# Predicting Plant Species Distribution Across an Alpine Rangeland in Northern Spain. A Comparison of Probabilistic Methods

J. Bedia[1,2*], J. Busqué[1] and J.M. Gutiérrez[2]

[1] *Centro de Investigación y Formación Agraria, Gob. de Cantabria (CIFA), 39600 Muriedas, Spain*

[2] *Instituto de Física de Cantabria (CSIC-UC), 39005 Santander, Spain*

[*] Corresponding author. Email: joaquin.bedia@unican.es

## ABSTRACT

Predictive models constitute an important tool in ecology. Using presence/absence data of 15 plant species of an alpine rangeland in northern Spain, and a set of 14 topographical and geomorphological descriptors of relatively easy acquisition, we examined and compared the performance of five state-of-the-art methods used in ecological modeling: Multiple Logistic Regression (MLR), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Classification and Regression Trees (CART) and Multivariate Adaptive Regression Splines (MARS). Validation of the models was carried-out computing the Area Under the ROC Curve (AUC) using leave-one-out cross validation and the resolution and reliability diagrams of the resulting probabilistic predictions. We also analyzed the binary presence/absence deterministic predictions obtained by setting two different probability thresholds: the species prevalence and a ROC-optimized value, and we computed the corresponding confusion matrices to calculate sensitivity, specificity, Cohen's kappa and the True Skill Statistic (TSS). The overall result of this comparison shows that the performance of each technique varies depending on the target species; in general, CART exhibited a poor performance and MLR was competitive with the more sophisticated ANN, MARS and SVM methods. The best predictive resolution was obtained in most cases by ANN followed by SVM and CART models; on the other hand, MLR and MARS were generally the best calibrated. We also present an ecological interpretation of results, with emphasis in the possible ways of improving our models. Most of the target species were accurately predicted evidencing that geomorphological and topographical variables are suitable descriptors at the scale of analysis.



*Key words: Multiple Logistic Regression, Artificial neural Networks, Support Vector Machines, Classification and Regression Tree, Multivariate Adaptive Regression Splines, plant prediction, alpine vegetation*

# 1. Introduction

Species' distribution models (Fielding and Bell 1997; Guisan and Zimmermann 2000; Guisan and Thuiller 2005; Elith and Leathwick 2009) constitute a very important tool in ecology and conservation biology for a number of theoretical and practical issues, such as ecological niche modeling (Drake et al. 2006), assessment of potential species' distributions (Garzón et al. 2006; Drake and Bossenbroek 2009), prediction of future biotic responses to global change (Thuiller 2003), nature reserve selection (Araújo et al. 2004) or wildlife management (Gude et al. 2009) among others. These models share a common general approach: they statistically relate the spatial distribution of certain species (either the presence/absence or the abundance) with a set of environmental descriptors acting as input variables of the model. It has been shown that the different techniques used may idiosyncratically differ in their performance across species (Thuiller 2003; Elith et al. 2006) and therefore, for each particular application, it is desirable to assess the performance of the different state-of-the-art models rather than sticking to a single modeling technique.

Traditional methods for predictive modeling, such as generalized linear or additive models (e.g. logistic regression), are parametric models based on *a priori* assumptions on the shape of the response of species to environmental factors. This approach may be too simplistic as species often exhibit varied and complex responses to environmental gradients (Oksanen and Minchin 2002). In this context, higher order interaction terms need to be included to deal with skewed or non unimodal response shapes, often leading to spurious and biologically unfeasible responses difficult to interpret ecologically (Guisan and Zimmermann 2000).

To circumvent these shortcomings, in more recent times a number of particular non-

parametric approaches have been introduced in the literature that do not make any previous assumption on the shape of species' responses to environmental predictors (see, e.g., the non-parametric multiplicative regression; McCune 2006). Some general purpose non-parametric models, such as Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) or Multivariate Adaptive Regression Splines (MARS) have been successfully applied to a broad class of prediction problems (Hastie et al. 2010), including ecological and biogeographical questions such as species distributions. These advances provide a broad spectrum of algorithms available in different statistical packages to be applied for deterministic or probabilistic species distribution modeling.

In this work, we focus on the prediction of the presence/absence of a set of 15 plant species and analyze five different prediction methods spanning the above categories, from the simplest Multiple Logistic Regression (MLR) to the recent and complex SVMs, including also ANNs, Classification and Regression Trees (CART; Breiman et al. (1984)) and Multivariate Adaptive Regression Splines (MARS; Friedman (1991)). MLR is a parametric technique belonging to the family of the generalized linear models (GLMs). It has been widely used in species distribution modeling (see Guisan et al. 2002, for a description and a review of their application in ecology), whereas the others are non parametric methods which have gained popularity among ecologists in more recent times. The predictive performance of MARS has been studied in comparison with GAM models for several freshwater fish species by Leathwick et al. (2006), revealing its capability to effectively identify the most parsimonious set of environmental predictors and robustly describe the distribution of species. Similarly, MARS showed better performance than MLR in a comparative study by Muñoz and Felicísimo (2004). CART models have revealed better performance than MLR for predicting the dis-

tribution of three major oak species in California (Vayssieres et al. 2000) and offer valuable properties such as their flexibility, ease of implementation and interpretation of results, producing a feature space partition fully described by a single tree (Hastie et al. 2010). However, it has been shown their poorer performance against GAMs and high order MLR models in the case of simulated species data whose dominant predictor had a linear response (Santika and Hutchinson 2009). ANNs constitute highly flexible function approximators for any kind of data, able to cope with non-linear structures, making them a potential tool for ecological modeling (Lek et al. 1996; Lek and Guegan 1999). ANNs outperformed linear regression in predicting trout abundance in mountain streams (Lek et al. 1996), but obtained similar results than MLR and discriminant analysis for predicting presence/absence of a river bird species in the Himalayas (Manel et al. 1999), although this was assessed computing confusion matrix-derived measures which may introduce problems associated with threshold effects (Fielding and Bell 1997). Finally, SVMs are a recently developed supervised learning technique used for regression and classification, as well as density estimation. They are considered as universal and powerful as ANNs (Cortes and Vapnik 1995) and, conceptually, they can be assimilated to the classical definition by Hutchinson (1957) of ecological niche as a multidimensional environmental space (Drake et al. 2006). Although their application in modeling species distributions is still infrequent, it has proved a useful tool for habitat niche definition in forecasting biological invasion by Zebra Mussels in North American freshwater bodies (Drake and Bossenbroek 2009) as a relevant example.

The aim of this paper is to compare the performance of these five predictive methods (MLR, ANN, SVM, CART and MARS) using the same presence/absence data set of 15 characteristic plant species of acidic alpine rangelands in Northern Spain. Our assessment

of model performance is based on several validation scores (both probabilistic and binary derived from the confusion matrix) and qualitative diagrams (calibration and resolution), revealing the advantages and shortcomings of the different techniques. The paper is organized as follows: In Sec. 2 we present the methodology, introducing the study area and the data used, the models to be considered and the evaluation and validation procedures undertaken. In Sec. 3 we present the results, focusing on the model performance and the resulting species distribution models. Finally, in Sec. 4 we present a discussion of the work from the point of view of its ecological significance, methodological aspects to be considered and possible ways of improvement of the predictive performance.

# 2. Method

*a. Study area and data collection*

The study site, *Riofrio* rangeland, is located in Cantabrian Range (Northern Spain) and covers an area of 570 hectares. The altitude ranges from ca. 1700 m.a.s.l to a maximum of 2536 m. It comprises all major vegetation zones above the tree-line of the silicious Cantabrian Range. It holds the SPA status (Special Protection Area) linked to the Natura 2000 Network of the European Union. During summers of 2007 and 2008 (June-September) we conducted an exhaustive survey resulting in a detailed vegetation map of the whole study area. Each homogeneous patch of vegetation with an extension larger than 100 $m^2$ was delimited by a polygon. At each polygon, we quantified the abundance of 15 predefined species in terms of % cover (see Table 2). These species were chosen because they play a major role in

the configuration of vegetation at the landscape scale, with an estimated summed cover representing more than 95% of total study area including bare soil and rocky outcrops. One of them, *Festuca-Agrostis*, is not a single species but a distinctive grassland type dominated by *Festuca* gr. *rubra* and *Agrostis capillaris*. For convenience, we will collectively refer to them simply as "species" hereafter. After discarding all polygons with very scarce or null plant cover (cliffs, lakes and screes), a final dataset of 415 polygons covering an area of 424.6 hectares was retained. Finally, abundance data were transformed into presence/absence data, considering any cover value $> 0$ as presence of the species (see Fig. 1 for the distribution of four illustrative species).

For each polygon, mean values of four topographical variables (altitude, slope, solar radiation and terrain convexity) were calculated from a 2 m resolution digital elevation model. Aspect was not used as environmental variable as it was highly correlated with solar radiation. We also used 10 geomorphological variables, extracted from a geomorphological map that we constructed based on field observations and interpretation of aerial photographs (Table 3).

*b. Models for data analysis*

In this section we briefly describe the mathematical formulation of the five modeling techniques used in the paper, and indicate the software used to fit the different models to the data described in the previous section.

All analyses were conducted in the R language and environment for statistical computing (R Development Core Team 2009). For the stepwise selection of variables we used the

function 'step'. Additionally, we used the packages "ROCR" (Sing et al. 2009) for the computation of AUC values and confusion matrices and "maptools" (Lewin-Koh et al. 2009) for map representation.

MLR (also called *logit model*) is a generalized linear model used for binomial regression which is commonly used as a benchmark to predict the probability of occurrence of an event $y$ (in our example species occurrences) by fitting data to the following formula:

$$y = f(\sum_{k=1}^{m} \alpha_k x_k), \tag{1}$$

where $\mathbf{X} = \{X_1, \ldots, X_m\}$ is a set of predictor variables (the environmental descriptors in Table 3 in our example) and $f(z) = 1/(1 + exp(-z))$ is the logistic, or sigmoid, function with output constrained to $(0, 1)$. The unknown parameters $\alpha_k$ are usually estimated by maximum likelihood leading to a simple optimization problem that can also be seen as a simple version of an ANN without intermediate hidden layers between the input and the output. MLR models were fitted using the iteratively reweighted least squares procedure implemented in R (R Development Core Team 2009).

