Journal of Vegetation Science 15: 285-292, 2004

Comparison of statistical methods commonly used in predictive modelling

Muñoz, Jesús[1*] & Felicísimo, Ángel M.[2]

[1]*Real Jardín Botánico, Plaza de Murillo 2, 28014 Madrid, Spain;*
[2]*Departamento de Expresión Gráfica, Universidad de Extremadura, 10071 Cáceres, Spain; E-mail amfeli@unex.es*
[*]*Corresponding author; E-mail jmunoz@ma-rjb.csic.es*

Abstract. Logistic Multiple Regression, Principal Component Regression and Classification and Regression Tree Analysis (CART), commonly used in ecological modelling using GIS, are compared with a relatively new statistical technique, Multivariate Adaptive Regression Splines (MARS), to test their accuracy, reliability, implementation within GIS and ease of use. All were applied to the same two data sets, covering a wide range of conditions common in predictive modelling, namely geographical range, scale, nature of the predictors and sampling method.

We ran two series of analyses to verify if model validation by an independent data set was required or cross-validation on a learning data set sufficed. Results show that validation by independent data sets is needed. Model accuracy was evaluated using the area under Receiver Operating Characteristics curve (AUC). This measure was used because it summarizes performance across all possible thresholds, and is independent of balance between classes.

MARS and Regression Tree Analysis achieved the best prediction success, although the CART model was difficult to use for cartographic purposes due to the high model complexity.

Keywords: Classification and Regression Tree; *Fagus*; *Grimmia*; Logistic regression; Multivariate Adaptive Regression Splines; Regression Tree Analysis.

Abbreviations: AUC = Area under the ROC curve; CART = Classification and Regression Trees; FN = False negative; FP = False positive; GAM = Generalized Additive Model; GIS = Geographic Information System; GLM = Generalized Linear Model; LMR = Logistic Multiple Regression; MARS = Multivariate Adaptive Regression Splines; NDVI = Normalized Difference Vegetation Index; PCR = Principal Components Regression; ROC = Receiver Operating Characteristics.

Introduction

Modelling studies have employed different statistical techniques to unravel the complexity of interactions between distributions and environmental factors. Those include Generalized Linear Models (GLM; Guisan et al. 1998), especially Logistic Multiple Regression (LMR; Narumalani et al. 1997; Felicísimo et al. 2002); Generalized Additive Models (GAM; Yee & Mitchell 1991) and Classification and Regression Trees (CART, also known as Regression Tree Analysis, RTA; Moore et al. 1991; Iverson & Prasad 1998). Recently, Guisan & Zimmermann (2000) made a comprehensive review of predictive modelling and noticed the lack of comparative studies in which more than two statistical techniques were applied to the same data set.

Most classical papers on predictive modelling are based on methodologies that assume a Gaussian relation between response and predictors, and also that the contribution to the response from the interactions among predictors is uniform across their range of values. Both assumptions are unwarranted in most cases (Austin & Cunningham 1981; Austin et al. 1990, 1994). Nevertheless, LMR with a quadratic function to represent Gaussian responses has often implied high predictive success. Further problems associated with classical regression analysis arise when many predictors are used. Such increase in the number of predictors implies an increase greater than exponential in the number of possible regression structures, and the almost inevitable problem with multicollinearity.

To sidestep these problems, analysts impose strong model assumptions, forcing the variables to act globally over the response by limiting or eliminating local changes in response or interactions. This strategy can no longer be justified if suitable predictors are likely to act and interact differently on the response variable across their range of values. The search for a model that handles the above problems lead us to experiment with a relatively new statistical technique employed in data-mining strategies in fields such as chemical engineering, marketing

campaigns or weather forecasting: multivariate adaptive regression splines (MARS; Friedman 1991).

The aim of this paper is to compare, *using the same data sets*, two techniques commonly used in modelling species/communities distributions (LMR and CART) and a relatively new technique (MARS). We test their efficiency, accuracy, reliability, implementation within GIS and ease of use in ecological modelling studies.

Material and Methods

Test data sets

The two data sets employed were selected to cover a variety of characteristics. They are considered representative of typical ecological data sets, thus allowing results to be extrapolated to other studies. The data sets cover different geographical ranges (continental vs regional), spatial resolutions (coarse vs high), nature of the predictors (direct vs indirect) and sampling methods (no sampling strategy vs stratified sampling).

