International Journal of Geo-Information

MDPI

Article

Development and Comparison of Species Distribution Models for Forest Inventories

Óscar Rodríguez de Rivera * and Antonio López-Quílez

Department of Statistics and Operational Research, Faculty of Mathematics, University of Valencia, 46100 Burjassot (València), Spain; antonio.lopez@uv.es

* Correspondence: osroderi@alumni.uv.es; Tel.: +44-(0)7858-714047

Academic Editors: Duccio Rocchini and Wolfgang Kainz

Received: 19 January 2017; Accepted: 14 June 2017; Published: 16 June 2017

Abstract: A comparison of several statistical techniques common in species distribution modeling was developed during this study to evaluate and obtain the statistical model most accurate to predict the distribution of different forest tree species (in our case presence/absence data) according environmental variables. During the process we have developed maximum entropy (MaxEnt), classification and regression trees (CART), multivariate adaptive regression splines (MARS), showing the statistical basis of each model and, at the same time, we have developed a specific additive model to compare and validate their capability. To compare different results, the area under the receiver operating characteristic (ROC) function (AUC) was used. Every AUC value obtained with those models is significant and all of the models could be useful to represent the distribution of each species. Moreover, the additive model with thin plate splines gave the best results. The worst capability was obtained with MARS. This model's performance was below average for several species. The additive model developed obtained better results because it allowed for changes and calibrations. In this case we were aware of all of the processes that occurred during the modeling. By contrast, models obtained using specific software, in general, perform like "hermetic machines", because it could sometimes be impossible to understand the stages that led to the final results.

Keywords: additive model; area under the curve; AUC; forest inventory; receiver operating characteristic; ROC; species distribution model

1. Introduction

Species distribution models (SDMs) are mathematical tools based on combination of observations of species occurrence or abundance with environmental variables. These tools are used to analyze species distributions across landscapes [1].

In SDM we usually follow the following processes: (1) compile the locations of the presence of the species; (2) from databases we obtain different values of environmental variables (precipitation, temperature, etc.) for the compiled locations; (3) these environmental variables fit the models to estimate the relationship between sites of occurrence or species richness; and (4) the models are tools to predict the variable of interest across the space or time of interest [2].

Species distribution models comprise three main components: an ecological model, data, and a statistical model [3]. The most pertinent point in statistical modeling is the selection of the mathematical model, because a wrong selection may reduce the predictive power. Ecological modeling experts have shown a keen interest in the effects of mathematical methods on the predictive capacity of distribution models (e.g., [4,5]). A group from California University's National Center for Ecological Analysis and Synthesis (NCEAS) carried out the most comprehensive study of modeling techniques to date [6]. Their research evaluated the predictive ability of sixteen methods with presence-only or presence-pseudo-absence data on six regions with more than 200 species. Results showed that new

methods, such as maximum entropy (MaxEnt), have greater predictive power than other methods, such as logistic regression (both adjusted generalized linear models, GLM, and adjusted generalized additive models, GAM). Subsequent studies have also obtained better predictive capacity for MaxEnt than for logistic regression [7–11].

Decision support tools for plant species selection for ecological/environmental management have been based on species distribution models (also called ecological niche models) that analyze the probability of the presence of the species as a function of environmental variables (e.g., precipitation, temperature, or soil properties). The idea of developing statistical models, with several variables, to predict the potential distribution of species could be a complex task requiring in-depth study of several statistical methods that provide fairly inconsequential results. Several tools are now available to facilitate this task. Several studies have compared the performance of different statistical approaches to predict species distributions, obtaining a variety of suggestions about model selection [5,6,12].

Some of the latest species distribution models only use the presence of the species in the modeling process. Other methods use presence/absence data or 'background' data. Logistic regression is the traditional approach to analyzing presence/absence data [2]. Our study uses a large dataset with presence-absence and, therefore, requires a method that can use these data; in other words, a method with presence-absence data. We understand that MaxEnt is not a presence-absence method; in fact, it uses the presence-only data and a user-defined number of randomly-selected points, combining these with the covariates to build an index of habitat suitability for each cell ranging from 0 (least suitable habitat) to 1 (most suitable habitat). Moreover, MaxEnt was included in this analysis because of is one of the most commonly used methods as a species distribution model, as we can see summarized in [13].

The aim of this paper is two-fold. The first aim is to present and explain, from a mathematical point of view, different common tools designed for species distribution modeling. Our main target is to show how these tools develop models based on regressions, and explore the advantages and disadvantages of each model. The second aim is to compare these models and decide which is the most accurate according to easily understood indicators.

This study uses real data on seventeen forest species obtained from the Spanish National Inventory, supported with environmental variables. These species, with presence-absence, were first located with geographical coordinates. We then generated distribution models with tools designed to create this kind of model. Finally, we developed an additive model with R and compared the results from it to evaluate the predictive capability of all of the models in an attempt to answer the following questions: Does any one statistical technique have a regularly greater predictive ability than the others for all types of relationships between environmental variables and the presence of the species? [14] Are species with a higher presence easier to predict than others that are less represented?

The paper proceeds as follows: In Section 2 (Material and Methods), we review the principal properties of different models and the way each model is evaluated in order to compare the prediction capability. In Section 3 (Results) we summarize the results of each distribution model and evaluate them to conclude which one we consider to be the most accurate and which has the best prediction capability. Finally in Section 4 (Discussion), we analyze the potential of that model.

