COMMUNITY ECOLOGY 10(2): 215-224, 2009
1585-8553/$20.00 © Akadémiai Kiadó, Budapest
DOI: 10.1556/ComEc.10.2009.2.12

Modelling the spatial distribution of tree species with fragmented populations from abundance data

L. Scarnati[1], F. Attorre[1,3], A. Farcomeni[2], F. Francesconi[1] and M. De Sanctis[1]

[1] *Department of Plant Biology, Sapienza University of Rome, P.le A. Moro, 5 - 00185 Rome, Italy*
[2] *Department of Experimental Medicine – Statistics Unit, Sapienza University of Rome, P.le A. Moro, 5 - 00185 Rome, Italy*
[3] *Corresponding author. E-mail: fabio.attorre@uniroma1.it*

Keywords: *Ilex aquifolium*, Gaussian processes with radial basis kernel functions, Multivariate adaptive regression splines, Potential areas, Random forest, Regression tree analysis, *Quercus suber*, Spatial modelling, Support vector regression, *Taxus baccata*.

Abstract: Spatial distribution modelling can be a useful tool for elaborating conservation strategies for tree species characterized by fragmented and sparse populations. We tested five statistical models—Support Vector Regression (SVR), Multivariate Adaptive Regression Splines (MARS), Gaussian processes with radial basis kernel functions (GP), Regression Tree Analysis (RTA) and Random Forests (RF)—for their predictive performances. To perform the evaluation, we applied these techniques to three tree species for which conservation measures should be elaborated and implemented: one Mediterranean species (*Quercus suber*) and two temperate species (*Ilex aquifolium* and *Taxus baccata*). Model evaluation was measured by MSE, Goodman-Kruskal and sensitivity statistics and map outputs based on the minimal predicted area criterion. All the models performed well, confirming the validity of this approach when dealing with species characterized by narrow and specialized niches and when adequate data (more than 40-50 samples) and environmental and climatic variables, recognized as important determinants of plant distribution patterns, are available. Based on the evaluation processes, RF resulted the most accurate algorithm thanks to bootstrap-resampling, trees averaging, randomization of predictors and smoother response surface.

Abbreviations: GP–Gaussian processes with radial basis kernel functions, IV–Importance Value, MARS–Multivariate Adaptive Regression Splines, RF–Random Forests, RTA–Regression Tree Analysis, SVR–Support Vector Regression.

Introduction

In the last few years a wide range of empirical models have been used to predict the distribution of tree species. They are now considered a useful tool for the conservation and management of forest habitats. Their use is based on the idea that known occurrences or measures of abundance of tree species can be related to environmental predictors using statistically derived response curves that aim to reflect the species' environmental responses (Guisan and Zimmerman 2000, Guisan and Thuiller 2005, Guisan et al. 2006). The fitted model is then used to project the niche into geographic space, providing a spatial prediction of the most suitable areas for a given species. Practical examples of this approach include the evaluation of the effect of climate change on the abundance and distribution of tree species (Iverson and Prasad 2002, Prasad et al. 2006, Attorre et al. 2007b, Benito Garzòn et al. 2008), the elaboration of management strategies (Hidalgo et al. 2008) and the assessment of the potential spreading areas of invasive alien species (Rouget et al. 2004). Many different methods are available for modelling the distribution tree species: models of environmental envelope such as ENFA (Hirzel et al. 2002) and BIOCLIM (Beaumont et al. 2005), classical regression models, such us generalized linear models and generalized additive models (Guisan et al. 2002, Lehmann et al. 2002), and machine-learning techniques that are able to deal with complex and non-linear relationships between predictors and response (Recknagel 2001). Among these latter methods are classification and regression trees (Iverson and Prasad 1998, Vayssieres et al. 2000), and its variants such as boosted regression trees (Elith et al. 2002) or random forest (Benito Garzòn et al. 2006), artificial neural networks (Pearson et al. 2002) and genetic algorithms (Peterson et al. 2001, 2002). Several studies, based on presence-only or presence-absence data, have been carried out to evaluate their performance (Elith et al. 2002, Pearson et al. 2002, Thuiller 2003, Segurado and Araujo 2004, Benito Garzòn et al. 2006, Hernandez et al. 2006, Tsoar et al. 2007). Based on the results of these studies we have compared the most promising methods to evaluate their efficiency in modelling abundance measures of tree species characterized, in the study area, by sparse and fragmented populations. The best method is then used to produce maps of the current potential distribution abundance of these species to support conservation strategies.

Study area

The study area is Central Italy and we have chosen *Ilex aquifolium*, *Taxus baccata* and *Quercus suber*. The first two

can be found mainly in the temperate beech woods of the Apennine and are relicts of the Cenozoic flora, characterized by warm-humid climatic conditions. They survived the glaciations of the Quaternary period in refugia areas, and may have followed *Fagus sylvatica* in the successive postglacial expansion. This process, possibly characterized by long-range dispersion events, determined their current fragmented presence and reduced consistency (Magri et al. 2006).

Quercus suber is an evergreen oak which grows in Mediterranean schlerophyllous forests. In central Italy, it grows only on the Tyrrhenian coast and is characterized by scattered populations from sea level up to 500 m, typically on siliceous acid soils. *Quercus suber* forests are linked to cork exploitation, however, the decline of this economic activity and the change of land use are causing their rapid decline.

Because of their high level of biodiversity, the two habitats have been included in Annex I of the Habitat Directive ("Beech forests of the Apennine with *Taxus* and *Ilex*" and "*Quercus suber* forests" - Council of Europe, 1992). For this reason, the production of reliable maps of potential spatial distribution of these species can be a valuable tool for the elaboration of conservation strategies.