ANNs are machine learning models inspired by the functioning of the brain (Hastie et al. 2010). Thus, an ANN is formed by an input layer, $\{X_1, \cdots, X_m\}$ an output layer with one, $Y$, or several variables $\{Y_1, \cdots, Y_n\}$ and an predefined number of hidden layers, connected to each other. In the most popular configuration (feedforward networks), each node in the hidden and output layers receives input from all the nodes in the preceding layer and computes an output as the sigmoid-filtered weighted sum of inputs. Thus, a feedforward network with a single hidden layer, as the one implemented in this study, computes :

$$y = f(\sum_{j} \beta_{ji} f(\sum_{k} \alpha_{ik} x_k)) \tag{2}$$

where $f$ is again the sigmoid function and $\alpha_{ik}$ and $\beta_{ji}$ are the parameters to be fitted to data. In this case, the optimization of the error function leads to a complex nonlinear problem which needs to be solved with particular algorithms (in our case, we used the R package "AMORE" by Castejón et al. (2007)). Note that, as mentioned before, when considering no hidden layer, (2) reduces to (1).

MARS is a nonparametric method for regression analysis developed in the early 90s by Friedman (1991) which allows approximating the underlying function through a set of adaptive piecewise linear regressions called "basis functions" in the following form:

$$y = \alpha_o + \sum_{k=1}^{K} \alpha_k b_k(\mathbf{x}), \tag{3}$$

where the slope of each piecewise $b_k(\mathbf{x})$ can change in a set of points $\mathbf{Z}_{ki} = \mathbf{z}_{ki}$, $i = 1, \ldots, m$ with $\mathbf{Z}_{ki} \subset \mathbf{X}$, called knots. The popularity of this technique is due to the efficient optimization procedure used for the iterative search for basis functions and knots. In this work, we used the implementation of MARS in the R package "mda" (Leisch et al. 2009).

Similarly, CART is based on classification trees formed by a collection of rules based on values of certain variables in the modeling data set. These rules define branches of the tree which are optimized following an efficient search process. The advantage of this method is the intuitive representation of the knowledge (i.e. the set of rules) but, on the other hand, it lacks a compact model representation. For CART implementation we used the R package "tree" (Ripley 2009).

Finally, SVMs are recently developed machine learning methods used for classification and regression (Schlkopf and Smola 2001). This technique maps the input vectors to a higher dimensional space where a maximal separating hyperplane is constructed by considering a

epsilon-insensitive metric, where the (absolute) error values lower than epsilon are mapped to zero.

The approximating function can be defined as:

$$y =< w; x > +b \tag{4}$$

where $<;>$ denotes the dot product (for the linear case) or a kernel function such as the gaussian kernel in the general case of non-linear classifiers. The parameters are obtained from data by solving the following optimization problem:

$$\text{minimize} \qquad \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{5}$$

$$\text{subject to} \qquad \begin{cases} y_i - < w; x_i > \leq \epsilon + \xi_i \\ < w; x_i > +b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \tag{6}$$

For SVM construction we used the R implementation in package "e1071" (Dimitriadou et al. 2009).

*c. Evaluation of the models: Scores*

There is no single score to measure the quality of a predictive model, and different indices provide different aspects of the relationships between observed and predicted values; the situation is even more complex in the case of probabilistic forecasts (see, e.g. Jolliffe and Stephenson 2003). Quantitative probabilistic predictions have several advantages over deterministic presence/absence ones, since they provide an ecologically relevant information introducing a notion of habitat suitability that can be projected in the geographical space

(Guisan and Thuiller 2005). For management applications, probabilistic measures provide more flexibility to final model users, who can set different probability thresholds best suited to their particular aims (Freeman and Moisen 2008; Gude et al. 2009). From the point of view of critical model testing, they can be used to generate accuracy measures that are independent from the species prevalence allowing the general interpretation and comparison of different models (Vaughan and Ormerod 2005).

In the case of binary deterministic predictions, there are two types of possible errors: false positives (FP, error type I), when the model predicts a positive case when it is actually a negative one and false negatives (FN, error type II) when, on the contrary, the model fails to predict a positive case. These values are usually arranged in a table that summarizes model performance, including also correctly predicted positives and negatives (true positives TP, and true negatives TN) known as *confusion matrix* (Fielding and Bell 1997). From them, a number of measures of classification accuracy can be derived: *Sensitivity* is the probability of a given case to be correctly classified, whereas *specificity* is the inverse of *Sensitivity*. Both measures are independent of each other when compared across models and are also independent of species prevalence (Allouche et al. 2006).

In the case of probabilistic predictions, the ROC (Receiver operating characteristics) curve is commonly used as a generalization of the above validation procedure to describe the accuracy of the model (Fig. 2). This curve is defined by plotting the $sensitivity(u)$ vs. $1 - specificity(u)$ values for the deterministic prediction given by a probability threshold $u$. Probabilities above/below this threshold are set to positive/negative (presence/absence). By varying the probability threshold, the system becomes either more conservative or more "adventurous" in its predictions. ROC curves describe the predictive ability of the system

under the whole range of probability thresholds providing a global measure of model performance. From the ROC curve, a numeric index can be obtained as a measure of model performance: the area enclosed under the ROC curve (AUC), which ranges from 1 (perfect prediction) to 0 (random prediction). It has been shown in previous ecological studies that AUC is independent of the species prevalence (Manel et al. 2001; Allouche et al. 2006) and it is to be preferred as a measure of model accuracy when interest is focused in comparing and ranking the performance of different classifiers (Fielding and Bell 1997; Allouche et al. 2006).

A high AUC value, which indicates good model discrimination, is not necessarily coupled to a high numerical accuracy of the predictions. 'Calibration', also known as 'conditional bias' or 'reliability', can be described as the level of agreement between predicted and observed probabilities of occurrence within the whole range of probability values. Calibration plots are the basic tool for its assessment (Vaughan and Ormerod 2005). They can be constructed as follows: prediction probability values are discretized into fixed intervals (e.g.: cases with predicted value between 0 and 0.1 fall in the first interval, between 0.1 and 0.2 in the second, and so on...). For each interval, the mean predicted value and the true fraction of positive cases (i.e. the species prevalence) are computed and plotted on the X and Y axes respectively (Fig. 3). The system is better calibrated the closer the curve is to the 45 degree diagonal, which indicates a perfect agreement between predicted probability and true prevalence. If the curve passes under the diagonal, it is an indication that the system is over-estimating the event and on the contrary, when the curve passes above the diagonal the probability of the event is being under-estimated.

Another important measure of the quality of a binary probabilistic prediction is the

resolution, which quantifies the deviation of the prediction from the true species prevalence. It can be represented by means of a resolution plot, in which a partition of the whole range of possible probabilities is represented by a histogram of the corresponding relative prevalences (Fig. 4). 'U'-shaped histograms are characteristic of good model resolutions (e.g. Fig. 4a), where most absence events are predicted with probability zero and most positive cases are predicted with probability 1. On the contrary, more ambiguous predictions are characterized by a relatively higher proportion of intermediate probabilities (e.g. Fig. 4b).

In applications in which a deterministic prediction is required, a particular probability threshold must be set. However, AUC does not provide a probability threshold for case classification, which must be selected based on the objectives of each particular case-study, depending if the aim is either minimizing FP or FN error rates or any other previous condition imposed by the user (Fielding and Bell 1997; Freeman and Moisen 2008). In order to illustrate model performance for deterministic predictions, we set two probability thresholds: the observed probability of occurrence of each species (i.e. their prevalence) and a ROC-optimized probability threshold (OPT) which is obtained by reading the point from the ROC curve at which the sum of sensitivity and specificity is maximized. The latter is equivalent to finding the point on the ROC curve whose tangent has a slope of one. It has been frequently applied in spite of its known tendency to overestimate the true occurrence of species with low prevalence (Manel et al. 2001; Freeman and Moisen 2008). From them, we derived the corresponding confusion matrices and the following summary statistics: sensitivity, specificity, Cohen's kappa and true skill statistic (TSS). The last two statistics have the advantage of correcting the overall accuracy of models by the accuracy expected to take place by chance alone (Fielding and Bell 1997; Manel et al. 2001; Allouche et al. 2006).

TSS has the additional advantage of being fully independent of the species prevalence and the size of the validation dataset, whereas kappa may introduce statistical artifacts to estimates of predictive accuracy as it responds in an unimodal fashion to prevalence (Allouche et al. 2006). In spite of its shortcomings, Cohen's kappa was also computed because of its widespread use in ecological literature.

*d. Evaluation of the models: procedure*

In order to avoid overfiting and to obtain robust estimates of model performance, we carried out a Leave-One-Out Cross Validation procedure (LOOCV), also known as "jackknife resampling" to compute the error (Verbyla and Litvaitis 1989; Fielding and Bell 1997). LOOCV is a resampling technique in which $n-1$ instances out of the total of $n$, are used as the training dataset and the remaining one is used for testing. The procedure is repeated $n$ times, one per observed instance, producing a more precise estimation of classification accuracy (Verbyla and Litvaitis 1989). Manel et al. (1999) used training and test datasets from separate geographical regions, showing that LOOCV models provide a suitable alternative to independent data set testing in order to assess model performance.