The ***Grimmia*** data set (1285 cases; 419 present, 866 absent) represents the distribution of species of the moss genus *Grimmia* in Latin America, from Mexico to Cape Hornos (Fig. 3). *Grimmia* was recently revised for Latin America and its taxonomy is well known worldwide (Muñoz 1999; Muñoz & Pando 2000). It is a genus typical of bare rocks in cool or cold regions, in intertropical regions only present at high altitudes (*Grimmia austrofunalis*, *G. navicularis* and *G. longirostris* recorded at 5300 m a.s.l. in the Andes). The dependent variable was presence/absence of plants of this genus at a given locality. The *Grimmia* data set was selected as an example to study model response at a continental scale. It was also selected because it is representative of many environmental studies: presences are based on museum collections, not on a sampling design and therefore suffer from biased sampling. Absences, not recorded in museums, were generated with a random grid in the GIS, and were approximately double the number of presences. Such proportion was chosen because with an equal number of absences/presences over such a large area, the model will tend to underestimate real absence. Randomly generated absences will inevitably increase the number of false negatives, but arising from the nature of the data this problem has no solution and will equally affect all statistical methods tested. Predictor variables were of three different types: altitude, climatic and normalized difference vegetation index (NDVI). Apart from altitude, they are all direct gradients, which can generate more general models (Guisan & Zimmermann 2000). Gridded altitude data were obtained from GTOPO30 (http://edcdaac.usgs.gov/gtopo30/gtopo30.html). Climatic data span the period 1961-1991 and were obtained from the Data Distribution Centre (DDC) of the Intergovernmental Panel on Climate Change (IPCC) at its web site, http://ipcc-ddc.cru.uea.ac.uk/. Each observation corresponds to the monthly 30-yr mean of the following variables: ground frost frequency ('frost', units: days·10), maximum temperature ('tmax', °C·10), minimum temperature ('tmin', °C·10), precipitation ('prec', 10·mm day^{-1} (=10 $L.m^{-2}.d^{-1}$)), radiation ('rad', $W.m^{-2}$) and wet day frequency ('wet', days·10). Climatic grids have a horizontal grid spacing of 0.5° (ca. 55 km at the Equator and 32 km at 55° S). NDVI was derived from the SPOT satellite remote sensing imagery (http://www.vgt.vito.be). We used four data sets, one per year season, extracted from a ten day global synthesis compiled from daily syntheses over the previous ten days at a resolution of 1 km (VGT-S10 SPOT products). NDVI was selected for its influence on factors such as albedo and heat exchange, as well as being a good descriptor of vegetation phenologic status or a surrogate for vegetation classes.

The ***Fagus*** data set (103 181 cases; ca. 50% each of presences and absences) was selected to represent high spatial resolution at a regional scale. The dependent variable was the presence/absence of *Fagus sylvatica* oligotrophic forest in the La Liébana region (Cantabria Province, NW Spain). A random sampling was performed in the GIS over a digitized vegetation map created at the Earth Sciences Department of the University of Cantabria (Spain). Equal numbers of samples were taken from the target forest type and the remaining vegetation classes. Predictor variables were derived from the digital elevation model generated from topographic maps, all of them have a horizontal grid spacing of 50 m. Altitude (constructed using Delaunay's triangulation algorithm), slope (derived using Sobel's operator; Horn 1981), potential insolation (constructed as a function of the sun's trajectory for standard date periods, which estimates the amount of time that each point receives direct sun radiation with 20 minutes temporal resolution; Fernández-Cepedal & Felicísimo 1987) and distance from the sea were used to estimate the oceanic-continental gradient given that other climatic data were not available. Complete details of this data set can be found in Felicísimo et al. (2002). Contrary to the *Grimmia* data set, all predictors except radiation must be considered indirect, which theoretically generate more local models.

Software included the GIS ArcInfo and ArcView 3.2 (ESRI Inc., http://www.esri.com/) and the packages SPSS 10.0 for LMR and Principal Component Regression, CART 4.0 (http://www.salford-systems.com) for Regression Trees and MARS 2.0 (http://www.salford-systems.com/) for multivariate adaptive regression splines.

Modelling methods

Methods used in predictive modelling consist of two main types: global parametric and local non-parametric.

