presence/absence map from the probability map according to a threshold by maximising the kappa statistic (Monserud and Leemans, 1992). The kappa statistic defines a similarity measure between the binary map and the available real or simulated biological evidence. The kappa values range from 0 to 1. In this application domain, values below 0.4 represent a low degree of similarity, between 0.4 and 0.55 an acceptable degree of similarity, between 0.55 and 0.70 good, from 0.70 to 0.85 very good, and above 0.85 excellent (Monserud and Leemans, 1992).

3. Results

In this paper we develop all the possible models permitted by the modelling framework in order to quantitatively and qualitatively compare the different final maps (Fig. 2,

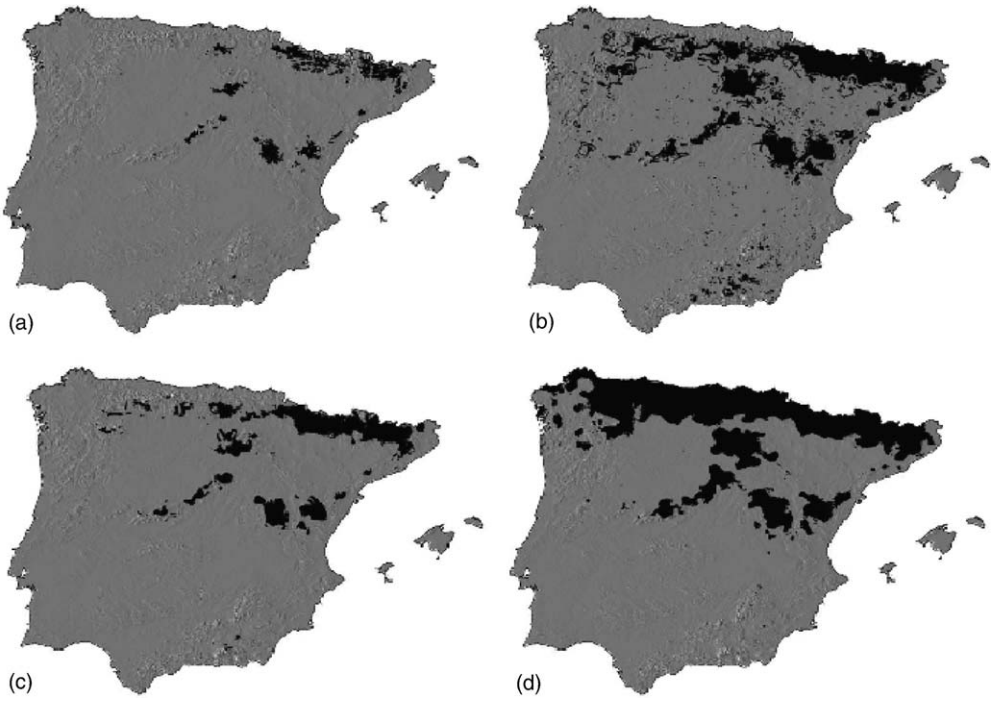


Fig. 2 – The real (a) and predicted distributions are compared in the figure, (b) regression and classification trees, (c) random forest and (d) neural networks.

Table 2 – Comparison of the accuracy prediction measures used to assess model performance

MODEL	CART	RF	NN
AUC	0.92	0.98	0.94
Kappa	0.57	0.62	0.60

AUC is used for estimating the prediction accuracy of habitat suitability and also the selection of the final model. The kappa statistic is used as an estimator of agreement of presence/absence prediction.

Tables 2 and 3); furthermore, we consider ROC plots in the analysis (Fig. 3). The results of the processes obtained for each of the predictive methods are the following:

CART: the tree was fully grown and then pruned according to the cost-complexity rule. Different tuning parameters (cp values) were tested, from 0 to 1, and the highest AUC was provided by cp = 0.1, with an AUC value of 0.92 (Table 2). The kappa statistic value used to cut off the final map was 0.57, with a threshold of 0.7, which generated the final presence/absence map (Fig. 2). Regression trees provide useful information on the variables used at each split. The variables used in the final tree model were: summer precipitation, total precipitation and minimum of average temperature of the coldest month.