Moreover, in addition to the *full* models, obtained by considering the full 14 input environmental variables described in Table 3, we also conducted a variable selection procedure for each of the species in order to obtain *reduced* models with optimum predictors. To this aim, we applied a stepwise logistic regression using the Akaike Information Criterion (AIC), obtaining the set of input variables displayed in Table 1 (see Sakamoto et al. 1986, for details on this method). Solar radiation and proportion of scree surface were the variables that most

often entered in the models (80%) followed by altitude and proportion of partially vegetated screes (73%). Terrain convexity, permanently waterlogged areas and solifluction terraces on fine materials were also often included in the models (67%). These variables were statistically very significant in almost all occasions. On the contrary, the variable fixed slopes did not enter in any model and the proportion of lake area entered only twice, although it was never statistically significant. We want to remark that more sophisticated variable/feature selection methods could be applied in this work (some examples are commented in Section 4b), but due to the limited number of variables available in our particular problems we preferred to use a benchmark method based on the most simple model for comparison purposes.

The comparison of the performance of the full and reduced models provides useful information about the sensitivity of the different methods on the number of input variables and overfiting.

# 3. Results

*a. Probabilistic predictions*

Table 1 shows the AUC values obtained after applying the LOOCV procedure to each model. In the case of MLR and ANN, the use of specific predictors always lead to better results, with the only significant exception of the ANN trained for *Juniperus nana*; in this case, there seem to be some nonlinear information hidden in the variables discarded by the stepwise logistic regression. Thus, some benefit could be obtained by using nonlinear feature selection algorithms, but the corresponding AUC is quite low and no much benefit

is expected. Therefore, from now on, only the stepwise models will be considered for MLR and ANN techniques. As opposite to the previous case, for SVM models the results of the general models are better for a number of species: *Erica arborea*, *Festuca-Agrostis*, *Vaccinium uliginosum*, *Festuca eskia* and *Juniperus nana*, although differences were small except for the latter species, which obtained a low AUC ($< 0.7$) in both cases. Therefore, a clear advantage of SVMs is their capability to efficiently work in problems with a high number of input variables, requiring no variable selection preprocessing.

MLR models obtained the highest AUC values for half of the species, although differences in AUC with the rest of the models were usually small. On the contrary, CART models obtained almost always the lowest AUCs. ANNs obtained the best AUC for *Carex nigra* and equalled the best result achieved by MLR for *Erica tetralix*. SVMs obtained good results and differences in AUC with the best methods were usually small; this model in particular, jointly with MLR, provided the best performance for *Euphorbia polygalifolia*. In general, differences in AUC were marginal between MLR, ANN, SVM and MARS, and only CART models performed notably worse (see Fig. 2 for the four illustrative species shown in boldface in Table 1).

Reliability and resolution diagrams of the four selected species for illustration of the general results are shown in Figures 3 and 4, respectively. In most cases, MLR and MARS models are the most reliable, showing less deviance from the diagonal (perfect calibration) than ANN and SVM models, although in some cases ANN models achieved fairly good calibration (e.g. *Luzula caespitosa*, Fig. 3c; *Genista obtusiramea*, Fig. 3d) and sometimes even notably better than MLR models (e.g. *Luzula caespitosa*, Fig. 3c). In most cases, SVM models produced a more irregular calibration, and CART models resulted very unreliable.

On the other hand, in the case of resolution, in most cases MLR models were considerably worse than the other methods, exhibiting more uniform, and even "n" shaped probability outputs, grouped around the uninformative 0.5 probability value. For almost all species, the best predictive resolution was achieved by ANN and CART models, with some exceptions such as *Luzula caespitosa* (Fig. 4c). As illustrative examples, *Juniperus nana*, a poorly predicted species, obtained a bad resolution for all models (Fig. 4b). Conversely, the ANN and MARS models of *Festuca eskia* (Fig. 4a) or the ANN and SVM models of *Luzula caespitosa* (Fig. 4c) achieved high predictive resolutions.

*b. Threshold-dependent deterministic predictions*

In some applications, probabilistic predictions need to be converted into deterministic ones by defining an appropriate probability threshold. As we have already mentioned in Section 2c, in this study we have considered two different thresholds: the prevalence of the species and a ROC-optimized probability threshold. In this case, a number of validation scores is commonly used in order to focus on different aspects of the prediction. Table 4 shows the results for the four species used as illustrative examples in this paper, although these results are generalizable to the 15 species modeled. TSS and Cohen's kappa scores obtained highly correlated results for both probability thresholds, an indication than in the present case study, the selection of any of the two statistics for model assessment is not determinant on final model choice in most of the situations. The highest Cohen's Kappa generally corresponded to the highest TSS in the set of species tested. However, the same can not be said for the probability threshold, whose selection often affected the final model choice.

The selection of the probability threshold is therefore a critical step in final deterministic outcome and special attention should be paid at this point. For species with low predictive scores, the optimized thresholds highly varied among different methods (e.g. *Juniperus nana*, Fig. 2), becoming under this particular circumstance an unreliable criterion for case classification.

*c. Model output mapping*

In order to analyze graphically the performance of the different methods, in this section we present several comparative geographical representations of illustrative species and models. For instance, Fig. 5 shows the predicted probability maps corresponding to the models with best AUC for the four representative species; a visual comparison with Fig. 1 (the observed presence/absence maps) gives an idea of the model capabilities.

The predicted probabilities can be ecologically translated into "habitat suitability" maps for each target species. Following the predicted probability intervals indicated by the grey scale tones in the maps, it is noticeable the high predictive resolution achieved for *Festuca eskia* and *Luzula caespitosa*, in contrast with the more intermediate probabilities predicted by model of *Juniperus nana*, which leads to a more ambiguous prediction of habitat suitability (note that this species had a low AUC score indicating a poor predictive skill from environmental descriptors).

Fig. 6 displays the mapped probabilities for the five models corresponding to the species *Genista obtusiramea*. Independently of the relative performance of each model in terms of AUC or TSS, it becomes evident that even for species distribution models with fairly good

accuracy, different modeling techniques provide rather different habitat suitability maps. Note that, according to Tables 1 and 4 the best models in this case are the MLR and SVM, which show similar mapping distributions in this figure.

# 4. Discussion and Conclusions

*a. Ecological interpretation of results*

Most validated models achieved AUC values characteristic of useful applications (0.7-0.9) and a high accuracy in some occasions ($> 0.9$, *Euphorbia polygalifolia*, *Festuca eskia*, *Genista obtusiramea*, *Juncus trifidus*, *Luzula caespitosa* and *Vaccinium uliginosum*), evidencing that topography and geomorphology are main controlling factors of vegetation distribution in the site at the scale of analysis. Only one species, *Juniperus nana*, obtained a low accuracy AUC value ($< 0.7$), an indication of the inadequacy of the models developed for its prediction. *Juniperus nana* is an ubiquitous species in the study area, although its presence is restricted to very low values of total cover in many places, favoured by micro-scale factors beyond the generic factors considered in this study. The estimated cover of *Juniperus nana* was lower than 5% in more than 80% of the polygons in which the species was present. The same can be said of *Vaccinium myrtillus*, another poorly predicted species which had an estimated cover lower than 5% in 77% of polygons in which it was present. In spite of this, we did not find any clear relationship between relative cover and prediction success, an indication that the predictive ability of input variables varies greatly depending on the target species. The inclusion of other factors not considered as predictors, such as grazing

389 pressure or history of fire disturbances may be important in order to improve model pre-
390 diction for some species. For instance, heathlands dominated by *Erica arborea* or *Calluna*
391 *vulgaris* have a long history of traditional management in which plant succession has been
392 arrested through regular burning (Webb 1998). In the same way, it must be reminded that
393 the study site has constituted for centuries an estival forage resource of prime importance for
394 local farmers. Herbivores are known to have a predominant role in community composition
395 in those ecosystems where they operate (Milchunas et al. 1988). It is likely that part of
396 this source of variation is explained by topographical variables (mainly slope) which partly
397 determine the geographical extent of ungulate distribution, but still part of it is probably
398 not accountable by physical descriptors, and other factors related to historical land uses,
399 farming practices or ungulate grazing behaviour should be explicitly considered (e.g. assem-
400 bly/shelter points, traditional pathways, location of preferred forage sources...). Similarly,
401 encroachment by the endemic broom species *Genista obtusiramea* is known to have taken
402 place very fast in recent decades, almost certainly favoured by the decay of traditional sheep
403 grazing that today does not exist anymore (*pers. comm.* of local farmers confirmed by own
404 data based on historical aerial photographs). These situations in which disturbance plays a
405 significant role in landscape composition contrast with the more stable conditions governing
406 the upper, less accessible parts of the study area where anthropogenic disturbance occurs at
407 a much lesser extent. In fact, models of plant species restricted to this zone (*Juncus trifidus*,
408 *Festuca eskia*, *Luzula caespitosa*, *Vaccinium uliginosum*) achieved very high predictive accu-
409 racy. This is connected with the concept of equilibrium. The assumption that species are in
410 pseudo-equilibrium with their environment has been recognised as a convenient theoretical
411 framework in species modeling (Guisan and Theurillat 2000), and it is in accordance with

the characteristics of the study site. Therefore, anthropogenic factors pose new challenges in species distribution models in alpine rangelands, requiring the definition of predictors able to effectively describe these processes and interactions at different temporal and spatial scales based on previous knowledge of site characteristics, historical land uses and ecological theory.