2. Materials and Methods

2.1. Species Occurrence Data

We have used the Spanish National Forest Inventory (NFI) dataset to elaborate our research project. NFI comprises a systematic grid with 91,889 plots, each of which is 0.2 ha in size. From this dataset, we started by choosing 17 forest species with presence/absence in each plot. The species analyzed with the percentage of presence are as follows: *Abies alba* Miller (<5%), *Castanea sativa* Miller (5%), *Fagus sylvatica* L. (5.5%), *Pinus halepensis* Miller (15%), *Pinus nigra* Arnold (9.5%), *Pinus pinea* L. (15%), *Pinus pinaster* Aiton (15%), *Pinus sylvestris* L. (12.3%), *Pinus uncinata* Turra (<5%), *Quercus*

canariensis Willd. (<5%), *Quercus faginea* Lam. (11%), *Quercus humilis* Miller (<5%), *Quercus ilex* L. (36%), *Quercus petraea* (Matt.) Liebl (<5%), *Quercus pyrenaica* Willd. (8.2%), *Quercus robur* L. (8.6%), and *Quercus suber* L. (5%). In this paper we only report the most characteristic species to evaluate the statistical processes used in the study.

2.2. Environmental Predictors

We have obtained the climatic data grids by applying the models for climatic estimation produced by [15] to the Shuttle Radar Topography Mission (STRM) 3-arc-second (≈90 m) elevation dataset [16]. These models interpolate monthly climate data from weather stations using latitude, longitude, and elevation as independent variables. We have analyzed 10 climatic predictors commonly used in tree species autoecology in Spain [17]: mean summer rainfall (SR), mean annual rainfall (R), mean summer temperature (ST), mean annual temperature (T), mean of maximum temperatures of the warmest month (MTWM), mean of minimum temperatures of the coldest month (MTCM), mean annual potential evapotranspiration (ETP), mean annual water surplus (SUP), and mean annual water deficit (DEF). Moreover, we have used the European Soil Database [18] to allocate each plot to a parent material class (calcareous or siliceous) (C). The distribution of calcareous parent materials is a very useful predictor of plant species distribution in Mediterranean ecosystems [19].

Model selection was based on the ease of working with each model and the possibility of repeating each process with the same characteristics using the species studied. Therefore, we constructed the models with one of the most widely-used models (MaxEnt), and others based on simple software developed by Saldorf System (San Diego, CA, USA) (CART and MARS). Finally, we built an additive model using R software.

The statistical methods used in this study are summarized below.

2.3. MaxEnt (Maximum Entropy)

MaxEnt [20,21] is an artificial intelligence method based on the statistical principle of maximum entropy. Models are limited by the value of the variables used to develop the problem. For example, the expected value (mean value predicted by the model) of each independent variable must match its empirical average (the mean value observed when sampling with an independent variable occurrence data item). MaxEnt obtains the maximum entropy probability of the distribution; in other words, the distribution nearest to the uniform distribution, with all of the conditions. Additionally, MaxEnt is based on the following points: (a) the presence of a species is represented by a likelihood function P on a set x of points in the study zone. P gives a positive value x everywhere so that the sum of P(x) is unity; (b) building a model of P with a group of constraints obtained from the empirical data of presence; (c) the restrictions are expressed as a simple function of known environmental variables, f(v); (d) in the MaxEnt method, the average forces of each function of each variable are close to the actual average of the variable zones of presence; and e) of the possible options available, a specific combination of features is selected to minimize the entropy function (measured by the Shannon index). The entropy function allows optimal selection of variables and functions based on their significance, and eliminates restrictions that do not provide the model with significance.

The general form of the probability function is, with i environmental variables:

$$P(x) = e^{\lambda f(x)} / Z_{\lambda}$$

where lambda is a weighting coefficient vector and f is the vector corresponding to the functions. Z is a normalization constant used to ensure that P(x) is the unit. The values P(x) obtained should be interpreted as relative suitability values. These values are normally processed by a logistic function that is adjusted to a more comprehensible level in the range between 0 (incompatible) and 1 (ideal).

Hypothetically, MaxEnt is most similar to generalized linear models and additive models. In what follows, we use the expressions of [22]. A commonly-used linear model is the Gaussian logit model, in which the logit of the predicted probability of occurrence is:

$$\alpha + \beta_1 f_1 + \gamma_1 f_1(x)^2 + \cdots + \beta_n f_n + \gamma_n f_n(x)^2$$

where the f_j are environmental predictors; α, β_j, and γ_j are fixed coefficients; and the logit function is defined by $\text{logit}(p) = \ln(p/(1-p))$. The above expression is no different in form as the log (rather than logit) of the likelihood of the pixel x in a MaxEnt expression with linear and quadratic structures. A common method for modeling interactions between variables in a linear model is to create product predictors, which is equivalent to the use of it in MaxEnt [21].

From a similar point of view, if the probability of presence/absence is modeled with an additive model using a logit link function, the logit of the predicted likelihood has the form:

$$g_1(f_1(x)) + \cdots + g_n(f_n(x))$$

where f_i are the environmental predictors. g_i are smooth functions fitted by the expression, with the quantity of smoothing measured by a measurement factor. This is a similar method as the log probability in a MaxEnt model, for pixel x, with threshold structures, and regularization has an equivalent effect to smoothing on the otherwise random functions g_i. In both cases, the form of the response curve to each environmental predictor is determined by the data.

During the process, MaxEnt generates different probability distributions, opening from a uniform scattering, and improves the fitting to the data. This improvement is defined as the average possibility of occurrence data, removing a constant, which means that the uniform distribution has a gain of zero.

Regardless of these similarities, several differences exist between generalized models and MaxEnt, leading them to create different results. When GLM/GAMs are developed to model the probability of presence, absences are needed. When applied to presence-only data, background pixels should be used as an alternative of true nonappearances [12,23]. However, the interpretation of the output is less clear-cut—it must be taken as a relative guide of ambient suitability. Dissimilarly, MaxEnt models a probability of presence over the pixels in the area of study, and on no account are pixels without records interpreted as no presences. Additionally, MaxEnt is a generative method, although GLM/GAMs are discriminative, and generative methods may give better likelihoods when the quantity of training data is insignificant [24].