RF: the final model was obtained by aggregating 500 base models. A different number of trees was also tested without significant differences. The number of variables used at each split (*mtry*) ranged between 1 and 14, obtaining the highest value of AUC = 0.98 for six variables. It is worth noting

Table 3 – A comparison of potential distribution areas, for the models (CART, RF and NN), with the actual distribution area of *Pinus sylvestris* L. in the Iberian Peninsula (Ruiz de la Torre, 2001)

MODEL	CART	RF	NN	Real
Area (km^2)	57900	32300	103800	8254

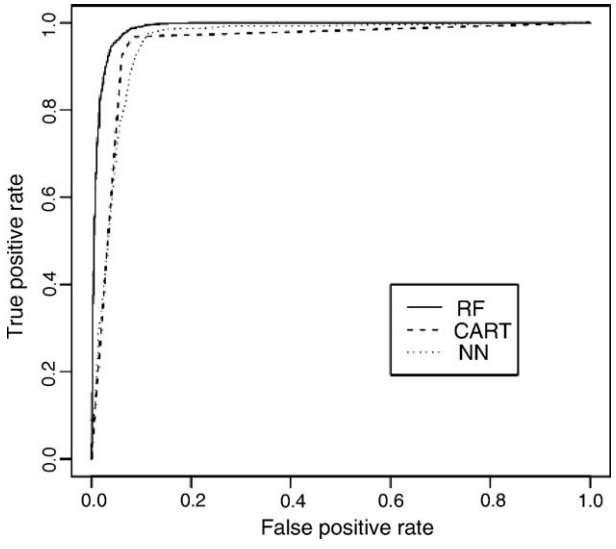


Fig. 3 – The ROC (receiver-operating) plot for random forest (solid line), neural networks (dotted line) and regression and classification trees (dashed line). For each model, the curves trace the true positive rate (or sensitivity) vs. the false positive rate (or 1-specificity) as a function of the threshold.

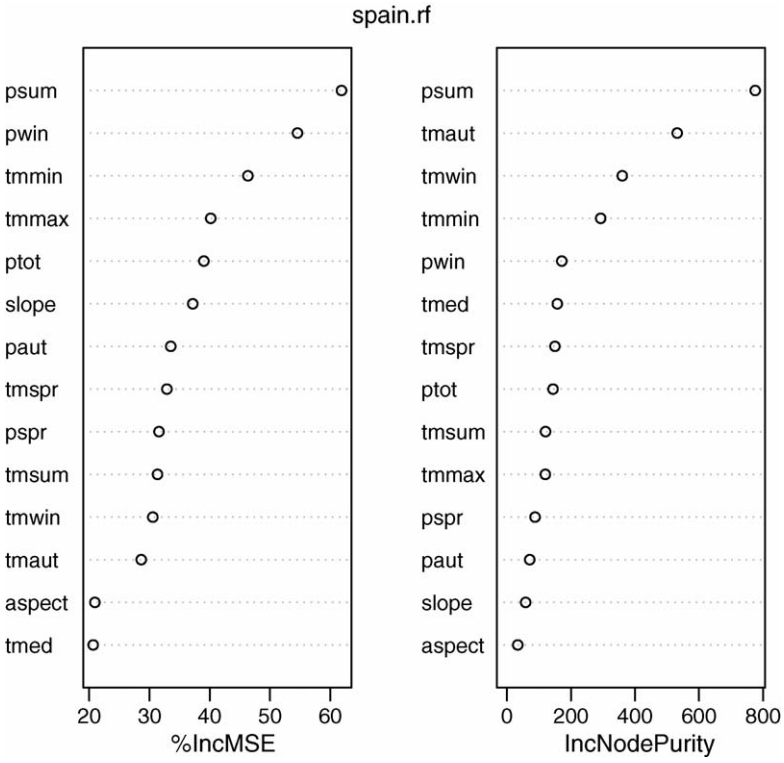


Fig. 4 – Variable importance plot generated by random forest algorithm. This plot shows the variable importance measured as increased node impurity (IncNodeImp) and also the mean square error (IncMSE). The variable full names are shown in Table 4.