*b. Considerations on variable selection procedure*

Variable selection is a crucial step as it affects the modeled spatial distribution of species (Araújo and Guisan 2006). Although the focus of this study was not on the different procedures of variable selection, our results show that a previous selection of variables improved predictive performance in almost all occasions, and only two ANN models of low performance (*Calluna vulgaris* and *Juniperus nana*) obtained higher AUCs when all variables were included. However, SVM models deserve special attention: of the 15 species modeled, the prediction of four of them (namely *Calluna vulgaris*, *Carex nigra*, *Genista obtusiramea* and *Vaccinium myrtillus*) was unaffected by variable simplification and other four of them had better performance when including all variables (*Erica arborea*, *Festuca-Agrostis*, *Festuca eskia* and *Juniperus nana*). This is in accordance with the findings of Drake et al. (2006) who observed that useful information can be obtained from SVM models by the addition of more environmental variables even if they are highly correlated, obtaining more consistent models without previous data reduction. On the other hand, Guisan et al. (2002) warned about the stepwise selection procedures based on AIC, as small variations in the response data may lead to vast changes in final model selection. This inconsistency among selected

variables was also detected by Manel et al. (1999) when they compared models for the whole study area and from partitioned regional data for the same river bird species. Furthermore, we are aware that the utilization of a linear procedure for variable selection in a non-linear context is not the most appropriate choice as important information might be lost in the process. Alternative variable selection procedures may prove useful in ecological applications, such as techniques related to the analysis of variable contribution in ANNs (Gevrey et al. 2003; Romero and Sopena 2008) or a genetic algorithm-based approach (D'Heyere et al. 2006). Another interesting alternative to stepwise AIC procedure in the case of GLMs is to previously identify variable interactions fitting a CART model (Guisan et al. 2002).

*c. Considerations on the spatial component*

Many ecological studies recognise explicitly the spatial heterogeneity of ecosystems and the spatially structured environmental factors as important properties controlling species distribution among regions and landscapes (Legendre and Fortin 1989; Wagner and Fortin 2005). One limitation shared by all our models lies in the correlative nature of vegetation data used for model construction, which implies the need to account for the spatial component. Our models are therefore spatially invariant as we did not include explicitly descriptors of the neighbouring spatial context (Guisan et al. 2006). A visual inspection of the geographical distribution of failed predictions (Fig. 7) suggests a spatial pattern of errors, at least in the three species with good to moderate predictive performance (7a,c and d), which are approximately distributed along the boundaries of the species' distributions within the site (Fig. 1). This suggests that the use of explicit spatial descriptors as input variables into

the models might improve their predictive ability (Dormann 2007). Recently developed methodologies such as Principal Coordinates of Neighbour Matrices (Borcard and Legendre 2002) allow the inclusion of the spatial component at all scales leaving an opened door to the improvement of our models eventually leading to more consistent predictions and more reliable spatial representations of the species' realized niches.

A detailed analysis of both variable selection and spatial factors are out of the scope of this paper and will be undertaken in a more general context including both point and spatial descriptors.

*d. Relative advantages of techniques tested*

Attending solely to AUC values, MLR seems to be the preferred predictive method in this case study, although SVMs, ANNs and specially MARS models obtained comparable or better AUCs in some occasions. Our results are in accordance with other previous comparative studies which show the relative competitiveness of predictions of MLR models when compared to other modeling techniques able to cope with skewed or multi-modal responses such as ANNs (Manel et al. 1999) or CART and MARS (Muñoz and Felicísimo 2004).

Resolution is an important characteristic of the predictive model, specially when a deterministic outcome is required, since the classifier will be less sensitive to probability threshold selection when the predictions are grouped around the values 0 and 1. In this sense, MLR models exhibited almost in all occasions the worst resolution. On the contrary, ANN models most frequently achieved the best predictive resolution. Thus, ANN results are expected to be more stable for varying probability thresholds because of their ability to effectively

separate case occurrences. CART models, in spite of their lower accuracy, also showed very good resolution. SVM models showed more variability although in general also attained fairly good predictive resolution. MARS classifiers, in spite of their good properties of accuracy and calibration did not achieve competitive results in this sense. Thus, our results reveal that a high predictive accuracy is not necessarily related to a good reliability, and an acceptable compromise between both should be achieved. Reliability is related to model calibration and, hence, is an important issue in the light of the "habitat suitability" concept (Guisan and Thuiller 2005). If research interests are focused on this, calibration is a key aspect that should be specifically addressed. High deviances from the observed probabilities mean unreliable habitat suitability maps that may misguide management actions, for example for the identification of potential species' distributions (e.g. Garzón et al. (2006)) or for supporting conservation planning and natural reserve selection (Araújo et al. 2004).

In general, the predictive technique chosen should consider all these factors and the decision should be based on the nature of the data to analyze, a sound understanding of the limitations and assumptions behind the theoretical background of each technique and the final practical aim of each researcher. A framework in which several modeling techniques are tested and compared is the recommended option provided the variability of results obtained. Model validation, whether using independent data sets or, as it is the case of this study, using any of the available resampling techniques, proves a vital step in model assessment if reliable measures of performance are to be obtained.

# 5. Aknowledgements

We thank Jesús Delgado and Sergio Cabañas their assistance in field mapping. This work was supported by the Spanish Institute for Agricultural and Food Research and Technology [INIA, RTA2005-00160-C02-00].

| Target species | Variables selected | MLR | | ANN | | SVM | | CART | MARS |
|---|---|---|---|---|---|---|---|---|---|
| | | Full | Step. | Full | Step. | Full | Step. | | |
| *Calluna vulgaris* | 1**,3,5,6*,7***,8**,11**,13*,14** | 0.69 | **0.72** | 0.63 | 0.62 | 0.60 | 0.60 | 0.70 | **0.7**2(1) |
| *Carex nigra* | 2***,3*,5,8**,12***,14*** | 0.83 | 0.85 | 0.82 | **0.87** | 0.86 | 0.86 | 0.80 | 0.85(3) |
| *Cytisus oromediterraneus* | 1,2,3***,5*,6**,8**,11,12***,13*,14* | 0.82 | 0.83 | 0.76 | 0.83 | 0.82 | 0.84 | 0.67 | **0.85**(2) |
| *Erica arborea* | 2,3*,5,7*,11***,12***,13*** | 0.85 | 0.88 | 0.82 | 0.82 | 0.88 | 0.87 | 0.67 | **0.89**(1) |
| *Erica tetralix* | 2***,3,4,5,6*,8*,9,11*,13***,14*** | 0.81 | **0.82** | 0.78 | **0.82** | 0.77 | 0.80 | 0.75 | 0.81(3) |
| *Euphorbia polygalipholia* | 2***,6,8,11***,12,13*** | 0.91 | **0.92** | 0.87 | 0.89 | 0.91 | **0.92** | 0.85 | 0.91(1) |
| *Festuca-Agrostis* | 1*,2***,3***,4,5,6***,9,12***,14* | **0.87** | **0.87** | 0.81 | 0.84 | 0.86 | 0.85 | 0.82 | **0.87**(1) |
| ***Festuca eskia*** | 1,2*,5,8***,11***,13***,14* | 0.95 | **0.96** | 0.94 | 0.94 | **0.96** | 0.95 | 0.89 | 0.95(1) |
| ***Genista obtusiramea*** | 1,2***,3*,11***,12*,13***,14** | 0.90 | **0.91** | 0.85 | 0.88 | 0.90 | 0.90 | 0.83 | 0.89(2) |
| *Juncus trifidus* | 1***,4*,5,6***,11***,13*** | 0.93 | 0.95 | 0.92 | 0.94 | 0.93 | 0.95 | 0.88 | **0.97**(1) |
| ***Juniperus nana*** | 5*,6**,7***,8,12**,14*** | 0.64 | 0.65 | **0.68** | 0.63 | 0.66 | 0.62 | 0.56 | **0.68**(3) |
| ***Luzula caespitosa*** | 2,3,5,6*,8,11**,13***,14** | 0.91 | **0.92** | 0.85 | 0.89 | 0.90 | 0.91 | 0.85 | 0.91(1) |
| *Nardus stricta* | 3**,6*,7,11,12***,13** | 0.81 | **0.84** | 0.77 | 0.78 | 0.80 | 0.81 | 0.72 | 0.77(1) |
| *Vaccinium myrtillus* | 1,2*,4*,5*,6,7***,11***,13***,14* | 0.79 | 0.81 | 0.72 | 0.74 | 0.80 | 0.80 | 0.72 | **0.83**(3) |
| *Vaccinium uliginosum* | 3*,4**,5**,6***,8*,11*** | 0.92 | **0.94** | 0.84 | 0.84 | 0.83 | 0.78 | 0.82 | 0.90(1) |

TABLE 1. Area under the ROC curve (AUC) for the different models and species resulting from leave-one-out cross validation. For MLR, ANN and SVM modeling techniques, columns indicate AUC values for both full models (all 14 input variables included) and stepwise models (only variables selected by the stepwise AIC procedure included). See Table 3 for variable codes. Statistical significance of variables in the MLR stepwise models is also indicated (Signif. codes: $P < 0.001^{(***)}$, $P < 0.01^{(**)}$, $P < 0.05^{(*)}$). The four target species used as examples in the figures and best AUC values obtained for each species are highlighted in boldface. In the MARS column, values in parenthesis indicate the order interaction used.

# REFERENCES


Allouche, O., A. Tsoar, and R. Kadmon, 2006: Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss). *Journal of Applied Ecology*, **43 (6)**, 1223–1232.

Araújo, M. B., M. Cabeza, W. Thuiller, L. Hannah, and P. H. Williams, 2004: Would climate change drive species out of reserves? an assessment of existing reserve-selection methods. *Global Change Biology*, **10 (9)**, 1618–1626.

Araújo, M. B. and A. Guisan, 2006: Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33 (10)**, 1677–1688.

Borcard, D. and P. Legendre, 2002: All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, **153 (1–2)**, 51–68.

Breiman, L., J. Friedman, R. A. Olshen, and C. J. Stone, 1984: *Classification and Regression Trees*. Wadsworth Publishing Company, Belmont, California, U.S.A.

Castejón, M., J. B. Ordieres, E. P. Vergara, F. J. M. de Pisón, A. V. Pernía, and F. Alba, 2007: *AMORE: A MORE flexible neural network package*. URL `http://wiki.r-project.org/rwiki/doku.php?id=packages:cran:amore`, r package version 0.2-11.

Cortes, C. and V. Vapnik, 1995: Support-vector networks. *Machine Learning*, **20 (3)**, 273–297.