Material and methods

Data set

Using bibliographical information and information provided by the staff of the protected areas, we identified all the locations of beech woods containing *Taxus* and *Ilex* and *Quercus suber* forests. In these areas, 65 sample plots for *Ilex*, 55 for *Taxus* and 85 for *Quercus suber* were carried out during the spring-summer period of 2007 for *Taxus* and *Ilex* and of 2008 for *Quercus suber*. Each plot had a radius of 15 meters and was identified by GPS coordinates (Fig. 1). In each plot, we measured, with a caliper, the diameter at breast height (1.30 m) of all the trees with a diameter equal to or over than 2.5 cm, using 5 cm diametrical classes.

As a measure of abundance the Importance Value (IV) was calculated according to the following formula:

Importance Value (x) = Density (x) + Dominance (x)
Density (x) = 100 * NS (x) / NS (all species)
Dominance (x) = 100 * BA (x) / BA (all species)

Where: x is one of the considered species, NS is the number of stems of a plot and BA is the basal area of the plot calculated using the diameter at breast height of each one of the stems. In monotypic stands, the IV could reach a maximum of 200.

For the environmental variables, we used climatic maps in GRID format with a spatial resolution of 500 m. These maps were obtained by interpolating precipitation and temperature data recorded in 300 meteorological stations and calculated as the average of the 1960 - 1990 period. Climatic variables were chosen among those believed to be more meaningful for their influence on the growth and distribution of tree species and considered representative of others more directly related to them, such as the number of growing degree days or actual evapotranspiration (Thuiller et al. 2003). We used:

Annual mean temperature (MeanT)
Minimum temperature of the coldest month (MinT)
Maximum temperature of the hottest month (MaxT)
Summer precipitation (PS)
Winter precipitation (PW)
Total annual precipitation (PTot)

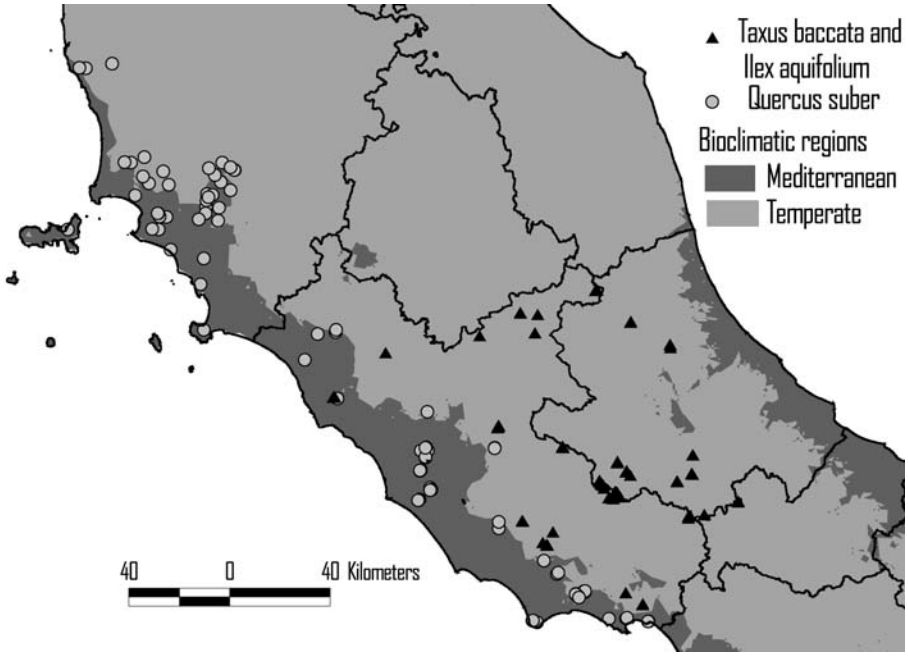

Figure 1. Sample plots and the study area in central Italy

Ilex symbol not seen!!.

The Universal kriging with external drift and covariates (altitude, slope, aspect, distance from coast, estimated solar radiation) was used as the interpolation method (Attorre et al. 2007a).

We also used slope (SLO) and geological maps, the latter as a surrogate of pedological information that was not available for the whole study area. The lithological substratum was measured and classified as either: sand, carbonatic, volcanic, arenaceous, clayey. In the statistical models, we parameterized the last variable using carbonatic as the baseline.

Statistical models

Various predictive models are tested in this study: Support Vector Regression (SVR), Multivariate Adaptive Regression Splines MARS), Gaussian processes with radial basis kernel functions (GP), Regression Tree Analysis (RTA) and Random Forests (RF). Data analysis was performed using the R statistical software (R Development Core Team 2008). R is a freeware that was developed by researchers who have also contributed novel statistical techniques in the form of packages that can be plugged into R. The predictive models tested are implemented in the following packages, available from the R website: *kernlab*, *earth*, *rpart*, *mgcv*, *randomForest*, *gbm*. The geographical analysis was performed with ArcGIS 9.2 (Environmental Systems Research Institute 2007).

We are aware of an important issue related to the specific formulation of the models we used. In fact, the specific parameterization choices we made are not the only ones available. For instance, one can use lowess or spline smoothers instead of kernels. The output obtained, within the same model, changes as one changes the details related to these possibly different parameterizations, yielding different outputs of species potential distribution (Thuiller 2003, Araújo and Guisan 2006). The existence of variability in a model outputs due to differences in its parameterization constitutes a source of uncertainty that could be evaluated by a sensitivity analysis, but this was clearly beyond the scope of this paper. We expect the variability between different models to be more substantial than the variability within each model with a different tuning. Furthermore, we have tried to set up the models with the most common and/or most intuitive parameterizations, in order to mimic a real situation of analysis for the prediction of species potential distribution.