that the AUC hardly changes for increasing *mtry* values. A kappa = 0.62 was obtained with a threshold of 0.8 for the final presence/absence map (Fig. 2). The variables used by RF, sorted according to an decreasing degree of importance in the modelling were: summer precipitation, autumn average temperature, winter average temperature, minimum average temperature of the coldest month, winter precipitation, annual average temperature, springtime average temperature, total precipitation, summer average temperature, maximum average temperature of the warmest month, springtime precipitation, autumn precipitation, slope and aspect (Table 4; Fig. 4).

Table 4 – Random forest variable importance order in the prediction

1	Psum	Summer precipitation
2	Tmaut	Autumn average temperature
3	Tmwin	Winter average temperature
4	Tmmin	Minimum average temperature of the coldest month
5	Pwin	Winter precipitation
6	Tmed	Annual average temperature
7	Tmsp	Springtime average temperature
8	Ptot	Total precipitation
9	Tmsum	Summer average temperature
10	Tmmax	Maximum average temperature of the warmest month
11	Pspr	Springtime precipitation
12	Paut	Autumn precipitation
13	Slope	Slope
14	Aspect	Aspect

NN: a number of neurons in the hidden layer from 10 to 60 was used to calibrate the model, with a final architecture of 40 neurons, which provides the highest AUC = 0.94 of the model (Table 2). The kappa = 0.60 for a threshold of 0.7, was obtained in correspondence to the final presence/absence map (Fig. 2).

The accuracy of models was assessed by ROC analysis (Fig. 3); high performance, with values always over 0.9, was obtained by the three models. The random forest algorithm, however, was clearly the most accurate, followed by neural networks, and then by the regression and classification tree models.

Based on the final presence/absence maps generated, we quantified the distribution area (in km^2) for the three models, and for the real distribution of the species. Significant differences were found in the predicted suitability area (Table 3) for the three models.

4. Discussion

The statistical learning modelling framework introduced in this study does not require assumption of normality of the variables and can deal with non-linear relationships. The procedures are independent from the scale resolution, geographical area and tree distribution. These features may have a substantial utility in ecology, for further applications in conservation and forest management. In particular, the approach may be used to model distribution shifts resulting from climate change. Moreover it constitutes a new approach with respect to the variety of models described in literature (i.e.

Hirtzel et al., 2002; Pearson et al., 2002; Thuiller, 2003), particularly for the incorporation of the random forest algorithm for species prediction. The results obtained with the random forest method for predicting habitat suitability are very encouraging, presenting the highest accuracy among the machine learning methods considered in this study.

CART have been previously used for species distribution (Moore et al., 1991; Iverson and Prasad, 1999; Iverson et al., 1999; Vayssières et al., 2000). One of the remarkable characteristics of CART is the simplicity involved in the modelling (De'Ath and Fabricius, 2000), which enables the variable importance for each node to be established. It was, however, the least accurate predictive model in this study (AUC = 0.92).

Neural networks have also been increasingly used for species distribution modelling (Benito Garzón et al., 2003; Thuiller, 2003; Linderman et al., 2004). The greatest shortcoming of NN is that it is hard to interpret their resulting structure, and their calibration may result mostly a "black art" to non-specialists (Caudill, 1991). NN do not easily show which variables and parameters are most important in the model construction. Furthermore, many tuneable parameters must be taken, implicitly or explicitly, into account: number of hidden layers, number of neurons in the hidden layers, weight decay, learning parameter, initial connections among the node weights, etc. When working with NNs, therefore, high predictive accuracy is attained with the use of only a careful experimental scheme which may prevent overfitting effects. In this study, the NN best model reached AUC = 0.94, a value slightly greater than that obtained with the use of the much simpler classification and regression trees.