For all species we use the model with the same variables, obtaining the following results shown below for each of the species under study.

2.4. MARS (Multivariate Adaptive Regression Splines)

MARS is a statistical method developed by Friedman [25]. It involves designing flexible models in which the data are adjusted to partial regressions. When models are nonlinear, they are approximated by partial linear regression, where the grade of the equation changes from one step to another, establishing a node between the end of one linear regression and the beginning of the next.

A node indicates the end of one partial regression and the beginning of another. Between two consecutive nodes, logically, the model is defined by a linear regression. The nodes are selected with the aid of a search procedure that generates a stepper algorithm. The model generated is overfitted, so the less relevant nodes are subsequently removed using a statistical approach known as generalized cross-validation. Finally, we only consider the most significant nodes. The function is a parameter interceptor β_0, and β_i is the weighted sum of one or more basic functions FB_i. Therefore, the model will consist of a weighted sum of selected basic expressions from a large number of basic expressions that link all of the values of the predictor. The model is generated as follows:

$$f(x) = \beta_0 + \sum \beta_i f_i FB_i$$

$$FB_i = \max(0, V - N)$$

$$FB_{i+1} = \max(0, N - V)$$

where FB is a basic function and acts as a new variable, V is the variable and N is the node.

Going deeper into the MARS algorithm, note that the models are constructed from double-sided truncated functions of the form (see Figure 1):

$$(x - t)_+ = (x - t; x > t/0; \text{other}$$

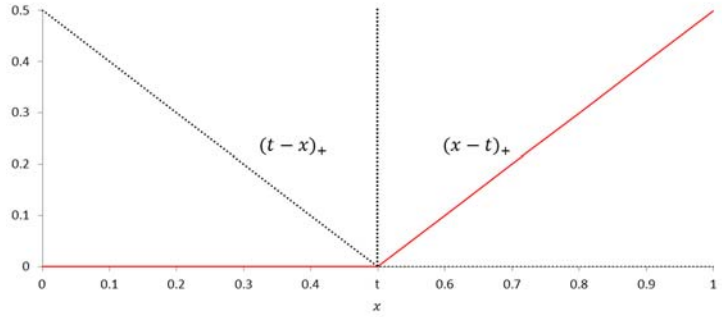


Figure 1. Representation of two basic functions, where parameter t is the knot. The (+) signs denotes that only positive values are considered.

Each expression is in linear pieces, with a lump in the value t, where each node is located at the end of one region of the data and starts at a different one [26]. The Salford Predictive Modeler Builder v6.6 (www.salford-systems.com) was used to generate these models.

2.5. CART (Classification and Regression Trees)

This method was established by Brieman et al. [27] and generates binary trees (the parent nodes are divided into two child nodes) by iterative partitions, in a process that can be repeated to attempt to turn each child node into a parent node. The algorithm searches for the optimal cutoff values among all of the independent variables to obtain an optimal set of binary divisions, so as to minimize the variance within each node and maximize it between different nodes; it is, therefore, possible that some variables will be unused. Once the tree that best classifies the cases has been identified, with no limits on complexity, the algorithm 'prunes', or simplifies, to avoid overfitting of the data. The result is a tree that establishes yes/no questions. Depending on the kind of dependent variable there can be two types of trees: regression (continuous dependent variable) and classification (discrete variable).

The aim of this method is to discriminate, estimate, or predict Y-based predictors X_1, ..., X_p by successive partitions or by sets of individuals, maximizing a measure of information content with respect to the response variable. In the validation phase this same design, a training matrix, or a similar, but independent, matrix (validation or test sample) can be used; in this case we use the same matrix.

The most important advantages of classification and/or regression trees are [28]: (a) structured knowledge is obtained in the form of classification rules or the values of a variable interval. This knowledge is easy to interpret and, in simple language, characterizes the classes or values of a variable interval; (b) as it is a nonparametric analysis (distribution-free procedure), it requires no distributional assumptions to validate probability; (c) it allows working with all types of predictor variables: binary, nominal, ordinal, and interval or ratio; (d) it allows unknown values for the predictor variables in the individuals, both in the construction phase and in the tree prediction; e) in the case of classification probability, it can be set to a priori classes; and (f) the observations can be weighed using an ad-hoc variable.

An expression known as recursive dividing is essential to the nonparametric statistical approach of classification and regression trees (CART) [27]. Supposing the data are given by $D = \{(X_i, Y_i),$

i = 1, 2, ..., n}, where Y_i are widths made on a uninterrupted response variable Y, and the X_i are measurements on an input r-vector X. We accept that Y is connected to X as in multiple regression, and the aim is to use a tree-based algorithm to predict Y from X.

Regression trees are built in a parallel way to classification trees, and the technique is generally stated as recursive-separating regression. In a classification tree, the class of a terminal knot is demarcated as the class that orders a plurality (generally in the two-class case) of all of the observations in that node, where ties are randomized. In a regression tree, the output is set to have the constant value $Y(\tau)$ at terminal node τ . Hence, the tree can be characterized as an r-dimensional histogram approximated of the regression surface, where r is the number of input variables, X_1, X_2, ..., X_r [29].

$$i(\tau) = \sum (Y_i - \bar{Y}_\tau)^2$$

where $\tilde{Y}_\tau$ is the average of Y_i for all annotations assigned to node τ.

To determine the type of split in any node we take, as our splitting strategy at node $\tau \in \check{T}$, the division that delivers the largest decrease in the value of i(T). The reduction in $i(\tau)$ due to a division into τ_L and τ_R is expressed by

$$\Delta i(\tau) = i(\tau) - i(\tau_L) - i(\tau_R)$$

The left daughter node and right daughter node emanating from a (parent) node τ are denoted by τ_L and τ_R, respectively.