Another issue is related to the fact that modelling techniques require records of absences since the use of presence-only data would bias the analysis and determine overoptimistic predictions of the potential distribution. However, in the case of rare tree species, absence data are difficult to obtain accurately: a given location may be classified in the ‘‘absence’’ set because for historical reasons the species is absent even though the habitat is suitable or because the habitat is truly unsuitable for the species; and only the latter cause is relevant for predictions. When no true absence data are available, one approach is to generate ‘pseudo-absences’ and to use them in the model as absence data for the species. There are many different methods for generating pseudo-absence (Zaniewski et al. 2002, Engler et al. 2004, Pearce and Boyce 2006). In this case, we have chosen random sampling without replacement, identifying a number of pseudo-absences equal to that of presences within known altitudinal ranges for the analysed species. The choice of the method for generating pseudo-absences may influence the performance of the models (see for instance Chefaoui and Lobo 2008). However, in this study simple random sampling has been chosen because in this way pseudo-absences can be regarded as a random sample from the background population, and the probabilistic properties of this random sample are known (s,ee for instance, Ward et al. 2009 on this issue). Further, the choice of the method is not relevant in our context, since it is unlikely to influence the comparative performance of the non-linear models we have used.

Support vector regression (SVR)

SVRs are considered to be a new generation of learning algorithms among machine-learning methods. SVR was proposed by Drucker et al. (1997) as a version of the Support Vector Machine for regression. SVR uses a functional relationship known as a kernel to map data onto a new hyperspace in which complicated patterns can be more simply represented (Müller et al. 2001). We used a Gaussian kernel, where with width considered as a hyperparameter and set as the estimated median width. As with any kernel method, in general the approach is almost insensitive to the choice of the kernel function, while it is sensitive to the choice of the bandwidth. Estimation of the width provides an easy and automatic choice of an operationally good width. See Hastie et al. (2001) for further discussion.

Vector Machines are gaining importance in modelling spatial distribution of species (Guo et al. 2005, Drake et al. 2006) because there is no theoretical requirement for observed data to be independent and they require less model tuning and fewer parameters than other more established methods.

Multivariate Adaptive Regression Splines (MARS)

MARS is similar to generalized additive models but it relies on fitting piecewise linear basis functions rather than smoothed functions (Leathwick et al. 2005) and includes a recursive simplification procedure. In other words, it fits linear segments, also called piecewise linear basis functions, to the data. It breaks the range of each predictor variable into subsets of the full range using ‘knots’ and allows the slope of the fitted linear segments between pairs of knots to vary while ensuring that the full fitted function is without breaks or sudden steps. The resulting estimate can be seen as the maximizer of a penalized least squares objective function. The Generalized Cross Validation (GCV) penalty per knot was set to 2 (Friedman 1991). MARS is commonly used in spatial ecology and in a recent comparison it performed particularly well for predicting occurrences in independent data sets (Elith et al. 2006).

Gaussian processes with radial basis kernel functions (GP)

This approach produces a so-called Radial Basis Functions network, whose main idea is to produce a smooth non-parametric estimate for the relationship between the response and the covariates (Williams and Barber 1998). GP gives more flexibility than a (global) linear regression approach, but not too much flexibility. At each point a different, but very simple, model is estimated. The resulting function is encouraged to be close to the observed point and to the neighbouring points, but it is also forced to be continuous and smooth. At each point a weighted average is produced, with weights being proportional to the kernel value when the kernel is centred at the point. The width for the kernel is considered as a hyperparameter and set as the estimated median width. We can expect predictions to be sensitive with respect to this choice, while being much less sensitive with respect to the kernel functional shape.

The GP approach is a promising tool in analyzing ecological data because this data is known to have a nonlinear structure, for example due to the presence of many zeros (zero-inflation). As far as we know, this is the first time it has been used in this research field.

Regression Tree Analysis (RTA)

RTA models have been widely used over the last few years to study the potential distribution of tree species abundance in the eastern United States (Iverson and Prasad 1998, 2002, Iverson et al. 1999). They are based on a recursive data partitioning algorithm that splits the data into subsets based on a single, best predictor variable. The algorithm proceeds by splitting these subsets using the remaining covariate values. The output is a tree with branches and terminal nodes. The predicted value at each terminal node is the average at that node, which can be considered as relatively homogeneous. In RTA the effect of the covariates is neither linear nor constant, producing a categorization of the (quantitative) covariate that best predict the observed response. Moreover, the hierarchical structure obtained provides a simple and effective way of understanding the covariates' impact on the observed response; the corresponding role in terms of response prediction can be based on the ordering produced by the increase in the percentage of response variability which is accounted for (Breiman et al. 1984). Our final RTA models were generated after pruning the full trees.

Random Forests (RF)

RF implements the automatic combination of tree predictors (Breiman 2001). In RF bootstrap samples are drawn to construct multiple trees and each tree is grown with a randomized subset of predictors. In our implementation we sampled 500 trees. This feature alleviates the problem of correlated variables because they may be extracted in turn, thus contributing to the aggregated tree model. Aggregation is obtained by averaging the trees. The RF algorithm also provides a measure of variable importance in the modelling: the importance is derived from the contribution of each variable accumulated along all nodes and all trees where it is used (Breiman 2001).

RF is receiving much attention in forecasting the effect of climate change on species distribution because growing a large number of small trees limits the generalization error. This means that it is very difficult to have over adaptation to the data, commonly known as overfitting (Prasad et al. 2006, Araujo and New 2007, Benito Garzòn et al. 2008).

Model validation

For all methods we carried out a *k*-fold cross validation, comparing the values observed and those predicted on the test set. We set k=10, but found no sensitivity to this choice for k ranging from 5 to 20. The cross validation was implemented as follows: first, the training data were randomly split into ten subsets of equal size, then each subset was in turn used for accuracy testing and the remaining nine for training. The procedure was repeated 1000 times. Finally, the total accuracy was estimated by averaging the accuracy of each test.