The Random forest model has not previously been used for predicting species habitat suitability previously. In this study, RF is the most accurate algorithm (AUC = 0.98). The RF variable importance measure indicated summer precipitation as the most influence variable in the modelling (Table 4). This is expected because of the Mediterranean climate of the Iberian Peninsula, and because of the *P. sylvestris* requirements as a north European conifer that find in the Iberian Peninsula its southernmost limit. RF has enabled to establish a measure of variable importance of each variable in the model construction and also the mean square error associated (Fig. 4). When compared with the actual distribution, the kappa statistic is used to assess the final map varied among the different models, ranging from 0.57 (CART) to 0.62 for the random forest algorithm (Table 2). Thus, these results also indicate random forest as the most accurate of the three methods used. As summarised in Table 3, the RF model is also the closest in presence area to the actual distribution.

Apart from the good results obtained in the evaluation of the models using AUC, another important aspect of the modelling involves the evaluation and biological interpretation of the results obtained. In the case of *P. sylvestris* in the Iberian Peninsula, the results are encouraging because they coincide with the bibliographic data collected. Data exists on the very recent historic presence of *P. sylvestris* in the Cantabrian Mountains, a mountain range in the North of the Peninsula, where the real distribution of this species is currently very limited. In this area, pollen studies indicate that the Scots pines practically disappeared as a result of anthropic action (Costa Tenorio et al., 1990; García Antón et al., 1997; Franco Múgica et al., 2001). Presence of *P. sylvestris* is indicated in the North of the Peninsula on all the three final maps (Fig. 3). Furthermore, the maps created with the three models (Fig. 3) present an extended potential area in the Central System mountain range, a result supported by palinological studies (Franco Múgica et al., 1998). Moreover, the results of our models can be compared with those obtained for the same species by other authors. Both the results obtained by Thuiller et al. (2003), and those obtained by Rouget et al. (2001) in their models for *P. sylvestris* in the NE of Spain generally coincide with the results of our study, thus presenting a very similar potential distribution area for Catalonia. In short, a larger potential distribution area for the Scots pine is evident in relation to what can be observed at present, and this is confirmed by the results of palinological studies. This area may have been reduced in recent years by competition from other species and by intense anthropic activity which, by means of fire management, has favoured the spread of pastures in the mountains of northern Spain (García Antón et al., 1997; Sánchez Goméz and Hannon, 1999).

In biological terms, considering the three final maps, the one generated with the use of neural networks is inaccurate in the distribution area; in addition, the neural networks model forecasts a much larger potential area than the other models, up to 103,800 km^2 (Table 3). The map designed with classification and regression trees presents an excessively dispersed area, if we consider that the study was based on distributions from forests and not on isolated sampling sites. The predicted occupation area is 57,900 km^2 (Table 3). The most statistically accurate map, the one designed with random forest, is indeed the one that better supports the biological knowledge of presence. Although the occupation area predicted by RF (32,300 km^2) is more extended in relation to the actual one, it is still the smallest between the three models.

This class of species suitability models could help to clarify certain doubts regarding primitive forests. We expect that modelling will need to consider additional information, in particular data on genetics, on ecophysiology data, and on interspecies competition. Depending on the scale of analysis, different tendencies can be described. It is therefore important to integrate studies at different scales and resolutions, and in different geographic areas. At the scale used in this study (1 km^2) for a large geographic area (Iberian Peninsula and Balearic Islands), the maps we have generated may detect significant tendencies as well as small refuges and migratory routes, of specific interest due to the relative geographic isolation of this peninsula. The isolation has been corroborated by means of genetic analyses (Prus-Glowacki and Stephan, 1994; Prus-Glowacki et al., 2003). The availability of refuges has been vital on the Iberian Peninsula for the conservation of flora during the colder periods.