Global parametric models use a strategy of global variable selection. Each variable enters the model as 'a whole' to explain its contribution to the response. This strategy is clearly inappropriate when the hypothesis is that variables interact in a non-homogeneous way across their range of values. However, global techniques are still appropriate for small data sets where the analyst is forced to use parametric modelling techniques because all points will influence almost every aspect of the model. As an example of global parametric model we have used LMR. Although widely used (e.g. Augustin et al. 1996; van de Rijt et al. 1996; Narumalani et al. 1997; Guisan et al. 1998; Mladenoff et al. 1999; Felicísimo et al. 2002), GLM and LMR have several important disadvantages. Ecologists frequently assume a unimodal and symmetric response to gradients, which real life obstinately tends to refute (Austin & Smith 1989; Yee & Mitchell 1991; also see Rydgren et al. in press). Such multi-modal or skewed distributions are sometimes dealt with using high-order polynomial functions, but this strategy heavily increases the risk of over-fitting – finding patterns that only apply to the training data – creating models that work almost perfectly with original data but have poor predictive ability with new data.

Secondly, in GLM the relationships between response and predictors are assumed to be linear, when real-world effects are generally more complex.

Our hypothesis is that in modelling organism/communities distributions, response is related to predictor variables in a non-linear and local fashion. *Local non-parametric models* are suitable under such a hypothesis as they use a strategy of local variable selection and reduction, and are flexible enough to allow non-linear relationships. From this type we have tested CART and MARS.

Classification and Regression Trees (CART; Breiman et al. 1984) is a rule based method that generates a binary tree through *binary recursive partitioning*, a process that splits a node based on yes/no answers about the values of the predictors. Each split is based on a single variable. Some variables may be used many times while others may not be used at all. The rule generated at each step maximizes the class purity within each of the two resulting subsets. Each subset is split further based on entirely different relationships. CART builds an overgrown tree based on the node purity criterion that is later pruned back via cross-validation to avoid over-fitting.

The main drawback of CART models, when used to predict organism distributions, is that with more than just a handful of predictor variables or cases to classify, the generated models can be extremely complex and difficult to interpret. This is exemplified by the work on Australian forests by Moore et al. (1991), generating a tree with 510 nodes for just ten predictors. In the present study, the optimal tree obtained for the *Fagus* data set (103 181 cases) has 1726 terminal nodes! Such complexity makes the tree impossible to interpret, whereas in many studies interpretability is a key issue. Moreover, implementation of such a tree within GIS is unworkable. Prediction maps are often a required outcome of modelling, and this shortcoming affects the use of CART when complexity grows beyond a reasonable limit.

Multivariate Adaptive Regression Splines (MARS; Friedman 1991) is a relatively novel technique that combines classical linear regression, mathematical construction of splines and binary recursive partitioning to produce a local model where relationships between response and predictors are either linear or non-linear. To do this, MARS approximates the underlying function through a set of adaptive piecewise linear regressions termed *basis functions* (BF) For example, the basis functions and the final function from the *Grimmia* model with 2nd. order interactions include mean number of frost days during April·10 (FROST04), mean number of frost days during August·10 (FROST08) and mean precipitation per day in April·10 (PREC04) (Fig. 3; below) are:

BF2 = max (0, 3318 – elevation);
BF3 = max (0, FROST04 – 67) * BF2;
BF4 = max (0, 67 – FROST04) * BF2;
BF6 = max (0, 6 – PREC04);
BF7 = max (0, FROST08 - 85) * BF2;

$$Y = 0.913 + 0.363496E\text{-}05 * BF3 - 0.449372E\text{-}05 * BF4 - 0.072 * BF6 - 0.271023E\text{-}05 * BF7 \quad (1)$$

Changes in slope of those basis functions occur at points called *knots* (values 3318, 67 and 6 in the above examples). The regression line is thus allowed to bend at the knots, which mark the end of one region of data and the beginning of another with different behaviour of the function (Fig. 1). Like splits in CART, knots are established in a forward/backward stepwise way. A model which clearly overfits the data is produced first. In subsequent steps, knots that contribute least to the efficiency of the model are discarded by backwards pruning steps. The best model is selected via cross-validation, a process that applies a penalty to each term (knot) added to the model to keep low complexity values.