D'Heyere, T., P. L. M. Goethals, and N. De Pauw, 2006: Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks. *Ecological Modelling*, **195 (1–2)**, 20–29.

Dimitriadou, E., K. Hornik, F. Leisch, D. Meyer, and A. Weingessel, 2009: *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. URL `http://CRAN.R-project.org/package=e1071`, r package version 1.5-22.

Dormann, C. F., 2007: Effects of incorporating spatial autocorrelation into the analysis of species distribution data. *Global Ecology and Biogeography*, **16 (2)**, 129–138.

Drake, J. M. and J. M. Bossenbroek, 2009: Profiling ecosystem vulnerability to invasion by zebra mussels with support vector machines. *Theoretical Ecology*, **2 (4)**, 189–198.

Drake, J. M., C. Randin, and A. Guisan, 2006: Modelling ecological niches with support vector machines. *Journal of Applied Ecology*, **43 (3)**, 424–432.

Elith, J. and J. R. Leathwick, 2009: Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology Evolution and Systematics*, **40**, 677–697.

Elith, J., et al., 2006: Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29 (2)**, 129–151.

Fielding, A. H. and J. F. Bell, 1997: A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24 (1)**, 38–49.

538 Freeman, E. A. and G. G. Moisen, 2008: A comparison of the performance of threshold
539 criteria for binary classification in terms of predicted prevalence and kappa. *Ecological*
540 *Modelling*, **217 (1–2)**, 48–58.

541 Friedman, J. H., 1991: Multivariate adaptive regression splines. *Annals of Statistics*, **19 (1)**,
542 1–67.

543 Garzón, M. B., R. Blazek, M. Neteler, R. S. de Dios, H. S. Ollero, and C. Furlanello, 2006:
544 Predicting habitat suitability with machine learning models: The potential area of pinus
545 sylvestris l. in the iberian peninsula. *Ecological Modelling*, **197 (3–4)**, 383–393.

546 Gevrey, M., L. Dimopoulos, and S. Lek, 2003: Review and comparison of methods to study
547 the contribution of variables in artificial neural network models. *Ecological Modelling*,
548 **160 (3)**, 249–264.

549 Gude, J. A., M. S. Mitchell, D. E. Ausband, C. A. Sime, and E. E. Bangs, 2009: Internal vali-
550 dation of predictive logistic regression models for decision-making in wildlife management.
551 *Wildlife Biology*, **15 (4)**, 352–369.

552 Guisan, A., T. C. Edwards, and T. Hastie, 2002: Generalized linear and generalized additive
553 models in studies of species distributions: setting the scene. *Ecological Modelling*, **157 (2–**
554 **3)**, 89–100.

555 Guisan, A., A. Lehmann, S. Ferrier, M. Austin, J. M. C. Overton, R. Aspinall, and T. Hastie,
556 2006: Making better biogeographical predictions of species' distributions. *Journal of Ap-*
557 *plied Ecology*, **43 (3)**, 386–392.

Guisan, A. and J. P. Theurillat, 2000: Equilibrium modeling of alpine plant distribution: how far can we go? *Phytocoenologia*, **30 (3–4)**, 353–384.

Guisan, A. and W. Thuiller, 2005: Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8 (9)**, 993–1009.

Guisan, A. and N. E. Zimmermann, 2000: Predictive habitat distribution models in ecology. *Ecological Modelling*, **135 (2–3)**, 147–186.

Hastie, T., R. Tibshirani, and J. Friedman, 2010: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. 2d ed., Springer Series in Statistics, Springer, URL `http://www-stat.stanford.edu/~tibs/ElemStatLearn/main.html`.

Hutchinson, G. E., 1957: Population studies - animal ecology and demography - concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 415–427.

Jolliffe, I. and D. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons.

Leathwick, J. R., J. Elith, and T. Hastie, 2006: Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological Modelling*, **199 (2)**, 188–196.

Legendre, P. and M. J. Fortin, 1989: Spatial pattern and ecological analysis. *Vegetatio*, **80 (2)**, 107–138.

Leisch, F., K. Hornik, and B. D. Ripley, 2009: *mda: Mixture and flexible discriminant analysis*. URL `http://CRAN.R-project.org/package=mda`, r package version 0.4-1.

Lek, S., M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga, and S. Aulagnier, 1996: Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, **90 (1)**, 39–52.

Lek, S. and J. F. Guegan, 1999: Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling*, **120 (2–3)**, 65–73.

Lewin-Koh, N. J., et al., 2009: *maptools: Tools for reading and handling spatial objects*. URL `http://CRAN.R-project.org/package=maptools`, r package version 0.7-29.

Manel, S., J. M. Dias, S. T. Buckton, and S. J. Ormerod, 1999: Alternative methods for predicting species distribution: an illustration with himalayan river birds. *Journal of Applied Ecology*, **36 (5)**, 734–747.

Manel, S., H. C. Williams, and S. J. Ormerod, 2001: Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, **38 (5)**, 921–931.

McCune, B., 2006: Non-parametric habitat models with automatic interactions. *Journal of Vegetation Science*, **17 (6)**, 819–830.

Milchunas, D. G., O. E. Sala, and W. K. Lauenroth, 1988: A generalized-model of the effects of grazing by large herbivores on grassland community structure. *American Naturalist*, **132 (1)**, 87–106.

Muñoz, J. and A. M. Felicísimo, 2004: Comparison of statistical methods commonly used in predictive modelling. *Journal of Vegetation Science*, **15 (2)**, 285–292.

Oksanen, J. and P. R. Minchin, 2002: Continuum theory revisited: what shape are species responses along ecological gradients? *Ecological Modelling*, **157 (2–3)**, 119–129.

R Development Core Team, 2009: *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing, URL `http://www.R-project.org`, ISBN 3-900051-07-0.

Ripley, B., 2009: *tree: Classification and regression trees*. URL `http://CRAN.R-project.org/package=tree`, r package version 1.0-27.

Romero, E. and J. M. Sopena, 2008: Performing feature selection with multilayer perceptrons. *IEEE Transactions on Neural Networks*, **19 (3)**, 431–441.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa, 1986: *Akaike information criterion statistics*. Reidel Publishing Company.

Santika, T. and M. F. Hutchinson, 2009: The effect of species response form on species distribution model prediction and inference. *Ecological Modelling*, **220 (19)**, 2365–2379.

Schlkopf, B. and A. Smola, 2001: *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Belmont, California, U.S.A.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer, 2009: *ROCR: Visualizing the performance of scoring classifiers.* URL `http://CRAN.R-project.org/package=ROCR`, r package version 1.0-4.

Thuiller, W., 2003: Biomod - optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology*, **9 (10)**, 1353–1362.

Vaughan, I. P. and S. J. Ormerod, 2005: The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, **42 (4)**, 720–730.

Vayssieres, M. P., R. E. Plant, and B. H. Allen-Diaz, 2000: Classification trees: An alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science*, **11 (5)**, 679–694.

Verbyla, D. L. and J. A. Litvaitis, 1989: Resampling methods for evaluating classification accuracy of wildlife habitat models. *Environmental Management*, **13 (6)**, 783–787.

Wagner, H. H. and M. J. Fortin, 2005: Spatial analysis of landscapes: Concepts and statistics. *Ecology*, **86 (8)**, 1975–1987.

Webb, N. R., 1998: The traditional management of european heathlands. *Journal of Applied Ecology*, **35 (6)**, 987–990.

# List of Tables

4 Accuracy of the models after computation of the deterministic binary response prediction. For simplicity, we illustrate only the four species used in the previous examples. Results presented correspond to the species prevalence and the ROC optimized (OPT) probability thresholds. OPT maximizes the sum of sensitivity (Sens) and specificity (Spec). True skill statistic (TSS) and Cohen's kappa ($K$) are also indicated. For each probability threshold, Best TSS results are presented in bold. Number in parenthesis next to MARS method indicate the order of interactions achieving the best result. 37

| **Target species** | **Prev** |
|---|---|
| *Calluna vulgaris* | 84 |
| *Carex nigra* | 25 |
| *Cytisus oromediterraneus* | 12 |
| *Erica arborea* | 16 |
| *Erica tetralix* | 26 |
| *Euphorbia polygalipholia* | 37 |
| *Festuca-Agrostis* | 54 |
| ***Festuca eskia*** | 50 |
| ***Genista obtusiramea*** | 28 |
| *Juncus trifidus* | 15 |
| ***Juniperus nana*** | 55 |
| ***Luzula caespitosa*** | 39 |
| *Nardus stricta* | 86 |
| *Vaccinium myrtillus* | 34 |
| *Vaccinium uliginosum* | 10 |

TABLE 2. Target species of this study and their prevalence, defined as the proportion in % of polygons in which the species is present (any cover value $> 0$ was considered a presence). The four species used in the paper for illustrative purposes are shown in boldface.

| Type | Code | Variable | Freq. | Units |
|---|---|---|---|---|
| Geomorphological | 1 | Boulder fields | 47 | |
| | 2 | Permanently waterlogged | 67 | |
| | 3 | Large solifluction terraces | 67 | |
| | 4 | Small Solifluction terraces | 33 | % |
| | 5 | Screes | 80 | polygon |
| | 6 | Screes partially fixed by vegetation | 73 | area |
| | 7 | Alluvial terrains | 33 | |
| | 8 | Rocky outcrops | 60 | |
| | 9 | Lakes | 13 | |
| | 10 | Fixed slopes | 0 | |
| Topographical | 11 | Solar radiation | 80 | $WH * m^{-2}$ |
| | 12 | Slope | 53 | Degrees |
| | 13 | Altitude | 73 | m |
| | 14 | Convexity | 67 | Non dimensional |