Species distributions are not only affected by climatic and topographic variables. The dispersal and colonization, migration rates of species, habitat fragmentation and historic factors, among others, have probably determined their current distribution. Modelling has been intensively used to calculate dispersal and migration rates of species (Labra et al., 2003; Takahashi and Kamitani, 2004; Pearson and Dawson, 2004; Soons and Ozinga, 2005). These calculations have been stepped up especially in the last years because of the global

warming that could lead shifts in species distributions. Nowadays, some models are trying to combine habitat suitability and kernel based approaches to estimate species dispersal rates in order to evaluate the migration of the species under climate change (Iverson et al., 2004). But no model was capable of integrating all the aforementioned factors that are affecting species distributions. The results of our models should therefore be interpreted with full knowledge of these inevitable limitations. Assuming these limitations, the modelling framework introduced in this study may offer new possibilities for mapping and analysing the potential vegetation at the different scale required by different geographic areas. The predictive approach is vital to decision-making in planning, resources management and conservation. It might also be relevant in the study of the potential movements and migration patterns in the projected scenarios of future climate change.

To conclude, the modelling framework presented here provided good results, with notably high and stable AUC values obtained by changing the tuning parameter achieved by the random forest learning method. To our knowledge this work represents the first time that RF is used for habitat prediction. Furthermore, with regard to the *P. sylvestris* map chosen to demonstrate the modelling strategy, we have shown that its occupation area has been restricted in the Iberian Peninsula (particularly in the mountains in the North and centre of the peninsula) in relation to its climatic capacity. The results obtained in this modelling framework are confirmed by pollen data, which indicate the presence of *P. sylvestris* in the recent past in the North of the Iberian Peninsula and a previous more extended distribution in the centre of Iberia.

Acknowledgements

This study was supported by the R+D project funded by the National Programme for Scientific Research, Development and Technological Innovation Plan of the Spanish Science and Technology Ministry (MARBOCLIM REN2003-03859). Furthermore, M. Benito Garzón has been granted a pre-doctorate scholarship from the Spanish Education and Science Ministry, and she wishes to thank ITC-irst, Italy, where part of the study was developed during a research internship. The authors also wish to acknowledge Javier Seoane's valuable comments on the early stage of the article.

REFERENCES

Anderson, R.P., Lew, D., Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol. Model. 162, 211–232.

Augustin, N.H., Cummins, R.P., French, D.D., 2001. Exploring vegetation dynamics using logistic regression and multinomial logit model. J. Appl. Ecol. 38 (5), 991–1103.

Bakkenes, M., Alkemade, J.R.M., Ihle, F., Leemans, R., Latour, B., 2002. Assessing effects of forecasted climate change on the diversity and distribution of European higher plants for 2050. Glob. Change Biol. 8, 390–407.

Beamount, L.J., Hughes, L., Poulsen, M., 2005. Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species' current and future distributions. Ecol. Model. 186, 250–269.

Benito Garzón, M., Maldonado Ruiz, F.J., Sánchez de Dios, R., Sáinz Ollero, H., 2003. Predicción de la potencialidad de los bosques esclerófilos españoles mediante redes neuronales artificiales. Graellsia 59 (2–3), 345–358.

Bishop, C., 1995. Neural Networks for Pattern Recognition. Oxford University Press, Oxford, 482 pp.

Bivand, R.S., 2000. Using the R statistical data analysis language on GRASS 5.0 GIS database files. Comput. Geosci. 26, 1043–1052.

Bivand, R.S., 2004. GRASS: Interface between GRASS 5.0 Geographical Information System and R. 29 pp. http://cran.r-project.org/src/contrib/Descriptions/GRASS.html.

Bivand, R.S., Neteler, M., 2000. Open source geocomputation: using the R data analysis language integrated with GRASS GIS and PostgreSQL data base systems. In: Proceedings of the fifth Conference on Geocomputation, University of Greenwich, UK, 23–25 August.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forest. Mach. Learn. 45, 5–32.

Breiman, L., 2002. Manual on Setting Up, using, and understanding Random Forests v3.1. 2002. http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, 368 pp.

Busby, J.R., 1991. BIOCLIM: a bioclimate analysis and prediction system. In: Margules, C.R., Austin, M.P. (Eds.), Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Melbourne, pp. 64–68.

Caudill, M., 1991. Neural networks training trips and techniques. AI Expert 6 (1), 56–61.