Another key improvement of MARS over global parametric models is the way it deals with interactions. In local modelling, interactions can no longer be treated as global. MARS considers interactions not between the

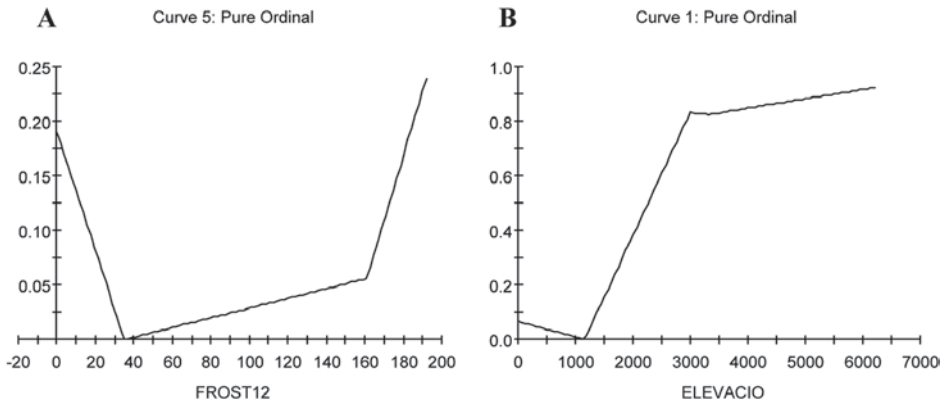

Fig. 1. Example of MARS functions showing the relationship between response and a single predictor: **A.** Effect of Frost12 (mean number of frost days during December·10) on the response. This almost flat-bottom function with three very different slope regions is difficult to model under the assumption of a linear relationship, because the positive and negative sloping sections cancel each other out. **B.** The basis function has a negative slope from sea level to 1157 m a.s.l., then ascends tol 3318 m a.s.l. where it reaches a semi-plateau value.

original predictors, but between the sub-regions of every basis function generated. A particular sub-region of a given basis function can interact with a particular sub-region of another basis function, but other regions of these basis functions might display none or a different interaction pattern. Examples are shown in Fig. 2, which illustrate the contribution to the response by interactions specified in Table 1.

Interactions among predictors

Common regression analysis rapidly becomes unreliable when dimensionality becomes high, a phenomenon known as the 'curse of dimensionality' (Hastie & Tibshirani 1990). An immediate consequence of the 'curse of dimensionality' is the severe limitation in the number of usable variables in a given analysis. Suppose we consider two regions for each predictor. With two predictors the number of regions to be considered will be four; with three predictors they will be eight (2^3) etc. In the present study, and limiting the number of regions to 2 – a number we do not know *a priori* –, the number of regions will be 32 with *Fagus* (five predictors), but $15{\cdot}10^{22}$ with the *Grimmia* data set (77 predictors) which is simply not feasible.

Table 1. MARS sample output for the *Grimmia* data set with 2nd. order interactions. See p. 3 for explanation of variables.

Basis function	Coefficient	Variable	Parent	Knot
0	0.913			
3	3.63E-006	FROST04	altitude	67
4	– 4.49E-006	FROST04	altitude	67
6	– 0.072	PREC04		6
7	– 2.71E-006	FROST08	altitude	85

Most statistical techniques perform poorly with high dimensionality, a problem that in predictive modelling literature is usually circumvented by limiting the number of variables employed through *a priori* selection, which is not always biologically warranted, or by eliminating the interactions from the model, as in Yee & Mitchell (1991). Some models for acceptable selection of predictors have been proposed, including the use of CART to identify the interactions, then fit GLM or GAM based on those variables (Guisan et al. 2002).

To compare the performance of each technique with regard to interactions, all statistical analyses were run with no interactions (i.e. main effects) and with interactions up to 5th order in MARS (when the results start to deteriorate). The same interactions that produced the best model in MARS were then used to run a new LMR analysis to compare against MARS performance.