TABLE 3. Summary of environmental variables used in the models. The geomorphological variables were calculated as proportion of polygon area by overlay of vegetation and geomorphological maps (scale 1 : 2,000 m). Solar radiation was calculated as total radiation received by the modeled surface of the study area from 31 April to 31 October 2007. Slope, altitude and terrain convexity (curvature) were computed from the 2 m resolution digital elevation model of the study area. Frequency column (Freq.) indicates percentage of stepwise models in which the variable was included.

| Species | Method | Prevalence | | | | OPT | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Sens** | **Spec** | **TSS** | ***K*** | **Sens** | **Spec** | **TSS** | ***K*** |
| | MLR | 0.87 | 0.91 | 0.78 | 0.78 | 0.99 | 0.67 | 0.66 | 0.65 |
| *Festuca* | ANN | 0.89 | 0.87 | 0.75 | 0.75 | 0.95 | 0.75 | **0.70** | 0.70 |
| *eskia* | SVM | 0.90 | 0.89 | 0.79 | 0.79 | 1.00 | 0.67 | 0.67 | 0.67 |
| | CART | 0.90 | 0.83 | 0.73 | 0.73 | 0.94 | 0.02 | 0.03 | 0.03 |
| | MARS(1) | 0.92 | 0.88 | **0.80** | 0.80 | 1.00 | 0.68 | 0.68 | 0.68 |
| | MLR | 0.89 | 0.80 | **0.69** | 0.61 | 0.96 | 0.73 | **0.69** | 0.58 |
| *Genista* | ANN | 0.83 | 0.84 | 0.67 | 0.62 | 0.85 | 0.73 | 0.58 | 0.50 |
| *obtusiramea* | SVM | 0.90 | 0.76 | 0.66 | 0.57 | 0.92 | 0.75 | 0.66 | 0.57 |
| | CART | 0.81 | 0.86 | 0.67 | 0.64 | 0.86 | 0.72 | 0.58 | 0.50 |
| | MARS(2) | 0.89 | 0.76 | 0.65 | 0.56 | 0.91 | 0.75 | 0.65 | 0.56 |
| | MLR | 0.60 | 0.61 | 0.21 | 0.21 | 0.56 | 0.65 | **0.21** | 0.21 |
| *Juniperus* | ANN | 0.49 | 0.72 | 0.21 | 0.21 | 0.99 | 0.04 | 0.02 | 0.03 |
| *nana* | SVM | 0.64 | 0.56 | 0.20 | 0.20 | 0.69 | 0.51 | 0.19 | 0.20 |
| | CART | 0.61 | 0.61 | 0.22 | 0.22 | 0.71 | 0.31 | 0.03 | 0.03 |
| | MARS(3) | 0.63 | 0.70 | **0.34** | 0.33 | 0.88 | 0.24 | 0.13 | 0.14 |
| | MLR | 0.84 | 0.84 | 0.68 | 0.67 | 0.93 | 0.73 | 0.66 | 0.62 |
| *Luzula* | ANN | 0.84 | 0.84 | 0.68 | 0.67 | 0.93 | 0.73 | 0.66 | 0.62 |
| *caespitosa* | SVM | 0.83 | 0.85 | **0.69** | 0.68 | 0.88 | 0.80 | **0.67** | 0.65 |
| | CART | 0.77 | 0.86 | 0.64 | 0.64 | 0.83 | 0.81 | 0.65 | 0.63 |
| | MARS(1) | 0.86 | 0.83 | 0.68 | 0.67 | 0.93 | 0.70 | 0.63 | 0.59 |

TABLE 4. Accuracy of the models after computation of the deterministic binary response prediction. For simplicity, we illustrate only the four species used in the previous examples. Results presented correspond to the species prevalence and the ROC optimized (OPT) probability thresholds. OPT maximizes the sum of sensitivity (Sens) and specificity (Spec). True skill statistic (TSS) and Cohen's kappa ($K$) are also indicated. For each probability threshold, Best TSS results are presented in bold. Number in parenthesis next to MARS method indicate the order of interactions achieving the best result.

# List of Figures

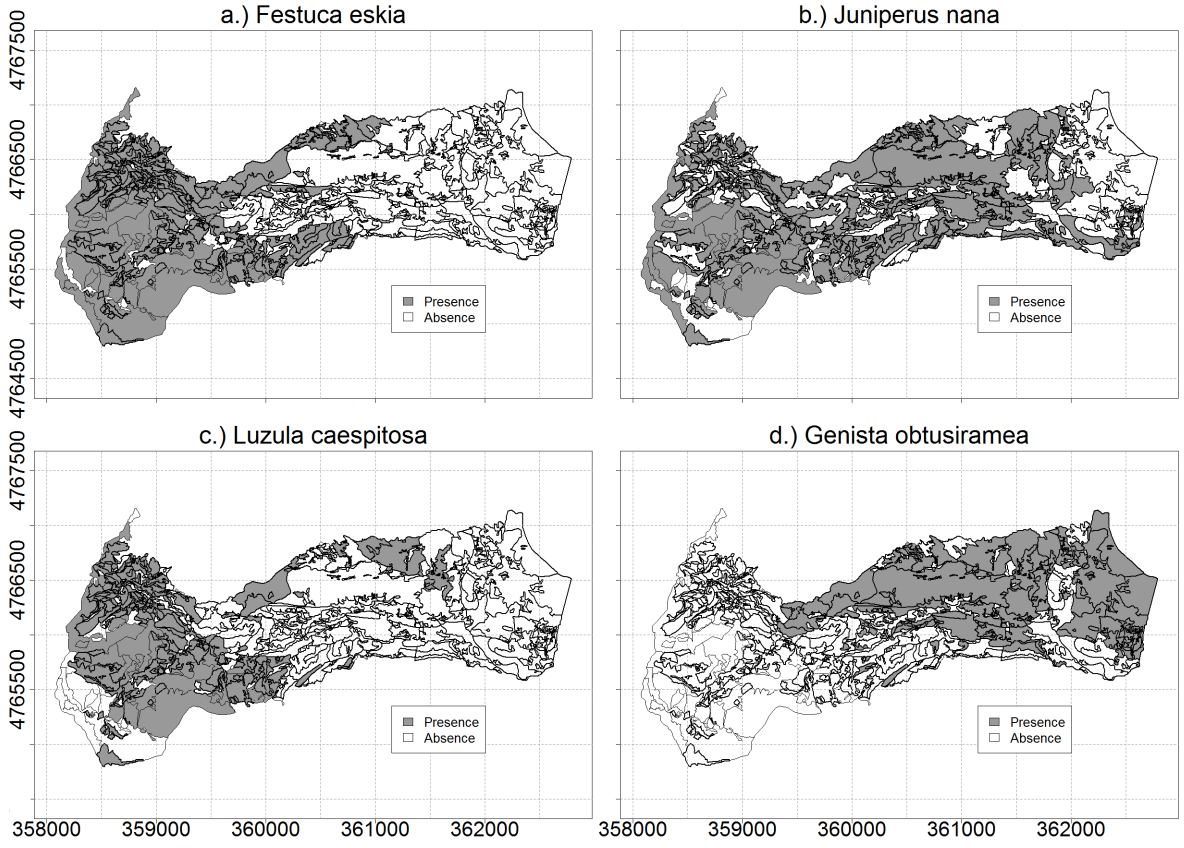


FIG. 1. Distribution maps of *Festuca eskia*, *Juniperus nana*, *Luzula caespitosa* and *Genista obtusiramea* in Riofrio rangeland. Units of axes are meters and correspond to the UTM ED-50 projected grid. Polygons excluded from the models due to very scarce or null plant cover are delimited by thin borders.