As suggested by Hernadez et al. (2006), in order to evaluate models we used multiple evaluation metrics accompanied with maps of suitable area predicted:

- The Mean Squared Error (MSE), which is the average of the squared differences between observed and predicted values. The MSE is commonly used in statistical literature to evaluate prediction performance.
- The average Goodman-Kruskal ordinal measure of association between the observed and predicted IV divided into classes according to the following breaks: 0.5, 3.5, 6.5, 10.5, 20.5, 30.5, 50.5. The Goodman-Kruskal index, like other non-parametric measures of association can complement MSE because it does not include an intrinsic hypothesis of symmetric loss**.**
- The sensitivity, which we defined as the proportion of presences correctly predicted among the observed presences. A predicted IV<0.5 was deemed as a predicted absence.

Due to the lack of true absence data, we preferred not to estimate the true-negative rate (the specificity, namely the proportion of true-negative predictions vs. the number of actual negative sites).

We also applied the criterion of the minimal predicted area (MPA) as defined by Engler et al. (2004) in order to compare the potential distribution maps obtained by each model. MPA is particularly useful when modelling the potential distribution of rare species by using presences and pseudo-absences. In fact, a model that predicts species presence everywhere could show the best evaluation (because all presences would then effectively be predicted as presences), but such a map would be useless. For this reason, according to the MPA, the best model based on presences and pseudo-

Table 1. Summary results for the 10-fold cross validation for *Ilex aquifolium*, *Taxus baccata* and *Quercus suber*.

	Min	1st Quartile	Mean	Median	3rd Quartile	Max	SD.
Ilex aquifolium							
				MSE			
SVR	3.17	115.05	443.09	310.69	616.16	2916.28	441.62
MARS	17.81	193.01	379.93	293.72	421.10	1928.85	342.80
RTA	3.06	147.55	433.16	301.08	562.41	2028.37	414.68
GP	30.40	170.42	357.14	278.96	426.77	2165.07	296.00
RF	15.13	118.09	292.96	197.65	305.12	1870.41	335.04
				Goodman-Kruskal			
SVR	0.13	0.67	0.75	0.77	0.85	1.00	0.14
MARS	0.00	0.63	0.72	0.74	0.82	1.00	0.15
RTA	0.00	0.61	0.72	0.74	0.84	1.00	0.17
GP	0.06	0.61	0.70	0.71	0.79	1.00	0.15
RF	0.13	0.69	0.77	0.78	0.86	1.00	0.13
				Sensitivity			
SVR	0.50	1.00	0.98	1.00	1.00	1.00	0.07
MARS	0.00	0.80	0.88	1.00	1.00	1.00	0.15
RTA	0.50	1.00	0.98	1.00	1.00	1.00	0.07
GP	0.60	1.00	1.00	1.00	1.00	1.00	0.02
RF	1.00	1.00	1.00	1.00	1.00	1.00	0.00
Taxus baccata							
				MSE			
SVR	0.52	93.36	279.70	229.77	407.30	1349.98	226.45
MARS	11.32	112.33	236.80	201.28	323.53	998.72	158.76
RTA	9.45	115.44	283.52	234.94	406.59	1228.55	204.68
GP	21.72	109.90	234.26	198.22	313.73	1076.19	159.68
RF	16.22	99.12	217.40	167.23	296.57	946.87	159.88
				Goodman-Kruskal			
SVR	0.17	0.63	0.72	0.72	0.81	1.00	0.13
MARS	0.18	0.65	0.73	0.73	0.82	1.00	0.12
RTA	0.00	0.61	0.70	0.71	0.81	1.00	0.15
GP	0.14	0.62	0.71	0.72	0.80	1.00	0.13
RF	0.28	0.67	0.75	0.76	0.83	1.00	0.12
				Sensitivity			
SVR	0.00	0.80	0.88	1.00	1.00	1.00	0.15
MARS	0.50	1.00	0.97	1.00	1.00	1.00	0.08
RTA	0.25	1.00	0.99	1.00	1.00	1.00	0.05
GP	0.50	1.00	0.98	1.00	1.00	1.00	0.06
RF	1.00	1.00	1.00	1.00	1.00	1.00	0.00
Quercus suber							
				MSE			
SVR	374.25	2291.07	3754.03	3537.15	4998.51	12712.22	1891.95
MARS	605.91	2348.78	3508.77	3308.29	4477.43	12213.25	1565.81
RTA	228.54	2814.41	4462.00	4242.89	5805.36	16749.04	2235.59
GP	646.74	2328.33	3357.73	3203.40	4196.69	10468.70	1378.64
RF	480.86	2072.59	3063.11	2837.29	3879.00	10183.71	1384.88
				Goodman-Kruskal			
SVR	0.40	0.69	0.76	0.77	0.85	1.00	0.12
MARS	0.39	0.67	0.76	0.77	0.85	1.00	0.12
RTA	0.00	0.67	0.76	0.77	0.86	1.00	0.13
GP	0.33	0.71	0.78	0.79	0.87	1.00	0.12
RF	0.40	0.73	0.80	0.82	0.89	1.00	0.11
				Sensitivity			
SVR	0.50	1.00	0.99	1.00	1.00	1.00	0.03
MARS	0.40	1.00	0.96	1.00	1.00	1.00	0.09
RTA	0.00	1.00	0.97	1.00	1.00	1.00	0.09
GP	1.00	1.00	1.00	1.00	1.00	1.00	0.00
RF	1.00	1.00	1.00	1.00	1.00	1.00	0.00

absences should predict the smallest possible potential area, while still covering a maximum number of the species occurrences. To calculate the MPA, for each species we then identified the threshold of predicted IV scores encompassing the 100% of the species occurrence (rule of parsimony) below which predictions were set as zero.