The best division at τ is the one that exploits $\Delta i(\tau)$. The consequence of employing such a splitting approach is that the best division will split up observations according to whether Y has a small or large value; in general, where divisions occur, we can see either $y(\tau_L) < y(\tau) < y(\tau_R)$, or its opposite, with $y(\tau_L)$ and $y(\tau_R)$ interchanged.

We note that discovery τ_L and τ_R to exploit $\Delta i(\tau)$ is similar to reducing $i(\tau_L) + i(\tau_R)$. Solving:

$$\min \tau_L, \tau_R \{p(\tau_L)s^2(\tau_L) + p(\tau_R)s^2(\tau_R)\}$$

where $p(\tau_L)$ and $p(\tau_R)$ are the proportions of observations in τ that divide to τ_L and τ_R, individually [28].

The Salford Predictive Modeler Builder v6.6 (www.salford-systems.com) was used to generate these models.

2.6. Generalized Additive Model with Thin Plate Splines (GAM.TP)

A generalized additive model (GAM) is a generalized linear model in which the linear predictor be determined by linearly on unidentified smooth expressions of some variables, and interest focuses on inference about these smooth expressions. Additive models were originally built by [28] to combine properties of linear models with additive models.

The generalized additive model replaces

$$\sum \beta_j X_j$$

with

$$\sum f_j(x_j)$$

where f_j is an unspecified ('non-parametric') function. It can be in a non-linear form:

$$E(Y \mid X_1, \ldots, X_p) = f(X_1, \ldots, X_p) = f_0 + f_1(X_1) + \cdots + f_p(X_p).$$

The function $f_j(x_j)$ is estimated in a flexible manner using a spline smoother [30].

A smoother is an instrument for summarizing the tendency of a dependent variable Y as an expression of one, or more, independent variables $X_1, \ldots, X_p$. It generates an estimate of the tendency that is less mutable than Y itself; therefore, the name 'smoother'. The most significant characteristic of

a smoother is its non-parametric nature, so the smooth function is also known as a non-parametric function. Its greatest difference from the GLM is that it does not undertake an inflexible form for the dependence of Y on $X_1, \ldots, X_p$. It allows an approach with the addition of expressions (expressions that have separated input estimates), not just with one indefinite expression only. For this reason it is the building block of the generalized additive model algorithm [31].

Testing the different types of splines reveals that the best model helped with the AIC value is the additive model with thin plate regression splines. The thin plate spline is the two-dimensional equivalent of the cubic spline in one dimension. It is the essential resolution to the biharmonic equation, and has the form

$$U(r) = r^2 \ln(r).$$

Assuming a dataset of points, a weighted mixture of thin plate splines concentrated about each point gives the interpolation expression that passes through the points precisely while reducing the so-called 'bending energy.' Bending energy is defined here as the integral over R^2 of the squares of the second derivatives:

$$I[f(x, y)] = \int\int ((f_{xx})^2 + 2(f_{xy})^2 + (f_{yy})^2) dxdy.$$

Regularization should be used to decrease the necessity that the interpolant pass through the data points exactly.

The designation of 'thin plate spline' is a physical analogy referring to the flexibility of a thin sheet of metal. In the physical situation, the deflection is in the z direction, at a right angle to the plane. In order to apply this impression to the problem of coordinate conversion, the lifting of the plate is interpreted as a dislocation of the x or y coordinates within the plane [32].

These splines are short rank isotropic smoothers of any number of variables. The splines are isotropic because any variation of the covariate co-ordinate system will not modify the output of smoothing. The low rank means that they have rarer coefficients than there are data to smooth. They are the default smoother for 's' terms due to there being a clear logic in which they are the ideal smoother of any given basis measurement/rank [33].

In this case, as we are building the model with R we used the mgcv package [33–37] to construct the additive model and the ROCR package [38] to obtain the validations, AUC values, and graphics.

2.7. Evaluation

The area under the receiver operating characteristic (ROC) function (AUC) is taken to be an important index because it provides a single measure of overall accuracy that is independent upon a particular threshold [39]. If the objective is to rank the classifiers, comparisons using ROC plots are more robust since they are not dependent of the values in a confusion matrix [40]. An ROC graph is a method for visualizing, establishing, and selecting classifiers based on their presentation. ROC curve analysis was developed during World War II as a tool in signal processing, and is now used in many branches of science. Standard references for ROC curve analysis are [40–45].

Although ROC graphs are conceptually simple, their application in research contexts gives rise to some complexities that are not obvious and their practical use entails some common misconceptions and pitfalls [46].

Some formulae typical in ROC curves are

$$\text{tp rate} \approx (\text{Positives correctly classified})/(\text{Total positives})$$

$$\text{fp rate} \approx (\text{Negatives incorrectly classified})/(\text{Total negatives}).$$

ROC graphs are two-dimensional graphs where the true positive rate is presented on the Y axis and the false positive rate is presented on the X axis. An ROC graph represents relative adjustments between profits (true positives) and expenses (false positives). Figure 2 shows the area under two ROC curves, A and B. Classifier A has a greater area and, therefore, better average performance.

Finally, it is probable for a low-AUC classifier to perform better in a specific region of the ROC space than a high-AUC classifier. Figure 2 shows an example of this: classifier B is generally worse than A, except at an fp rate > 0.6 where B has an insignificant advantage. However, in practice the AUC performs very well and is often used when a general measure of predictiveness is desired.

In order to analyze suitability of the different models, we have used 10% of the data available for each species.