Dealing with multicollinearity

Multicollinearity occurs when one or more variables are exact or near exact linear functions of other variables in the data set. This is a common problem associated with organism/communities modelling (Brown 1994; De Veaux & Ungar 1994). Assessment of model performance when multicollinearity may be an issue was done by including the following analyses (all LMR are forward conditional stepwise LMR with *P*-to-enter = 0.05 and *P*-to-remove = 0.1, no polynomial functions used due to the high number of predictors): LMR with original variables; LMR with Varimax orthogonalized PCA factors (Principal Component Regression); LMR using as variables the basis functions generated in the MARS main effects (i.e. no-interactions) analysis (five basis functions in *Grimmia* data set and ten in *Fagus*

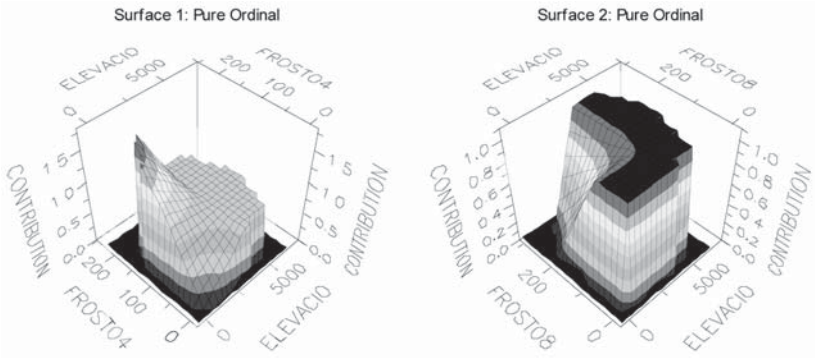

Fig. 2. Example of MARS functions obtained from the *Grimmia* data set with 2nd. order interactions among predictors, showing the contribution to the response from such interactions: **A**. Altitude and Frost04 (mean number of frost days during April·10) only interact below 3318 m a.s.l., but not above this elevation. **B**. Altitude only interacts with Frost08 (mean number of frost days during August·10) below 3318 m a.s.l. and then only when there are less than 21 frost days.

data set); CART with original variables; CART with the same PCA factors as above; CART with the same MARS basis functions as above and MARS with the original variables.

We used two LMR projection methods to avoid multicollinearity problems: principal component regression (PCR), already used in predictive modelling (see references in Guisan & Zimmermann 2000) and LMR using as variables the basis functions generated in the MARS main effects analysis. PCR uses the projections of the predictor variables onto a subset of PCs in place of the predictor variables themselves. Since the PCs are linearly uncorrelated (Varimax rotation), there are no multicollinearities between the projection coefficients. To our knowledge, MARS basis functions has not previously been combined with LMR to circumvent multicollinearity.

Evaluation of the models

Assessment of model performance was done by a method that was independent of any threshold: the area under the Receiver Operating Characteristic (ROC) curve, commonly termed AUC. The ROC curve is recommended for comparing two class classifiers, as it does not merely summarize performance at a single arbitrarily selected decision threshold, but across all possible decision thresholds (Fielding & Bell 1997). It plots the sensitivity (i.e. true positives) vs (1– specificity) (i.e. false positives). An ideal classifier hugs the left side and top side of the graph and the area under the curve equals one. A random classifier should achieve ca. 0.5. Whilst most analyses use 0.5 as the decision threshold to consider a case as present or absent, this value is arbitrary and it does not necessarily give a more accurate model (Fielding & Bell 1997; Manel et al. 1999). AUC eliminates this problem by considering the performance of the model across all possible threshold values. Moreover, ROC curves are invariant under changing distributions of the binary classes (presence/absence), as they actually plot the 'percentage of class-1 observations' vs 'percentage of class-0 observations' and are therefore independent of the balance between the two classes. Hanley & McNeil (1983) have shown that when dealing with a single scoring model, the AUC is equal to the empirically observed probability of a class-1 observation attaining a higher score than a class-0 observation. They have also shown that the AUC is actually equivalent to the normalized Mann-Whitney two-sample statistic, which makes it equivalent to the Wilcoxon statistic.

AUC is a measure of model accuracy, but it does not provide a rule for the classification of cases (Fielding & Bell 1997). The final decision about which threshold should be selected in a particular study depends upon its objectives. The relative importance of FP (False positives, Type I) and FN (False negatives, Type II) error rates must be individually considered in each study, and this decision is independent of model accuracy. If the purpose of the study is to identify sites where we need to be certain that an organism will be found, we must select the threshold that minimizes FP error rates. Contrarily, if the aim is conservation of the same organism, the threshold must be chosen to minimize FN error rates.

Finally, and in order to have a reliable estimate of the prediction power of each model, we use two approaches:

1. *Tenfold cross-validation*: The original data set is randomly divided into ten mutually exclusive subsets (the folds) of about equal size. A subset is removed and the remaining cases are used to generate a model. This model is afterwards applied to the removed section and its empirical error calculated. The process is repeated with the remaining nine subsets and the mean empirical error is used as the final estimate of the total error of the model. We report results obtained by applying the optimal models to the entire (i.e. learning) data sets.