Variable importance

Variables predicted to be important in determining the spatial distribution of species were identified only for MARS, RTA and RF. In fact the SVR and GP do not directly provide such information. For RTA, the importance of a variable is measured as the total reduction in MSE achieved by all splits on that variable. In RF, the importance of a predictor is usually evaluated by randomly permuting its values. The RF is fitted on the original data set with the only difference given by the permuted predictor and the MSE is evaluated. The operation is repeated many times and variable importance is estimated as the mean difference in MSE between the model fit using the original data and using the permuted data. The mean difference is then normalized using the standard error so that the final importance measures can be used for ranking. For MARS, variable importance is calculated by refitting the model after dropping all terms involving the variable in question, calculating the reduction in goodness-of-fit and normalizing the results. In order to make a comparison between techniques, variable importance measures were all standardized.

Results

We compared the five techniques by assessing MSE, Goodman-Kruskal, Sensitivity and Minimal Predicted Area on each of the three species for each model. The three validation indexes clearly show that the best model is RF, which has the lowest MSE and the highest Goodman-Kruskal coefficient and, according to the sensitivity index, is the best in predicting the presence of the three species (Table 1). No marked differences were found among the other methods, even though GP was the second best six times in terms of performance. Based on the predictions, we quantified the MPA of the three species for the five models. According to the parsimony criterion for the MPA, the lowest predicted

Table 2. Minimal Predicted Area (km^2) for the tree species according to the five models.

	SVR	MARS	GP	RTA	RF
Ilex aquifolium	2512	6824	7287	3687	5992
Taxus baccata	1929	3684	3795	2909	3156
Quercus suber	14906	20001	21048	13534	20362

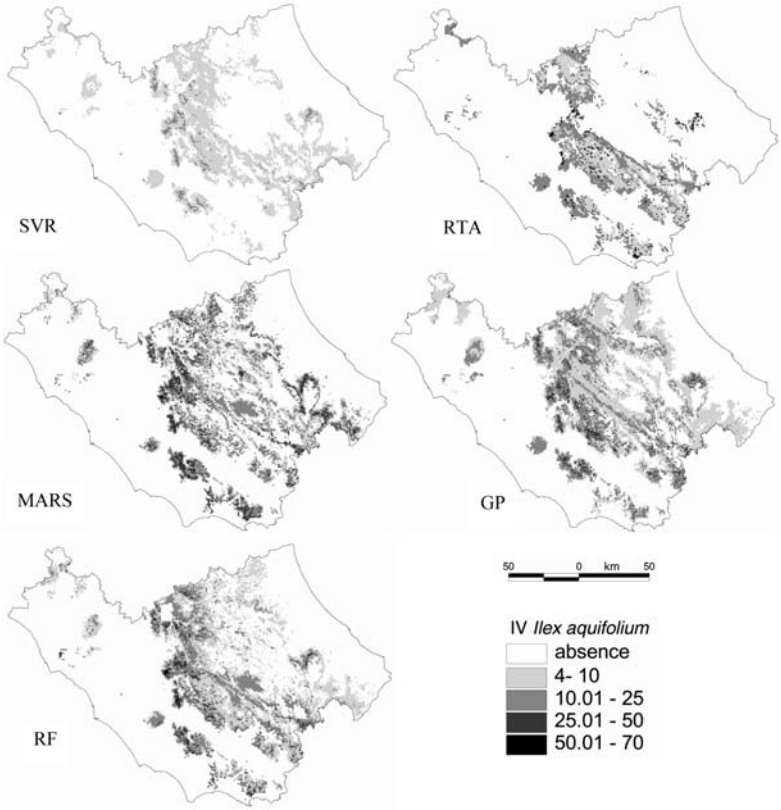

Figure 2. Current potential distribution of *Ilex aquifolium* abundance measured as Importance Value, according to the five models.

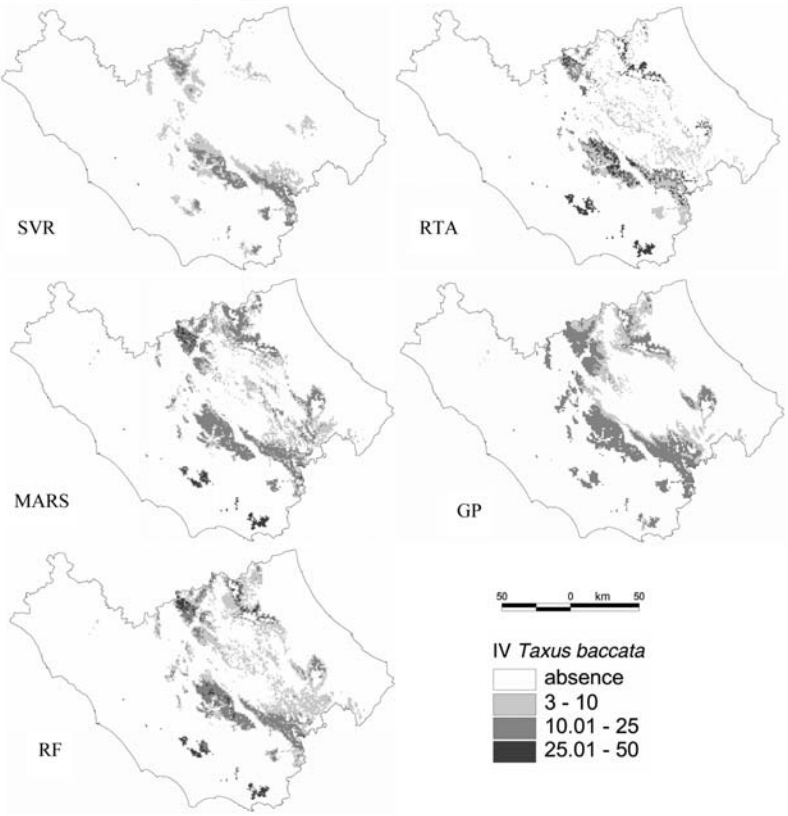

Figure 3. Current potential distribution of *Taxus baccata* abundance measured as Importance Value, according to the five models.