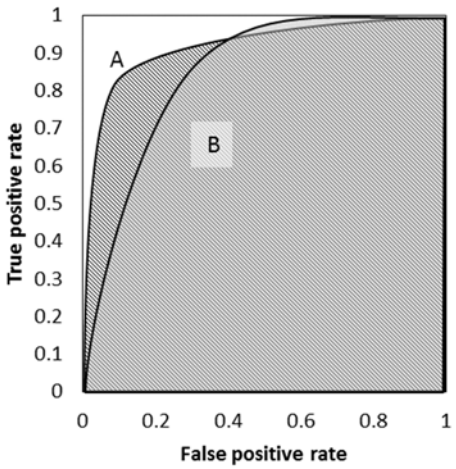


Figure 2. Example of two ROC curves and area under the curves (AUC).

3. Results

The study performed with the 17 species reveals some important results. For the sake of simplicity, we present here only the results summarizing the species using the two species with the highest and lowest results.

A two-way ANOVA shows that all the environmental variables included in the models are significant for all measures of performance ($p < 0.05$). All models designed have good predictions and obtain high AUC values.

Analyzing predictability, based on the AUC, we have obtained that for all the species analyzed, MARS and MaxEnt are the models with the lowest predictability and, consequently, with the lowest AUC average. However, CART and GAM generally have the highest AUC values.

Below are the AUC values obtained in the verification process for the species analyzed. The following ecological modeling methods are compared: MARS and MaxEnt, CART, and GAM.TP. The scatterplot graph shows the different models' behavior, demonstrating that all have good predictability based on their AUC values.

Figure 3 shows that MaxEnt, CART, and GAM.TP have AUC values near to 1, and that all of the models have good results for predicting species distribution. For the different species, all of the statistical models show similar behavior and performed in the same way. Comparing model predictability, the AUC values in GAM.TP have better results, on average, than the others. A comparison of one particular species with the highest AUC results and another with the lowest reveals no important differences between the models. As we can see in the table below, AUC values are very similar across all models.

In Table 1, we can see that the average for the best species modeled was 0.986 with a deviation value of 0.029. Moreover, one of the species with the lowest AUC value was *Pinus pinea* with an average of 0.876; in this case the deviation was 0.019. In both cases MARS obtained the lowest AUC values. In contrast, the best values were obtained with GAM.TP, although the few differences between this model and the others were not very significant. Finally, we can see Quercus ilex, a species with greater presence in the dataset, and also the species with the lowest AUC for all the models.

Figure 4a (*Quercus canariensis*) and 4b (*Pinus pinea*) show the results. Models are represented in different colors to facilitate understanding: MaxEnt (red), MARS (green), CART (blue), and GAM.TP (black). Figure 4a shows that every model has very good results, but GAM.TP exceeds the others. However, Figure 4b shows that MARS obtains worse results than the others and, for this reason, the line is well below the rest.

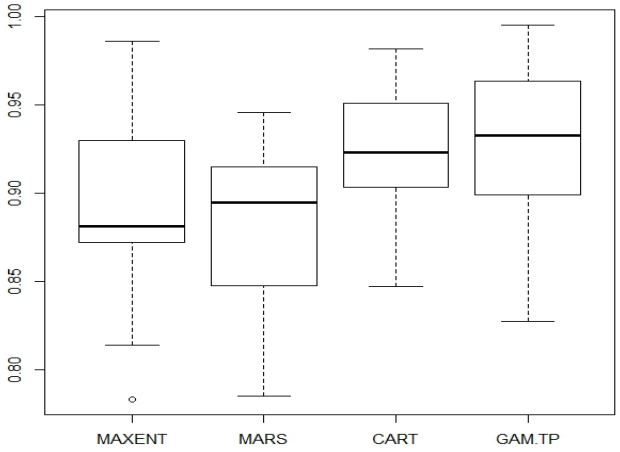
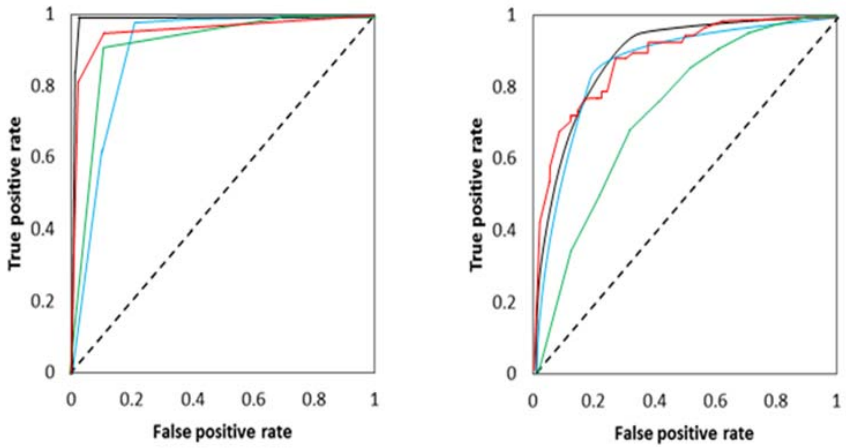


Figure 3. Average values of AUC for MaxEnt, MARS, CART, and GAM.TP models.

Table 1. Average of values of AUC for each model with *Quercus canariensis* and *Pinus pinea*.

Species	MAXENT	MARS	CART	GAM.TP
Q. canariensis	0.986	0.920	0.970	0.995
P. pinea	0.884	0.847	0.874	0.899
Q. ilex	0.783	0.814	0.847	0.834



(**a**) (**b**)

Figure 4. Curve ROC of different models with *Quercus canariensis* (**right**, **a**) and *Pinus pinea* (**left**, **b**). Models are represented in different colors to facilitate understanding: MaxEnt (red), MARS (green), CART (blue), and GAM.TP (black).