2. *Use of independent data sets for training and evaluation*: Both the *Grimmia* and *Fagus* data sets were split randomly into one *training* and one *evaluation* data set, containing approximately 70% and 30% of the total cases, respectively. The training data sets were used to generate the models which were then tested with the independent evaluation data sets.

Table 2. Results with the *Grimmia* data set (best-fit model in bold). LMR = Logistic Multiple Regression; PCA = Principal Component Analysis; CART = Classification and Regression Trees; AUC = Area under ROC curve; MARS = Multiple Adaptive Regression Splines; CI = Confidence interval (95%).

	Evaluation by ten-fold cross-validation		Evaluation by independent data set	
Model	AUC	CI	AUC	CI
LMR with original variables	0.960	0.951-0.970	0.953	0.933-0.972
LMR with original variables, same interactions as MARS best model	0.963	0.954-0.972	0.951	0.931-0.971
LMR with PCA factors	0.892	0.874-0.910	0.879	0.844-0.914
LMR with MARS BFs	0.939	0.926-0.951	0.942	0.920-0.963
CART with original variables	0.970	0.960-0.980	0.912	0.876-0.947
CART with PCA factors	0.950	0.937-0.962	0.845	0.801-0.889
CART with MARS BFs	0.956	0.945-0.967	0.927	0.898-0.957
MARS with original variables, no interactions	0.977	0.970-0.984	0.931	0.899-0.964
MARS with PCA factors	0.933	0.919-0.946	0.897	0.865-0.928
MARS with original variables 2nd. order interactions	0.957	0.947-0.967	**0.954**	**0.930-0.978**
MARS with original variables 3rd. order interactions	0.980	0.974-0.987	0.945	0.915-0.974
MARS with original variables 4th. order interactions	0.982	0.976-0.988	0.939	0.908-0.969

Results

Model performance

Tables 2 and 3 present the results of the different statistical techniques applied to the *Grimmia* and *Fagus* data sets, respectively. For both data sets differences between evaluation via 10-fold cross-validation and independent data are evident. Confidence intervals rarely overlap between the two approaches, confirming the results of Manel et al. (1999) that evaluation via independent test samples is needed. The following discussion therefore applies to the evaluation by the independent data set.

With the *Grimmia* data set the best performance was achieved by the MARS with 2nd. order interactions model (Fig. 3), with AUC = 0.954, although LMR results were very similar. Improvement of MARS over LMR is evident for *Fagus*, with an increase from 0.778 to 0.909 in AUC. In this data set, however, the best results were obtained with CART (AUC = 0.946). Unfortunately, the complexity of the tree which generates such results, with 1726 terminal nodes, makes this model inadequate for cartographic purposes, a frustration already pointed out by Guisan & Zimmermann (2000). The MARS model, very similar in performance to the CART model, can then be considered as a surrogate mapping technique (Fig. 4).

Interactions among predictors

The best results were achieved with models allowing for higher order interactions. MARS was better than CART in that the models generated were easier to interpret, and interactions between predictors could be more easily understood and explained in terms of their biological relevance. In the particular case of the *Grimmia* data set, Fig. 2 shows that between 'Elevation' and 'Frost04' (Fig. 2A) or 'Frost08' (Fig. 2B) there is interaction only below 3318 m a.s.l. and not above this altitude, where in both cases Contribution = 1. Fig. 2A should be interpreted as that in South America, below 3318 m a.s.l., contribution of the interaction to the model (Eq. 1) increases linearly above 6.7 frost days in April (basis function BF3, units are days·10), according to the idea of *Grimmia* as a genus typical of cool or cold regions. Similarly, Fig. 2B shows that contribution to the model (Eq. 1) equals one below 3318 m a.s.l. only when frost days in August are less than 8.5 (basis function BF7, units are days·10). As BF7 enters negatively in the model (Eq. 1, coefficient – 0.271023E-05), it means that probability of finding *Grimmia* increases when the number of frost days in August increases.