Carpenter, G., Gillison, A.N., Winter, J., 1993. DOMAIN: a flexible modeling procedure for mapping potential distributions of plants and animals. Biodivers. Conserv. 2, 667–680.

Costa Tenorio, M., Garcia Anton, M., Morla Juaristi, C., Sainz Ollero, H., 1990. La evolución de los bosques de la Península Ibérica: Una interpretación basada en datos paleobiogeográficos. *Ecología*, fuera de serie no 1, pp. 31–58, Madrid.

De'Ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81 (11), 3178–3192.

Dedecker, A.P., Goethals, P.L.M., Gabriels, W., De Pauw, N., 2004. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). Ecol. Model. 174, 161–173.

Debeljak, M., Džeroski, S., Jerina, K., Kobler, A., Adamič, M., 2001. Habitat suitability modelling for red deer (*Cervus elaphus* L.) in South-central Slovenia with classification trees. Ecol. Model. 138, 321–330.

Duckworth, J.C., Bunce, R.G.H., Malloch, A.J.C., 2000. Vegetation gradients in Atlantic Europe: the use of existing phytosociological data in preliminary investigations on the potential effects of climate change on Brithis vegetation. Glob. Ecol. Biogeogr. 9, 187–199.

Dudik, M., Philips, S.J., Shapire, R.E., 2004. A maximum entropy approach to species distribution modelling. In: Proceedings of the 21st International Conference on Machine Learning.

Džeroski, S., Drumm, D., 2003. Using regression trees to identify the habitat preference of the sea cucumber (Holothuria leucospilota) on Rarotonga, Cook Islands. Ecol. Model. 170, 219–226.

Ellison, A.M., 2004. Bayesian inference in ecology. Ecol. Lett. 7 (6), 509–520.

Farjon, A., 1984. Pines. In: Brill, E.J. (Ed.), Drawings and Descriptions of the Genus Pinus. Leiden, The Netherlands, p. 220.

Fleishman, E., Mac Nally, R., Fay, J.P., Murphy, D.D., 2001. Modeling and predicting species occurrence using broad-scale environmental variables: an example with butterflies of the Great Basin. Conserv. Biol. 15 (6), 1674–1685.

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49.

Folch i Guillén, R., 1981. Dominis de vegetació del Països Catalans, in: Ketres (Ed.), La vegetació dels Països Catalans. Institució Catalana d'Historia Natural, 513 pp.

Franco Múgica, F., García Antón, M., Maldonado Ruiz, J., Morla Juaristi, C., Sainz Ollero, H., 2001. The Holocene history of *Pinus* forests in the Spanish Northern Meseta. Holocene 11 (3), 343–358.

Furlanello, C., Neteler, M., Merler, S., Menegon, S., Fontanari, S., Donini, A., Rizzoli, A., Chemini., C., 2003. GIS and the Random Forest Predictor: Integration in R for Tick-borne Disease Risk Assessment, in: K. Hornik, F. Leisch (Ed.), Proceedings of the DSC-03 International Workshop on Distributed Statistical Computing. Vienna, Austria, March 20–22.

Franco Múgica, F., García Antón, M., Sainz Ollero, H., 1998. Vegetation dynamics and human impact in the Sierra of Guadarrama, Central System, Spain. Holocene 8 (1), 69–82.

García Antón, M., Franco Múgica, F., Maldonado Ruiz, J., Morla Juaristi, C., Sainz Ollero, H., 1997. New data concerning the evolution of the vegetation in Lillo Pinewood (León, Spain). J. Biogeogr. 26, 929–934.

Gómez-Campo, C., Malato-Béliz, J., 1985. The Iberian Peninsula. In: Plant Conservation in the Mediterranean Area. Junk Publishers, Dordrecht, 269 pp.

Guisan, A., Weiss, S.B., Weiss, A.D., 1999. GLM versus CCA spatial modelling of plant species distribution. Plant Ecol. 143, 107–122.

Guisan, A., Zimmerman, N.E., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.

Hewitt, G.M., 1999. Post-Glacial re-colonization of European biota. Biol. J. Linn. Soc. 68, 87–112.