Analyzing the results, species with the highest number of presences have lower values in the predictions due to a wide range of the environmental variables. In contrast, the species with the most absences have the highest AUC values, perhaps due to the representative environmental characteristics that give rise to the presence of these species.

If we analyze presence-absence from the dataset and compare it with the AUC average, we find that the relationship between AUC and percentage of presence is negative (based on the correlation index), with a value of −0.75. Species with the highest percentage of presence have lower AUC values than other less-represented species in the area of study.

Figure 5 represents the relationship between the presence percentage and the AUC value, showing it to be negative; when the percentage of presence increases, the AUC value decreases. A simple trend line clearly shows the behavior between these values. The trend line is represented with the expression $y = -0.0038x + 0.9365$, and an R^2 value of 0.508.

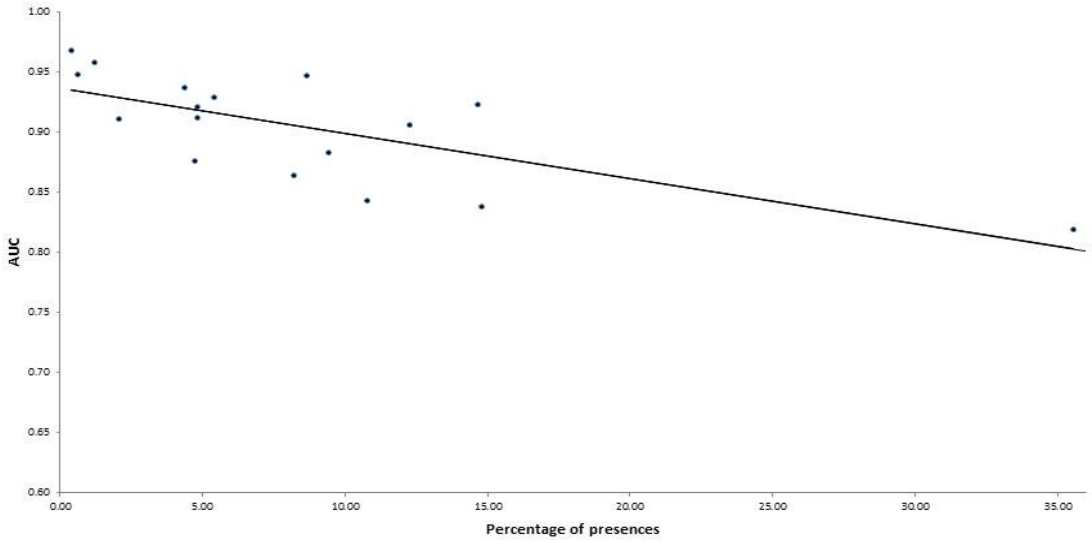


Figure 5. Dispersion graph with the percentage of the number of presences (X axis) and AUC values (Y axis) with regression.

In summary, every AUC value obtained with those models is significant and all the models could be useful to represent the distribution of each species. Overall, the additive model with thin plate splines gave the best results. MaxEnt, CART, and GAM.TP with thin plates splines obtained similar AUC values. The worst capability was obtained with MARS. This model's performance was below the average for several species. The models we developed obtained better results because they allowed for changes and calibrations. In this case we were aware of all of the processes that occurred during the modeling. By contrast, models obtained using specific software, in general, perform like "hermetic machines" because it could sometimes be impossible to understand the stages leading toward the final results.

4. Discussion

Our modeling framework examines how applicable some of the most widely used models are to species, the presences of which are largely set by the physical environment.

As we can see, all of the techniques developed here proved capable of successfully predicting species distribution. Factually, all of the models obtain similar performance based in the AUC and, over all, all of the methods show good results for predicting species distribution. Thus, there are no important differences between the different techniques developed in this analysis. Moreover, we can highlight interesting points based in our results to try to clarify and support our model selection:

GAM.TP performed better overall than MaxEnt and MARS, even though these differences are not substantial when compared with regression trees. This result differed from other comparative analyses [5,14,47,48], where linear models and additive models performed better than classification trees. We establish that, despite the dissimilarities in model suppositions, all statistical techniques seem to provide the best predictions for additive models.

As we said in the introduction, MaxEnt was included in this analysis due to the popularity of the method. We can understand that, in some points, it is not comparable to presence/absence models. Moreover, the similarity in results gave us some chances to compare the models. As we said, MaxEnt is presence-only data; for this reason the interpretation of the output is less clear than the models for presence/absence data. MaxEnt output must be taken as a relative guide of environmental suitability. On the other hand, presence/absence models could be more reliable because these models use information from the real absences.

From the user perspective, if we compare the different modeling techniques, CART and MARS required the least amount of user guidance (probably because they were developed in tools designed in a friendly environment). Moreover these tools are less flexible than the other statistical techniques developed in this research and, also, there is a high complexity if the user tries to manipulate the default settings in order to improve the accuracy of the outputs.

GAM.TP requires some knowledge of statistical techniques due to the amount of possibilities that it offers for building one's own model. Moreover, this flexibility and the possibilities offered by this approach make this approach more attractive. Currently, several tutorials available for several packages in R make the possibility to elaborate advanced statistical models more affordable without master knowledge (although we recommend a deeper research in statistical techniques before applying any model to avoid misunderstanding and frustration).

In conclusion, additive models with thin plate splines may be considered one of the greatest methods to analyze species distribution models working with presence-absence data, comparable to MaxEnt, CART, and MARS. Our results show a better fit and more flexibility in the design.

Looking at the quality of the data and the possibility to work with presence/absence values, and also with a systematic survey, we can confirm, looking our results, that the information obtained from the absences could be more important than the presences. Analyzing this result from an ecological perspective, absences deliver more information about the species due to the combination of several environmental predictors.