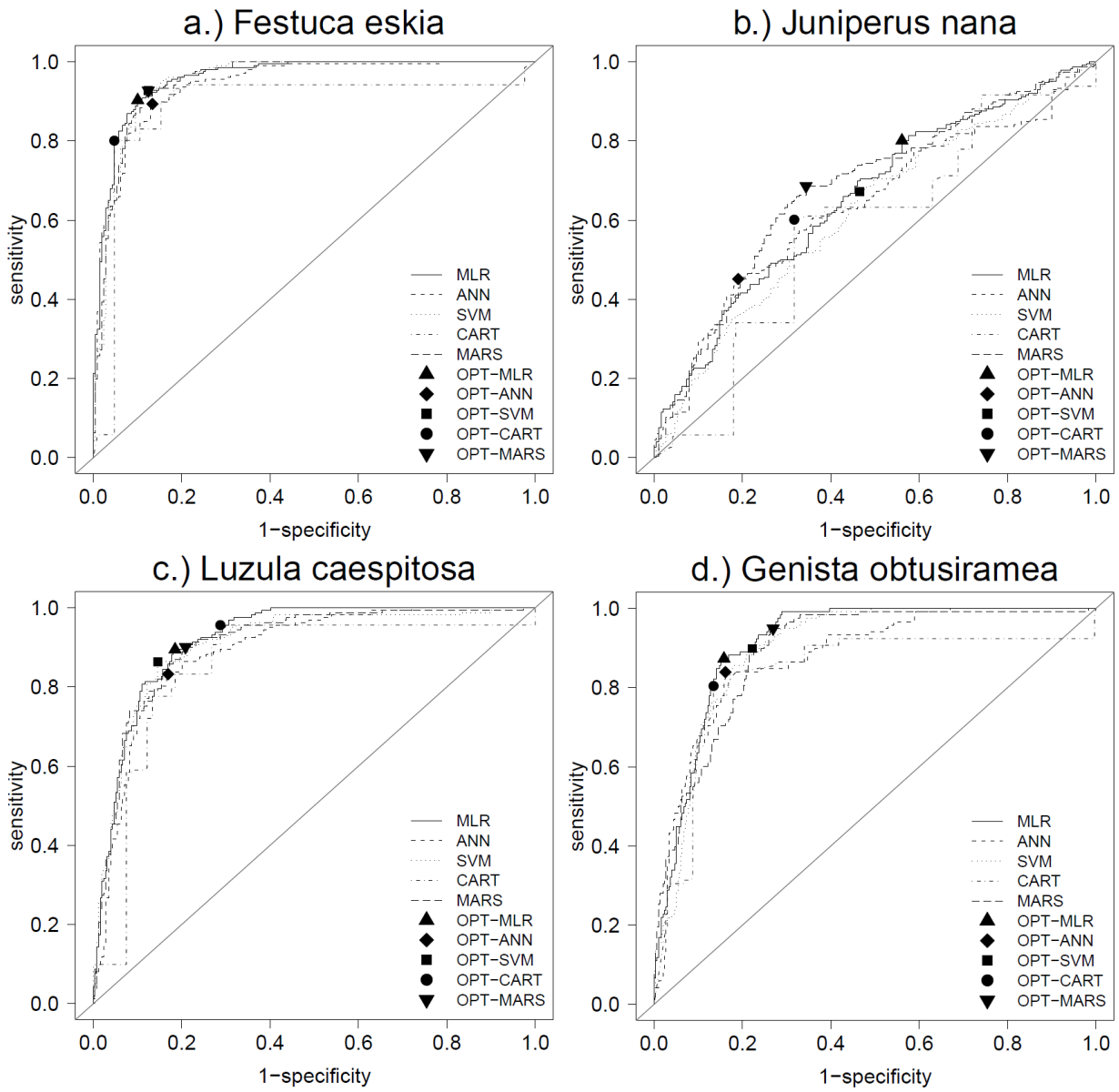


FIG. 2. Receiver-operating characteristics (ROC) curves corresponding to the species a) *Festuca eskia*, b) *Juniperus nana*, c) *Luzula caespitosa* and d) *Nardus stricta*. MLR, ANN and CART curves correspond to their stepwise versions. Optimised probability thresholds (OPTs) are indicated on their respective curves.

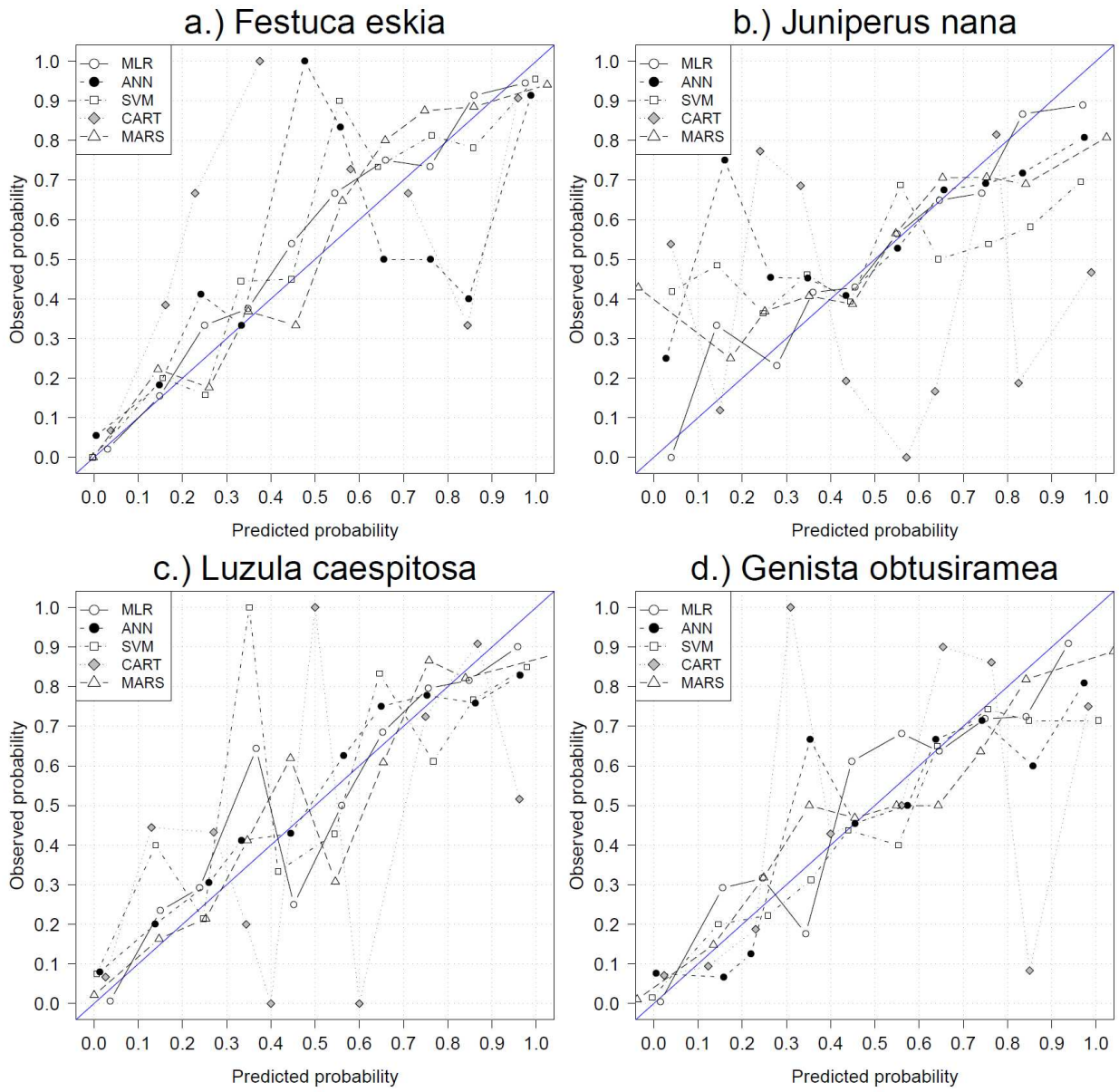


FIG. 3. Calibration diagrams of the models constructed for a) *Festuca eskia*, b) *Juniperus nana*, c) *Luzula caespitosa* and d) *Genista obtusiramea*. Results of the five modeling techniques are plotted in the same diagram. Incomplete lines indicate that some probability intervals were not predicted by the model (e.g. probability values between 0.4 and 0.5 were not predicted by CART model for *Festuca eskia*). Values below the diagonal indicate over-estimated probabilities and values above it under-estimated predictions.

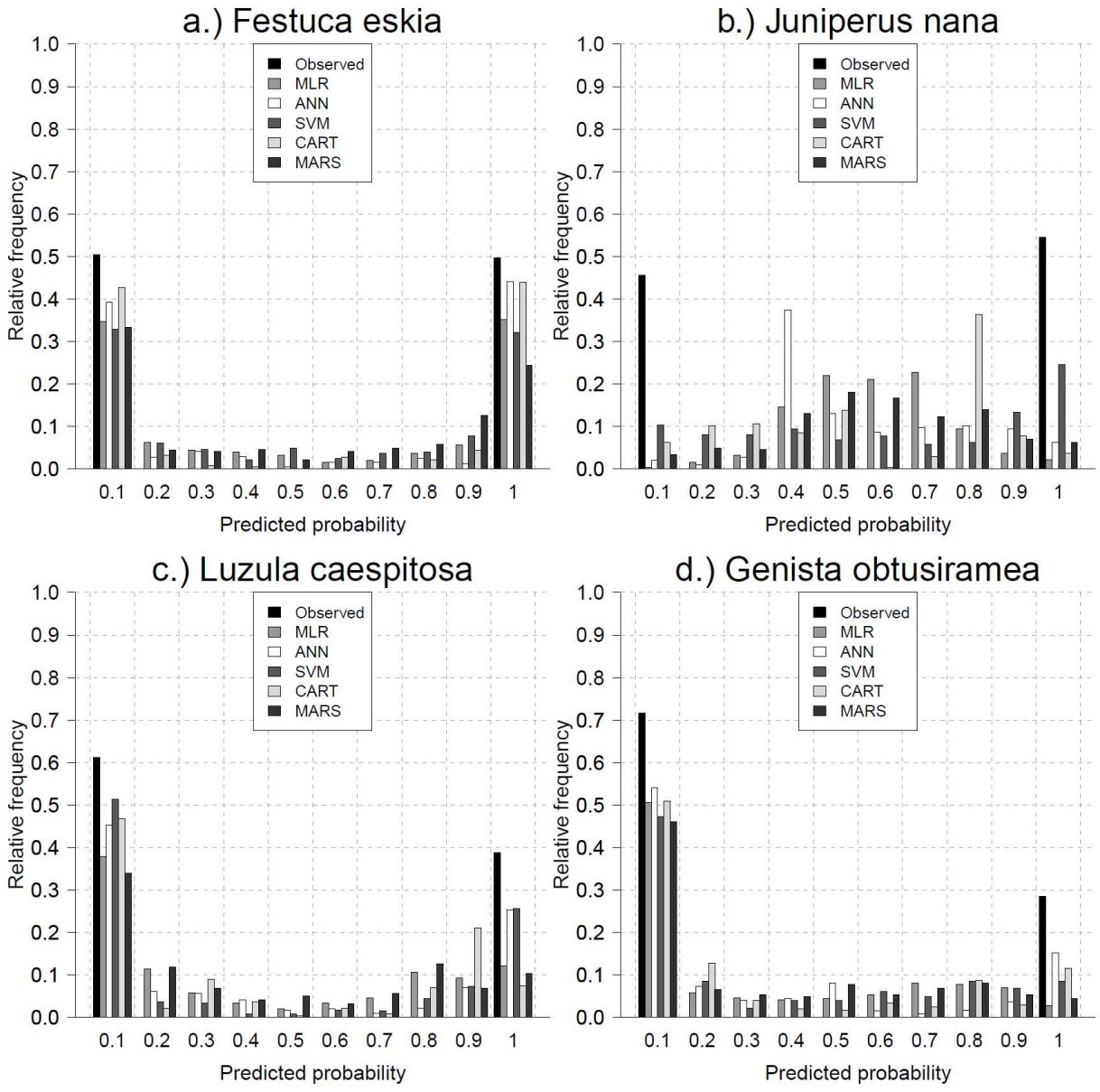


FIG. 4. Resolution diagram of the models constructed for a) *Festuca eskia*, b) *Juniperus nana*, c) *Luzula caespitosa* and d) *Genista obtusiramea*. The black bars for the 0 and 1 values indicate the observed presence/absence relative frequency of the species.