Multicollinearity

Tables 2 and 3 show that the projection methods used to circumvent multicollinearity, which *a priori* should outperform the methods using the original predictors (De Veaux & Ungar 1994), always reduced predictive power. This was an unexpected result, because the data sets included a large number of predictors that might intuitively be considered highly correlated. The reason of this apparent contradiction is that interactions among predictors, demonstrated in the previous section, are masked when using projected variables (De Veaux & Ungar 1994).

Table 3. Results with the *Fagus* data set (best-fit model in bold). LMR = Logistic Multiple Regression; PCA = Principal Component Analysis; CART = Classification and Regression Trees; AUC = Area under ROC curve; MARS = Multiple Adaptive Regression Splines; CI = Confidence interval (95%). [a] Model used to generate the map shown in Fig. 4.

	Evaluation by ten-fold cross-validation		Evaluation by independent data set	
Model	AUC	CI	AUC	CI
LMR with original variables	0.793	0.790-0.796	0.778	0.772-0.783
LMR with PCA factors	0.755	0.752-0.758	0.757	0.751-0.762
LMR with MARS BFs	0.898	0.896-0.900	0.897	0.894-0.901
CART with original variables	0.979	0.978-0.980	**0.946**	**0.943-0.949**
CART with PCA factors	0.805	0.802-0.808	0.794	0.789-0.799
CART with MARS BFs	0.976	0.975-0.977	**0.946**	**0.943-0.949**
MARS with original variables	0.896	0.894-0.898	0.909[a]	0.906-0.912
MARS with PCA factors	0.792	0.790-0.795	0.789	0.784-0.794
MARS with original variables 2nd. order interactions	0.907	0.905-0.909	0.906	0.903-0.912

Discussion

Our results show that distributions of species and communities can be better defined if we accept that they may follow multivariate non-linear mappings. At present there are many sources of predictors which may interact or correlate with each other in an unknown fashion, especially if we are considering remote sensing data. Simplifications, both in the number of predictors employed or assuming linear relationships with the response, may render misleading models that perform poorly on real world data. Our results support the view that global parametric methods are inferior in such modelling studies.

LMR has been successfully used in predictive modelling to predict distributions of species and communities, despite its drawbacks such as its inability to deal with skewed or multi-modal responses. CART, which may produce better numerical predictions on new data, generates complex models that can lead to no or spurious model interpretations, and is occasionally of no use for cartographic purposes. One goal of predictive modelling is to generate 'potential habitat distribution maps'. According to Guisan & Zimmermann (2000), such maps are cartographic representations of (1) the probability of occurrence; (2) the most probable abundance; (3) the predicted occurrence; or (4) the most probable entity. Cartographic implementation is therefore crucial.

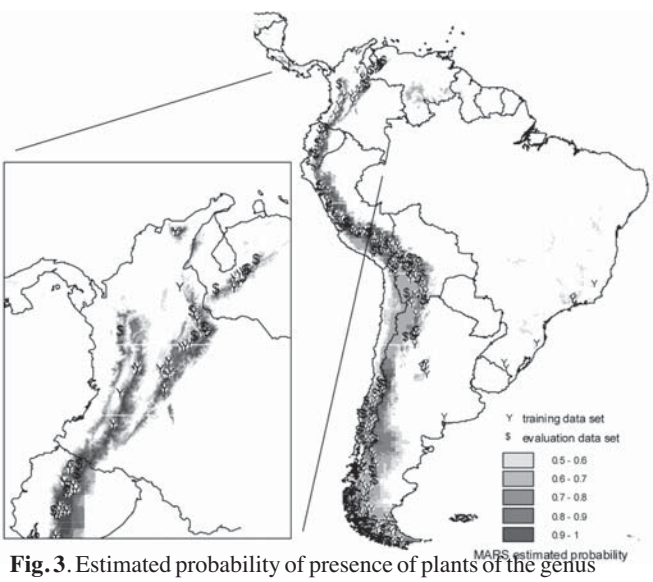

Fig. 3. Estimated probability of presence of plants of the genus *Grimmia* in South America according to MARS with 2nd. order interactions model. Inset shows in detail the northern tip of the Andean range.