From an ecological perspective, analyzing the variables used in all of the models, we can see some differences between the variables' importance, depending of the model used. Comparing the species used before, we can see in Table 2 that different models have different variables' weight. In our case, MARS and CART have the same set of the most important variables for both species (SUP, MTCM, C, SR). Moreover, MaxEnt uses different variables for the different species: MTCM, C, R, and ETP with *Q. canariensis*, and ST, SUP, MTCM, and SR in *P. pinea* model. Finally, with GAM. TP, the most important variables were MTCM, SR, R, and WD in *Q. canariensis*, and in *P. pinea*, all of the variables have similar weight, but the more important four were R, SR, MTWM, and ST.

Table 2. Summary of the most important variables for each model with *Quercus canariensis* and *Pinus pinea*.

Species	**MAXENT**	**MARS**	**CART**	**GAM.TP**
Q. canariensis	MTCM	SUP	SUP	MTCM
	C	MTCM	MTCM	SR
	R	C	C	R
	ETP	SR	SR	WD
P. pinea	ST	SUP	SUP	R
	SUP	MTCM	MTCM	SR
	MTCM	C	C	MTWM
	SR	SR	SR	ST

As we have seen in the results (Table 1 and Figure 5), species that are less represented (i.e., with more absences), have better predictability than species with more presences. This situation shows us the importance of absences in predictive models. These absences give us several pieces of

information about the suitability of species and defining absence areas. If we analyze these models as management tools, this information is essential regarding the species selection and, in our case, for forest management.

Finally, we understand that there are more advanced approaches that can be applied in species distribution models, most of them through the Bayesian approach (i.e., R-INLA (Integrated nested Laplace approximation) can be compiled with the stochastic partial differential equations (SPDE) approach [49] which, through a discretization of a continuous Gaussian field, can cope efficiently with variables characterized by a complex spatial structure). However, our objective was to show the interesting opportunities that these explanatory techniques offer and to assess the relationships between environmental variables.

Author Contributions: Óscar Rodríguez de Rivera and Antonio López-Quílez conceived and designed the experiments; Óscar Rodríguez de Rivera performed the experiments; Óscar Rodríguez de Rivera analyzed the data; Óscar Rodríguez de Rivera and Antonio López-Quílez wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Elith, J.; Leathwick, J.R. Species distribution models: Ecological explanation and prediction across space and time. *Annu. Rev. Ecol. Evol. Syst.* **2009**, *40*, 677–697. [CrossRef]
2. Hijmans, R.J.; Elith, J. Species Distribution Modelling with R. The R Foundation for Statistical Computing, 2015. Available online: http://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf (accessed on 4 May 2016).
3. Austin, M.P. Spatial prediction of species distribution: An interface between ecological theory and statistical modeling. *Ecol. Model.* **2002**, *157*, 101–118. [CrossRef]
4. Muñoz, J.; Felicísimo, A.M. Comparison of statistical methods commonly used in predictive modeling. *J. Veg. Sci.* **2004**, *15*, 285–292. [CrossRef]
5. Segurado, P.; Araújo, M.B. An evaluation of methods for modeling species distributions. *J. Biogeogr.* **2004**, *31*, 1555–1568. [CrossRef]
6. Elith, J.; Graham, H.C.; Anderson, P.R.; Dudík, M.; Ferrier, S.; Guisan, A.; Hijmans, J.R.; Huettmann, F.; Leathwick, R.J.; Lehmann, A.; et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* **2006**, *29*, 129–151. [CrossRef]
7. Gibson, L.; Barrett, B.; Burbidge, A. Dealing with uncertain absences in habitat modeling: A case study of a rare ground-dwelling parrot. *Divers. Distrib.* **2007**, *13*, 704–713. [CrossRef]
8. Elith, J.; Graham, C.H. Do they? How do they? Why do they differ? On finding reasons for differing performances of species distribution models. *Ecography* **2009**, *32*, 66–77. [CrossRef]
9. Roura-Pascual, N.; Brotons, L.; Peterson, A.; Thuiller, W. Consensual predictions of potential distributional areas for invasive species: A case study of Argentine ants in the Iberian Peninsula. *Biol. Invasions* **2009**, *11*, 1017–1031. [CrossRef]
10. Tognelli, M.F.; Roig-Junent, S.A.; Marvaldi, A.E.; Flores, G.E.; Lobo, J.M. An evaluation of methods for modeling distribution of Patagonian insects. *Rev. Chil. Hist. Nat.* **2009**, *82*, 347–360. [CrossRef]
11. Marini, M.; Barbet-Massin, M.; Lopes, L.; Jiguet, F. Predicting the occurrence of rare Brazilian birds with species distribution models. *J. Ornithol.* **2010**, *151*, 857–866. [CrossRef]
12. Ferrier, S.; Watson, G.; Pearce, J.; Drielsma, M. Extended statistical approaches to modeling spatial pattern in biodiversity in northeast New South Wales. 1. Species-level modeling. *Biodivers. Conserv.* **2002**, *11*, 2275–2307. [CrossRef]
13. Elith, J.; Phillips, S.J.; Hastie, T.; Dudík, M.; Chee, Y.E.; Yates, C.J. A statistical explanation of MaxEnt for ecologists. *Divers. Distrib.* **2011**, *17*, 43–57. [CrossRef]
14. Meynard, C.N.; Quinn, J.F. Predicting species distributions: A critical comparison of the most common statistical models using artificial species. *J. Biogeogr.* **2007**, *34*, 1455–1469. [CrossRef]
15. Sánchez Palomares, O.; Sánchez Serrano, F.; Carretero, M.P. *Modelos y Cartografía de Estimaciones Climáticas Termopluviométricas Para la España Peninsular*; Instituto Nacional de Investigaciones Agrarias: Madrid, Spain, 1999; p. 192.