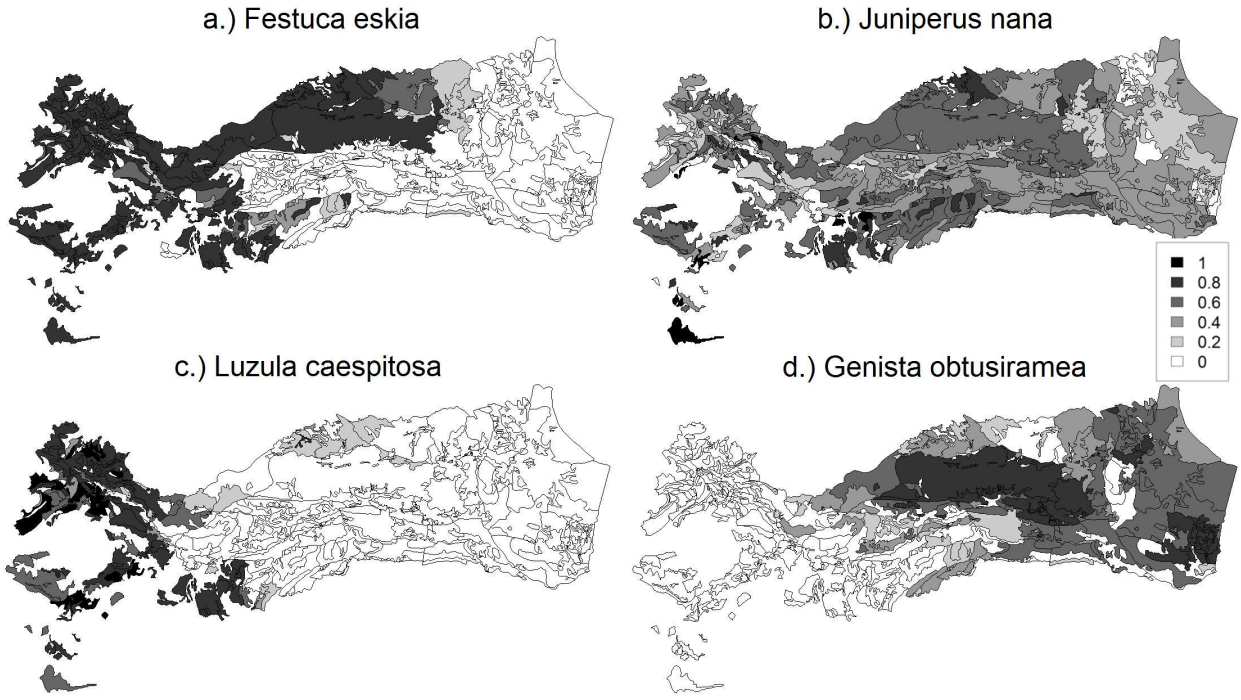


FIG. 5. Predicted probability maps of four species corresponding to models of varying accuracy: a.) ANN model of *Festuca eskia*, b.) MARS model of *Juniperus nana*, c.) SVM model of *Luzula caespitosa* and d.) MLR model of *Genista obtusiramea*. Probability intervals are represented by grayscale tones as indicated by the legend in the right hand side of the figure.

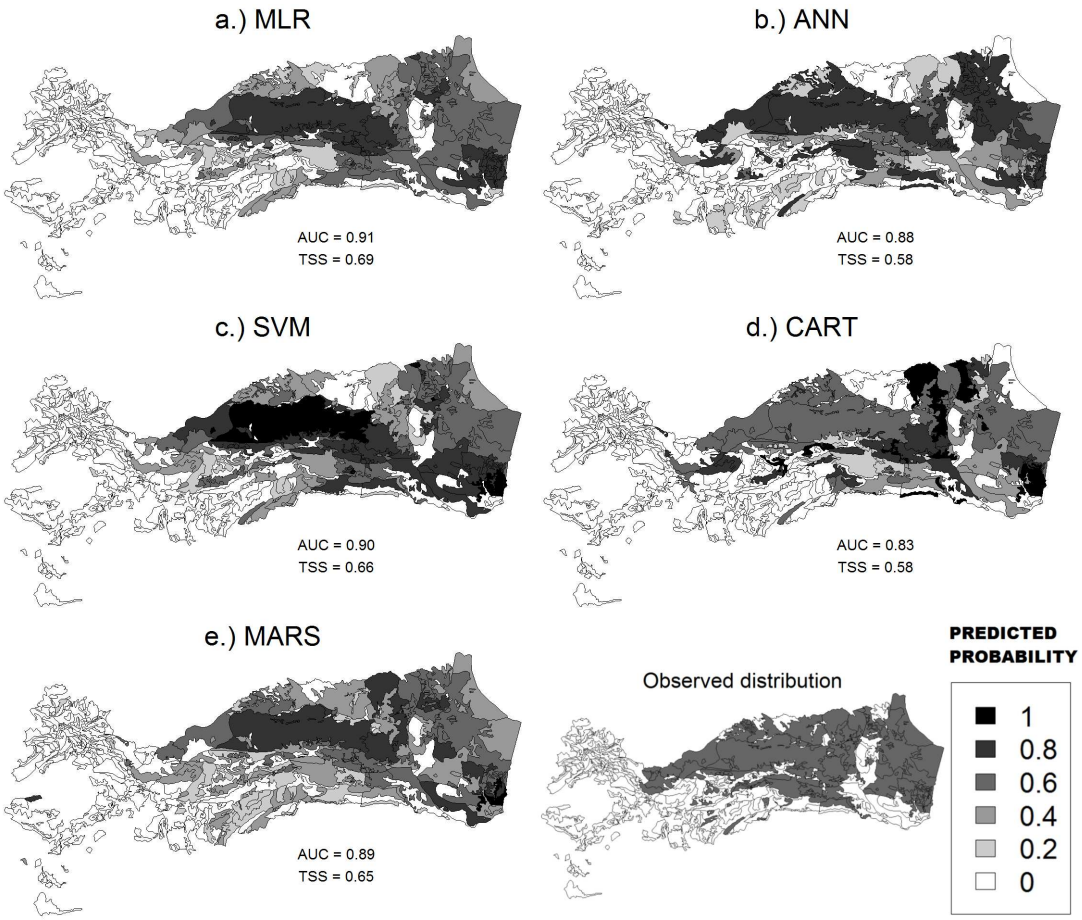


FIG. 6. Predicted probability maps of the species *Genista obtusiramea* according to the different modeling techniques tested: a.) multiple logistic regression, b.) artificial neural network, c.) support vector machine, d.) classification and regression tree and e.) multivariate adaptive regression splines.

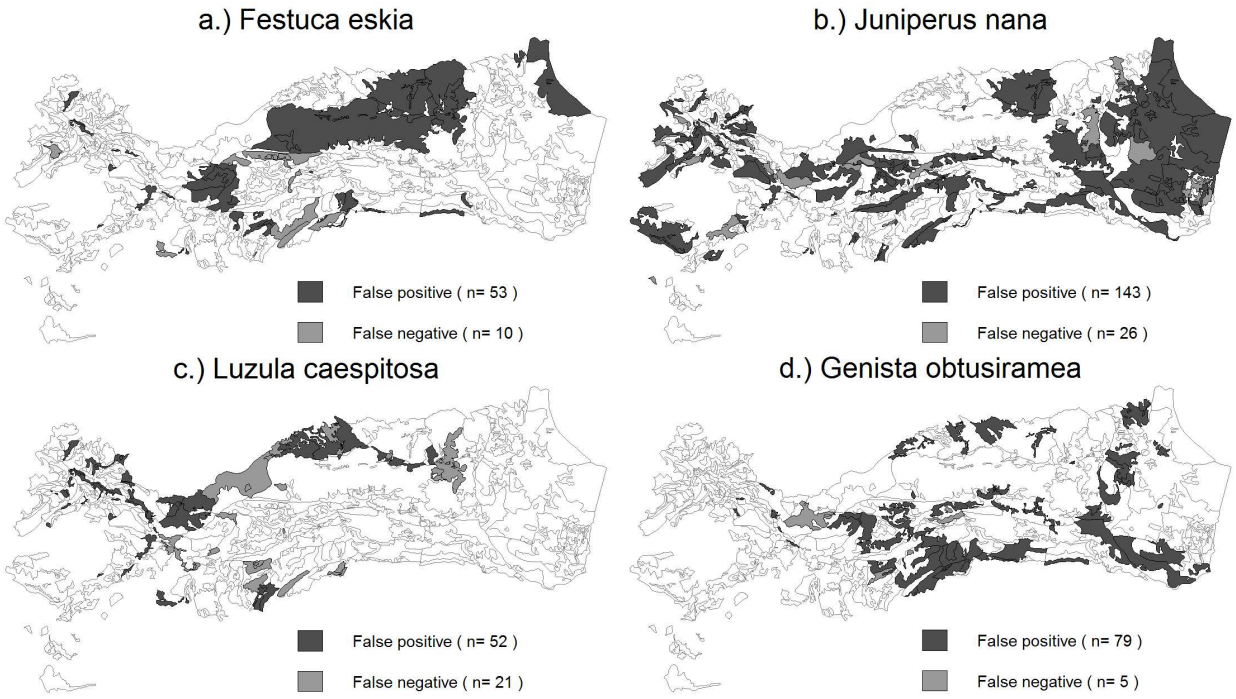


FIG. 7. Maps of wrong predictions (false positives and false negatives) of models displayed in Fig. 5. Deterministic predictions have been obtained using the optimized probability threshold (OPT) as cut-off value.