Table 3. Variable Importance predicted by the five models for *Ilex aquifolium, Taxus baccata* and *Quercus suber*.

Ilex aquifolium			
	MARS (Rank)	RTA (Rank)	RF (Rank)
Ptot	69 (2)	79 (3)	100 (1)
PS	4 (5)	100 (1)	60 (3)
PW	0	82 (2)	96 (2)
MeanT	0	31 (9)	31 (8)
MinT	18 (4)	72 (4)	50 (5)
MaxT	100 (1)	37 (6)	54 (6)
Slo	40 (3)	68 (5)	43 (7)
Sand	0	0	0
Vulcanic	0	32 (8)	1 (10)
Arenaceous	0	0	0
Carbonatic	3 (6)	0	0
Clayey	0	0	2 (9)
Taxus baccata			
Ptot	69 (2)	79 (3)	100 (1)
PS	4 (6)	100 (1)	33 (3)
PW	0	82 (2)	61 (2)
MeanT	0	31 (8)	12 (5)
MinT	18 (4)	72 (4)	11 (6)
MaxT	100 (1)	37 (6)	18 (4)
Slo	40 (5)	68 (5)	3 (7)
Sand	0	0	2 (8)
Vulcanic	0	32 (9)	0
Arenaceous	0	0	0
Carbonatic	3 (7)	0	0
Clayey	0	0	0
Quercus suber			
Ptot	41 (2)	88 (2)	73 (2)
PS	40 (3)	78 (3)	100 (1)
PW	0	100 (1)	72 (3)
MeanT	100 (1)	23 (6)	62 85)
MinT	0	43 (5)	67 (4)
MaxT	0	17 (6)	16 (7)
Slo	33 (4)	55 (4)	22 (6)
Sand	0	0	0
Vulcanic	0	10 (7)	7 (9)
Arenaceous	9 (5)	3 (8)	14 (8)
Carbonatic	0	0	0
Clayey	0	0	0

IVs were identified for each species: 4 for *Ilex*, 3 for *Taxus* and 6 for *Quercus suber*. Below these limits, predicted IVs were not used to produce the maps of potential distribution. Significant differences were found in the predicted suitability areas (Table 2). In particular SVR and RTA predicted areas noticeably smaller than the other three methods (GP, MARS and RF). The maps of the three species (Figs 2-4) reflected the validation measures of Table 1: SVR clearly under-predicts both the spatial distribution and IV values with respect to the other methods; RF tends to produce a larger and smoother predicted distribution map than RTA, while GP showed results comparable to that of a well established method such as MARS (Elith and Leathwick 2007). Through the visual assessment of the potential distribution maps, it was possible to notice that *Quercus suber* has its climatic optimum in the Mediterranean region but distant away from the coast, confirming the status of mesomediterranean species (Hidalgo et al. 2008). *Ilex aquifolium* and *Taxus baccata* characterized the Temperate region of the Apennines but the former shows a potential spatial distribution surrounding, at lower altitude, that of *Taxus* with a small overlapping. The two regression-tree methods (RTA and RF) identified precipitation variables as important for species distributions: annual, summer and winter precipitations have been ranked as the top three predictors (Table 3). MARS, instead, uses mainly mean and maximum temperatures to predict IV values. For all models the contribution of lithological variables has been proved to be minimal or not significant, probably because of a too coarse spatial resolution of data.

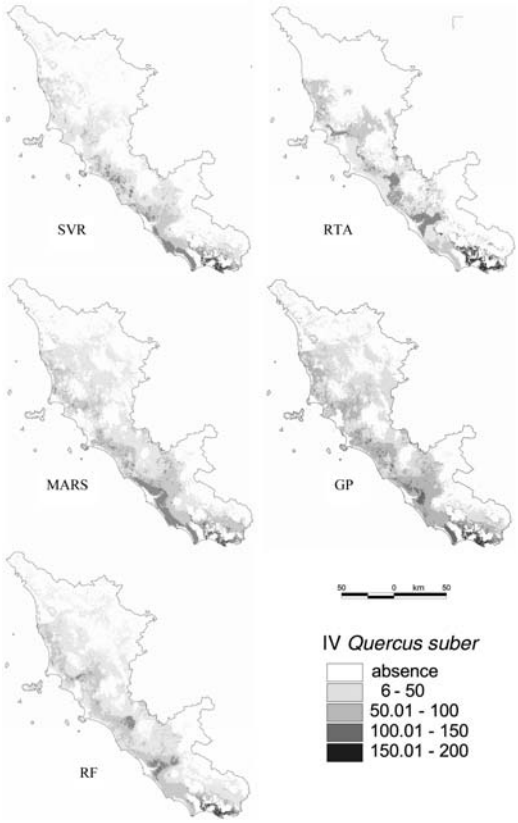

Figure 4. Current potential distribution of *Quercus suber* abundance measured as Importance Value, according to the five models.

Discussion

Over the last few years several studies have been conducted to analyze the performance of new spatial models in predicting tree species distribution: they used presence-absence data derived from the rasterization of forest maps (Benito Garzòn et al. 2006) and from forest inventories (Guisan et al. 2007) or abundance measures derived from forest inventories and averaged on GRID cells with a spatial resolution of 20 × 20 km (Prasad et al. 2006). Since many authors have highlighted that model performances depend on the ecological characteristics of the species, number of observations and spatial resolution of data (Drake et al. 2006, Elith et al. 2006, Guisan et al. 2006, Hernandez et al. 2006, Tsoar et al. 2007) we decided to compare several models in order to verify their efficiency in a specific case involving tree species characterized by fragmented and sparse population and which are the object of nature conservation policies such as the Habitat Directive. The potential distribution maps produced by the best model can then be used to support the elaboration of conservation strategies for these forest habitats together with other field data such as demography and interactions with soil parameters and other species (Guisan and Thuiller 2005). In spite of considerable differences in the complexity of the modelling algorithm, the five models examined in this study showed good performances with relatively small differences in predictive accuracy (Table 1). There are three reasons for these results: 1) all the chosen models are able to fit complex responses and select a relevant set of variables; 2) thanks to an extensive field campaign, conducted for all the species, it was possible to collect a number of sample plots greater than the critical threshold of 40-50, with which models should be trained (Farber and Kadmon 2003, Drake et al. 2006, Hernandez et al. 2006); 3) the three analyzed species are ecologically specialized and more easily modelled than generalist species with a wider environmental space (Segurado and Araujo 2004, Thuiller et al. 2004, Luoto et al. 2005, Elith et al. 2006, Hernandez et al. 2006).