Hirtzel, A.H., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83 (7), 2027–2036.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. University of Michigan Press, Ann Arbor, 183 pp.

Iverson, L.R., Schwartz, M.W., Prasad, A.M., 2004. Potential colonization of newly available tree-species habitat under climate change: an analysis for five eastern US species. Landscape Ecol. 19, 787–799.

Iverson, L.R., Prasad, A.M., Schwartz, M.W., 1999. Modelling potential future individual tree-species distribution in the eastern United Estates under a climate change scenario: a case study with Pinus virginiana. Ecol. Model. 115, 77–93.

Iverson, L.R., Prasad, A.M., 1999. Predicting abundance for 80 tree species following climate change in the eastern United States. Ecol. Monogr. 68 (4), 465–485.

Labra, F.A., Lagos, N.A., Marquet, P.A., 2003. Dispersal and transient dynamics in metapopulations. Ecol. Lett. 6, 197–204.

Lehmann, A., Overton, J.M.C., Leathwick, J.R., 2003. GRASP: generalized regression analysis and spatial prediction. Ecol. Model. 160, 165–183.

Lek, S., Guegan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecol. Model. 120, 65–73.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. The Newsletter of R Project 2/3, 18–22.

Liu, C., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28, 385–393.

Linderman, M., Liu, J., Qi, J., An, L., Ouyang, Z., Yang, J., Tan, T., 2004. Using artificial neural networks to map the spatial distribution understorey bamboo from remote sensing data. Int. J. Remote Sens. 25 (9), 1685–1700.

Loidi, J., Bascones, J.C., 1995. Memoria del mapa de series de vegetación de Navarra. 1:200.000. Gobierno de Navarra. Dpto. de Ordenación del Territorio y Medio Ambiente. 99 pp.

Luoto, M., Pöyry, J., Heikkinen, R.K., Saarinen, K., 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. Glob. Ecol. Biogeogr. 14, 575–584.

Manel, S., Williams, H.C., Ormerod, S.J., 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38, 921–931.

McPherson, J.M., Jetz, W., Rogers, D.J., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? J. Appl. Ecol. 41, 811–823.

Miller, J., Franklin, J., 2002. Modelling distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecol. Model. 157, 227–247.

Mitasova, H., Mitas, L., 1993. Interpolation by regularized spline with tension. I. Theory and implementation. Math. Geol. 25, 641–655.

Monserud, R.A., Leemans, R., 1992. Comparing global vegetation maps with the Kappa statistics. Ecol. Model. 62, 275–293.

Moore, D.M., Lees, B.G., Davey, S.M., 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographical information system. Environ. Manag. 15 (1), 59–71.

Neteler, M., Mitasova, H., 2004. Open Source GIS: A GRASS GIS Approach, 2nd ed. Kluwer Academic Publishers/Springer, Boston, 420 pp.

Ottaviani, D., Lasinio, G.J., Boitani, L., 2004. Two statistical methods to validate habitat suitability models using presence-only data. Ecol. Model. 179 (4), 417–443.

Pearson, R.G., Dawson, T.P., Berry, P.M., Harrison, P.A., 2002. SPECIES: a spatial evaluation of climate impact on the envelope of species. Ecol. Model. 154, 289–300.

Pearson, R.G., Dawson, T.P., 2004. Long-distance plant dispersal and habitat fragmentation: identifying conservation targets for spatial landscape planning under climate change. Biol. Conserv. 123, 389–401.

Pearson, R.G., Dawson, T.P., Liu, C., 2004. Modelling species distributions in Britain: a hierarchical integration of climate and land-cover data. Ecography 27, 285–298.

Peterson, A.T., Ortega-Huerta, M.A., Bartley, J., Sánchez-Cordero, V., Soberón, J., Buddemeier, R.H., Stockwell, D.R.B., 2002. Future projections for Mexican faunas under global climate scenarios. Nature 416, 626–629.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259.