16. Farr, T.G.; Rosen, P.A.; Caro, E.; Crippen, R.; Duren, R.; Hensley, S.; Kobrick, M.; Paller, M.; Rodriguez, E.; Roth, L.; et al. The Shuttle Radar Topography Mission. *Rev. Geophys.* **2007**, *45*, RG2004. [CrossRef]
17. Alonso Ponce, R.; López Senespleda, E.; Sánchez Palomares, O. A novel application of the ecological field theory to the definition of physiographic and climatic potential areas of forest species. *Eur. J. For. Res.* **2010**, *129*, 119–131. [CrossRef]
18. Panagos, P.; Van Liedekerke, M.; Jones, A.; Montanarella, L. European Soil Data Centre: Response to European policy support and public data requirements. *Land Use Policy* **2012**, *29*, 329–338. [CrossRef]
19. Gastón, A.; Soriano, C.; Gómez-Miguel, V. Lithologic data improve plant species distribution models based on coarse-grained occurrence data. *For. Syst.* **2009**, *18*, 42–49.
20. Elith, J.; Burgman, M.A. Predictions and their validation: Rare plants in the Central Highlands, Victoria, Australia. In *Predicting Species Occurrences: Issues of Accuracy and Scale*; Scott, J.M., Heglund, P.J., Morrison, M.L., Raphael, M.G., Wall, W.A., Samson, F.B., Eds.; Island Press: Covelo, CA, USA, 2002; pp. 303–314.
21. Phillips, S.J.; Dudík, M. Modeling of species distributions with Maxent: New extensions and a comprehensive evaluation. *Ecography* **2008**, *31*, 161–175. [CrossRef]
22. Yee, T.W.; Mitchell, N.D. Generalized additive models in plant ecology. *J. Veg. Sci.* **1991**, *2*, 587–602. [CrossRef]
23. Ferrier, S.; Watson, G. *An Evaluation of the Effectiveness of Environmental Surrogates and Modelling Techniques in Predicting the Distribution of Biological Diversity*; Environment Australia: Canberra, Australia, 1997.
24. Ng, A.Y.; Jordan, M.I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Adv. Neural Inform. Process. Syst.* **2001**, *14*, 605–610.
25. Friedman, J.H. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*, 1–141. [CrossRef]
26. Nedjah, N.; Luiza de Macedo, M. *Fuzzy Systems Engineering: Theory and Practice*; Springer: New York, NY, USA, 2005.
27. Breiman, L.; Friedman, F.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth and Brooks: Pacific Grove, CA, USA, 1984.
28. Schiattino, I.; Silva, C. Árboles de Clasificación y Regresión: Modelos Cart. *Cienc. Trab.* **2008**, *10*, 161–166.
29. Izenman, A. *Modern Multivariate Statistical Techniques*; Springer: New York, NY, USA, 2008.
30. Hastie, T.; Tibshirani, R.J. *Generalized Additive Models*; Chapman & Hall/CRC Press: London, UK, 1990.
31. Liu, H. *Generalized Additive Model*; Department of Mathematics and Statistics University of Minnesota Duluth: Duluth, MN, USA, 2008.
32. Donato, G.; Belongie, S. Approximate Thin Plate Spline Mappings. In Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark, 28–31 May 2002; Springer: Copenhagen, Denmark, 2002; pp. 531–542.
33. Wood, S.N. Thin-plate regression splines. *J. R. Stat. Soc. B* **2003**, *65*, 95–114. [CrossRef]
34. Wood, S.N. Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Stat. Soc. B* **2000**, *62*, 413–428. [CrossRef]
35. Wood, S.N. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.* **2004**, *99*, 673–686. [CrossRef]
36. Wood, S.N. *Generalized Additive Models: An Introduction with R*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2006.
37. Wood, S.N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. B* **2011**, *73*, 3–36. [CrossRef]
38. Sing, T.; Sander, O.; Beerenwinkel, B.; Lenaguer, T. ROCR: Visualizing the Performance of Scoring Classifiers. R Package Version 1.0-4. 2012. Available online: http://CRAN.R-project.org/package=ROCR (accessed on 27 April 2015).
39. Deleo, J.M. Receiver operating characteristic laboratory (ROCLAB): Software for developing decision strategies that account for uncertainty. In Proceedings of the Second International Symposium on Uncertainty Modelling and Analysis, College Park, MD, USA, 17–20 March 1993; IEEE Computer Society Press: Washington, DC, USA, 1995; pp. 318–325.
40. Fielding, A.H.; Bell, J.F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.* **1997**, *24*, 38–49. [CrossRef]
41. Metz, C.E. Basic principles of ROC analysis. *Semin. Nucl. Med.* **1978**, *8*, 283–298. [CrossRef]

42. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef] [PubMed]
43. Murphy, A.H.; Winkler, R.L. Diagnostic verification of probability forecasts. *Int. J. Forecast.* **1992**, *7*, 435–455. [CrossRef]
44. Pearce, J.; Ferrier, S. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecol. Model.* **2000**, *133*, 225–245. [CrossRef]
45. Marzban, C. The ROC curve and the area under it as performance measures. *Weather Forecast.* **2004**, *19*, 1106–1114. [CrossRef]
46. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2005**, *27*, 861–874. [CrossRef]
47. Thuiller, W. BIOMOD—Optimizing predictions of species distributions and projecting potential future shifts under global change. *Glob. Chang. Biol.* **2003**, *9*, 1353–1362. [CrossRef]
48. Phillips, S.J.; Anderson, R.P.; Schapire, R.P. Maximum entropy modeling of species geographic distributions. *Ecol. Model.* **2006**, *190*, 231–259. [CrossRef]
49. Lindgren, F.; Rue, H.; Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. B* **2011**, *73*, 423–498. [CrossRef]