By integrating the results obtained with the validation measures with the MPA and the visual examination of the output maps (Figs 2-4), it was possible to rank the methods according to their overall performance. In fact, thanks to bootstrap-resampling, tree averaging and randomization of predictors RF proved to be superior with respect to the other methods, confirming that it is one of the most promising methods in modelling the spatial distribution of species (Araujo and New 2007). It provides, also, a smoother response surface with no jumping-classes effect typical of RTA. Further, it necessitates almost no tuning. SVR, even though comparable to the other methods in terms of validation performance, produced smaller and more fragmented maps. GP and MARS performed similarly but the former showed a better overall efficiency in predicting species presences and IV values (Table 1). In conclusion, the application of spatial methods to produce maps of the potential distribution of abundance data has been confirmed to be a useful tool to support the elaboration of conservation strategies, especially in the case of ecologically specialized species characterized by fragmented and sparse populations. Obviously we stress that modelling should be integrated with detailed collection of field data, including data on species demography and biotic interactions, if it is to be fully useful for conservation purposes (Guisan and Thuiller 2005, Scarnati et al. 2009). Moreover, we believe that, besides the development and use of complex and suitable statistical tools, efforts should be made to measure factors which might potentially have a more direct influence on plant distribution and abundance, such as soil parameters, site history, disturbance, dispersal limitation, biotic interactions and human influences. These factors are usually ignored in distribution modelling because they are still difficult to measure in a spatially explicit way, but their incorporation could further bridge the gap between modellers and practitioners.

Acknowledgements: Work carried out in the framework of the Biodiversity Observatory of the Lazio Region (Italy) with the support of the Parks Lazio Agency. We are grateful to the staff of the Protected Areas who helped us during the field work for the sample areas identification. Finally, we would like to thank two anonymous reviewers whose comments greatly improved this paper.

References

Araújo, M. B. and A. Guisan. 2006. Five (or so) challenges for species distribution modeling. *J. Biogeogr.* 33: 1677-1688.

Araújo, M. B. and M. New. 2007. Ensemble forecasting of species distributions. *Trends Ecol. Evol.* 22 (1): 42-47.

Attorre, F., M. Alfò, M. De Sanctis, F. Francesconi and F. Bruno. 2007a. Comparison of interpolation methods for mapping climatic and bioclimatic variables at regional scale. International *J. Climatol*. 27: 1825-1843.

Attorre, F., F. Francesconi, N. Taleb, P. Scholte, A. Saed, M. Alfo and F. Bruno. 2007b. Will dragonblood survive the next period of climate change? Current and future potential distribution of *Dracaena cinnabari* (Socotra, Yemen). *Biol. Conserv.* 138: 430-439.

Beaumont, L.J., L. Hughes, and M. Poulsen. 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 Garzòn, M., R. Blazek, M. Neteler, R. Sánchez de Dios, H. Sainz Ollero and C. Furlanello. 2006. Machine learning models for predicting species habitat distribution suitability: An example with *Pinus sylvestris* L. for the Iberian Peninsula. *Ecol. Model.* 197: 383-393.

Benito Garzòn, M., R. Sánchez de Dios and H. Sainz Ollero. 2008. Effects of climate change on the distribution of Iberian tree species. *Appl. Veg. Sci.* 11: 169-178.

Breiman, L. 2001. Random forests. *Machine Learning* 45: 5-32.

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

Chefaoui, R.M. and J.M. Lobo. 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. *Ecol. Model.* 210: 478-486.

Drake, J.M., C. Randin and A. Guisan. 2006. Modelling ecological niches with support vector machines. *J. Appl. Ecol.* 43: 424-432.

Drucker, H., C.J.C. Burges, L. Kaufman, A. Smola and V. Vapnik. 1997. *Support Vector Regression Machines.* Advances in Neural Information Processing Systems 9, NIPS 1996, pp. 155-161.

Elith, J., M.A. Burgman and H.M. Regan. 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecol. Model.* 157: 313-330.

Elith, J., C. H. Graham, R. P. Anderson, M. Dudýk, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J.R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. McC. Overton, A. Townsend Peterson, S. J. Phillips, K. Richardson, R. Scachetti-Pereira, R. E. Schapire, J. Soberòn, S. Williams, M. S. Wisz and N.E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography* 29: 129-151.

Elith, J. and J. Leathwick. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Divers. Distrib.* 13: 265-275.

Engler, R., A. Guisan and L. Rechsteiner. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *J. Appl. Ecol.* 41: 263-274.

Farber, O. and R. Kadmon. 2003. Assessment of alternative approaches for bioclimatic modelling with special emphasis on the Mahalanobis distance. *Ecol. Model.* 160: 115-130.

Friedman, J. 1991. Multivariate adaptive regression splines. *Ann. Stat.* 19: 1-141

Guisan, A. and N. E. Zimmerman. 2000. Predictive habitat distribution models in ecology. *Ecol. Model.* 135: 147-186.

Guisan A., T.C. Edwards and T. Hastie. 2002. Generalized linear and generalized additive models in studies of species distribution: setting the scene. *Ecol. Model.* 157: 89-100.

Guisan, A. and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. *Ecol. Lett.* 8: 993-1009.