Prus-Glowacki, W., Stephan, B.R., 1994. Genetic variation of *Pinus sylvestris* from Spain in relation to other European populations. Silvae Genet. 43, 7–14.

Prus-Glowacki, W., Stephan, B.R., Brujas, E., Alia, R., Marciniak, A., 2003. Genetic differentiation of autochthonous populations of *Pinus sylvestris* (Pinaceae) from the Iberian Peninsula. Plant Syst. Evol. 239, 55–66.

R Development Core Team, 2004. R: a language and environment for statistical computing. R Foundation for

Statistical Computing. Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org/.
Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecol. Model. 146, 303–310.
Rivas Martínez, S., 1987. Mapa de series de vegetación de España 1:400.000 y memoria. ICONA. Ministerio de Agricultura. Pesca y Alimentación, Madrid, 268 pp.
Robertson, M.P., Peter, C.I., Villet, M.H., Ripley, B.S., 2003. Comparing models for predicting species' potential distributions: a case study using correlative and mechanism predictive modelling techniques. Ecol. Model. 164, 153–167.
Rouget, M., Richardson, D.M., Lavorel, S., Vayreda, J., Gracia, C., Milton, S.Z., 2001. Determinants of distribution of six Pinus species in Catalonia, Spain. J. Veg. Sci. 12, 491–502.
Ruby, J.L., 1967. The correspondence between genetic, morphological and climatic variation patterns in Scoth Pine. Silvae Genet. 16, 50–56.
Ruiz de la Torre, J., (dir.), 2001. *Mapa Forestal de España*. Escala 1:200.000. ICONA. Ministerio de Agricultura. Pesca y Alimentación, Madrid.
Sánchez Goméz, M.F., Hannon, G.E., 1999. High-altitude vegetation pattern on the Iberian Mountain Chain (north-central Spain) during the Holocene. Holocene 9 (1), 39–57.
Sánchez Palomares, O., Sánchez Serrano, F., Carretero Carretero, P., 1999. *Modelos y Cartografía de estimaciones climáticas termopluviométricas para España peninsular*. INIA. Ministerio de Agricultura. Pesca y Alimentación, 192 pp.
Segurado, P., Araujo, M.B., 2004. An evaluation of methods for modelling species distributions. J. Biogeogr. 31 (10), 1555–1568.
Seoane, J., Bustamante, J., Diaz-Delgado, R., 2004. Competing roles for landscape, vegetation, topography and climate in predictive models of bird distribution. Ecol. Model. 171, 209–222.
Seoane, J., Carrascal, L.M., Alonso, C.L., Palomino, D., 2005. Species-specific traits associated to prediction errors in bird habitat suitability modelling. Ecol. Model. 185, 299–308.
Soons, M.B., Ozinga, W.A., 2005. How important is long-distance seed dispersal for the regional survival of plant species? Divers. Distrib. 11, 165–172.
Takahashi, K., Kamitani, T., 2004. Effect of dispersal capacity on forest plant migration at landscape scale. J. Ecol. 92, 778–785.
Therneau, T., Atkinson, E., 1997. An introduction to recursive partioning using the rpart routine. Tech. Rep. 61, Section of Biostatistics, Mayo Clinic, Rochester.
Thuiller, W., 2003. BIOMOD: optimizing predictions of specie distributions and projecting potential future shifts under climate change. Glob. Change Biol. 9, 1353–1362.
Thuiller, W., Vayredo, J., Pino, J., Sabate, S., Lavorel, S., Gracia, C., 2003. Large-scale environmental correlates of forest tree distribution in Catalonia (NE Spain). Global Ecol. Biogeogr. 12, 313–325.
Vayssièrs, M.P., Richard, R.E., Allen-Diaz, B.H., 2000. Classification trees: an alternative non-parametric approach for predicting species distribution. J. Veg. Sci. 11, 679–694.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, 4th ed. Springer, 495 pp.
Walker, P.A., Cocks, K.D., 1991. HABITAT: a procedure for modeling a disjoint environmental envelope for a plant or animal species. Global Ecol. Biogeogr. 1, 108–118.