Guisan, A., O. Broennimann, R. Engler, M. Vust, N.G. Yoccoz, A. Lehmann and N.E. Zimmermann. 2006. Using niche-based models to improve the sampling of rare species. *Conserv. Biol.* 20: 501–511.

Guisan, A., N. E. Zimmermann, J. Elith, C. H. Graham, S. Phillips and A. T. Peterson. 2007. What matters for predicting the occurrences of trees: techniques, data or species characteristics? *Ecol. Monogr.* 77: 615–630.

Guo, Q., M. Kelly and C. H. Graham. 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. *Ecol. Model.* 182: 75-90.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning.* Springer, New York.

Hernandez, P. A., C. H. Graham, L. L. Master and D. L. Albert. 2006. The effect of sample size and species characteristics on performance of different species distribution modelling methods. *Ecography* 29: 773-785.

Hidalgo, P.J., M.J. Marìn, J. Quiijada and J.M. Moreira. 2008. A spatial distribution model of cork oak (*Quercus suber*) in southwestern Spain: a suitable tool for reforestation. *Forest Ecol. Manage.* 255: 25-34.

Hirzel, A. H., J. Hausser, D. Chessel and N. Perrin. 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? *Ecology* 83: 2027-2036.

Iverson, L.R. and A.M. Prasad. 1998. Predicting abundance of 80 tree species following climate change in the Eastern United States. *Ecol. Monogr.* 68: 465-485.

Iverson, L.R. and A.M. Prasad. 2002. Potential redistribution of tree species habitat under five climate change scenarios in the Eastern United States. *Forest Ecol. Manage.* 155: 205-222.

Iverson, L.R., A.M. Prasad and M.K. Schwartz. 1999. Modelling potential future individual tree species distributions in the Eastern United States under a climate change scenario: a case study with *Pinus virginiana*. *Ecol. Model.* 115: 77-93.

Leathwick, J. R., D. Rowe, J. Richardson, J. Elith and T. Hastie. 2005. Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biol.* 50: 2034-2052.

Lehmann, A., J. M. Overton and M. P. Austin. 2002. Regression models for spatial prediction: their role for biodiversity and conservation. *Biodivers. Conserv.* 11: 2085-2092.

Luoto, M., J. Pöyry, R. K. Heikkinen and K. Saarinen. 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecol. Biogeogr.* 14: 575-584.

Magri, D., G. G. Vendramin, B. Comps, I. Dupanloup, T. Geburek, D. Gomory, M. Latalowa, T. Litt, L. Paule, J. M. Roure, I. Tantau, W. O. Van der Knaap, R. J. Petit and J. L. De Beaulieu. 2006. A new scenario for the Quaternary history of European beech populations: palaeobotanical evidence and genetic consequences. *New Phytol.* 171: 199-221.

Müller, K. R., S. Mika, G. Rätsch, K. Tsuda. and B. Schölkopf. 2001. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks* 12: 181-202.

Pearce, J. and M. Boyce. 2006. Modelling distribution and abundance with presence-only data. *J. Appl. Ecol.* 43: 405-412.

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

Peterson, A.T., V. Sanchez-Cordero, J. Soberòn, J. Bartley, R. W. Buddemeier and A. G. Navarro-Sigüenza. 2001. Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecol. Model.* 144: 21-30.

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

Prasad, A. M., L. R. Iverson and A. Liaw. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9: 181-199.

Recknagel, F. 2001. Applications of machine learning to ecological modelling. *Ecol. Model.* 146: 303-310.

Rouget, M., D. M. Richardson, J. L. Nel, D. C. Le Maitre, B. Egoh and T. Mgidi. 2004. Mapping the potential ranges of major plant invaders in South Africa, Lesotho and Swaziland using climatic suitability. *Divers. Distrib.* 10: 475-484.

Scarnati, L., F. Attorre, M. De Sanctis, A. Farcomeni, F. Francesconi, M. Mancini and F. Bruno. 2009. A multiple approach for the evaluation of the spatial distribution and dynamics of a forest habitat: the case of Apennine beech forests with *Taxus baccata* and *Ilex aquifolium*. *Biodivers. Conserv.* Doi: 10.1007/s10531-009-9629-z

Segurado, P. and M. B. Araujo. 2004. An evaluation of methods for modelling species distributions. *J. Biogeogr.* 31: 1555-1568.

Thuiller, W. 2003. BIOMOD – Optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biol.* 9: 1353-1362.

Thuiller, W., J. Vayreda, J. Pino, S. Sabate, S. Lavorel and C. Gracia. 2003. Large-scale environmental correlates of forest tree distributions in Catalonia (NE Spain). *Global Ecol. Biogeogr.* 12: 313-325.

Thuiller, W., S. Lavorel, G.F. Midgley, S. Lavergne and A.G. Rebelo. 2004. Relating plant traits and species distributions along bioclimatic gradients for 88 *Leucadendron* species in the Cape Floristic Region. *Ecology* 85: 1688-1699.

Tsoar, A., O. Allouche, O. Steinitz, D. Rotem and R. Kadmon. 2007. A comparative evaluation of presence only methods for modelling species distribution. *Divers. Distrib.* 13: 397-405.

Vayssieres, M.P., R.E. Richard and B.H. Allen-Diaz. 2000. Classification trees: an alternative non-parametric approach for predicting species distribution. *J. Veg. Sci.* 11: 679-694.

Ward, G., T. Hastie, S. Barry, J. Elith, and J. Leathwick. 2009. Presence-only data and the EM algorithm. *Biometrics* 65: 554-563.

Williams, C. K. I. and D. Barber. 1998: Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20: 1342-1351.

Zaniewski, A.E., A. Lehmann and J. Overton. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecol. Model.* 157: 261-280.

Received February 23, 2009
Revised September 15, 2009
Accepted October 12, 2009