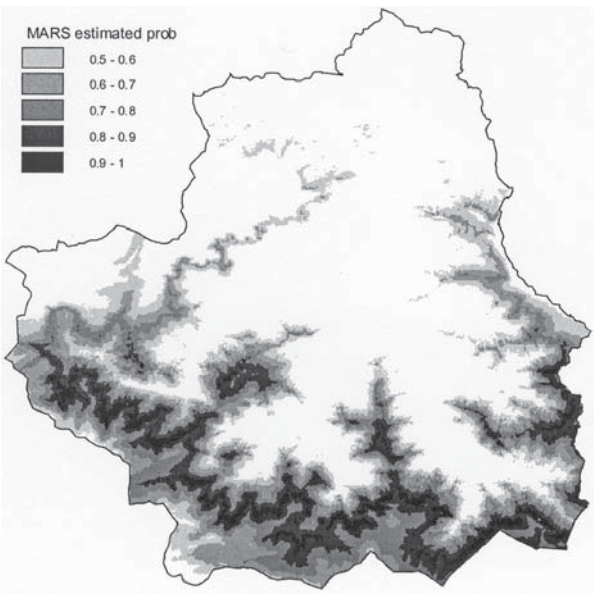

Fig. 4. Predicted map representing the probability of occurrence of oligotrophic forest with *Fagus sylvatica* in La Liébana (NW Spain).

Contrary to all of these disadvantages, MARS is better suited to model situations that include a high number of variables, non-linearity, multicollinearity and/or a high degree of interaction among predictors. MARS has been shown to perform as well as and more consistently than other methods with two very different data sets. Moreover, it is extremely easy to implement within GIS, and we conclude that it can be seen as an alternative that extends the use of Generalized Methods and avoids unsupported inferences derived from uncritical use of standard procedures.

References

Augustin, N.H., Mugglestone, M.A. & Buckland, S.T. 1996. An autologistic model for the spatial distribution of wildlife. *J. Appl. Ecol.* 33: 339-347.

Austin, M.P. & Cunningham, R.B. 1981. Observational analysis of environmental gradients. *Proc. Ecol. Soc. Aust.* 11: 109-119.

Austin, M.P. & Smith, T.M. 1989. A new model for the continuum concept. *Vegetatio* 83: 35-47.

Austin, M.P., Nicholls, A.O. & Margules, C.R. 1990. Measurement of the realized qualitative niche: environmental niche of five *Eucalyptus* species. *Ecol. Monogr.* 60: 161-177.

Austin, M.P., Nicholls, A.O., Doherty, M.D. & Meyers, J.A. 1994. Determining species response functions to an environmental gradient by means of a beta-function. *J. Veg. Sci.* 5: 215-228.

Breiman, L., Friedman, F., Olshen, R. & Stone, C. 1984. *Classification and regression trees*. Wadsworth, Pacific Grove, CA, US.

Brown, D.G. 1994. Predicting vegetation types at treeline using topography and biophysical disturbance variables. *J. Veg. Sci.* 5: 641-656.

De Veaux, R.D. & Ungar, L.H. 1994. Multicollinearity: A tale of two nonparametric regressions. In: P. Cheeseman, P. & Oldford, R.W. (eds.) *Selecting models from data: AI and Statistics IV*, pp. 293-302. Springer-Verlag. New York, NY, US.

Felicísimo, A.M., Francés, E., Fernández, J.M., González-Díez, A. & Varas, J. 2002. Modeling the potential distribution of forests with a GIS. *Photogramm. Engin. Remote Sens.* 68: 455-461.

Fernández-Cepedal, G. & Felicísimo, A.M. 1987. Método de cálculo de la radiación solar incidente en áreas con apantallamiento topográfico. *Rev. Biol. Univ. Oviedo* 5: 109-119.

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

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

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

Guisan, A., Theurillat, J.-P. & Kienast, F. 1998. Predicting the potential distribution of plant species in an alpine environment. *J. Veg. Sci.* 9: 65-74.

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

Hanley, J.A. & McNeil, B.J. 1983. A method of comparing the areas under Receiver Operating Characteristic curves derived from the same cases. *Radiology* 148: 839-843.

Hastie, T.J. & Tibshirani, R.J. 1990. *Generalized additive models*. Chapman & Hall, London, UK.

Horn, B.K.P. 1981. Hill-shading and the reflectance map. *Proc. Inst. Electr. Electron. Engin.* 69: 14-47.

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

Manel, S., Dias, J.-M. & Ormerod, S.J. 1999. Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. *Ecol. Model.* 120: 337-347.

Mladenoff, D.J., Sickley, T.R. & Wydeven, A.P. 1999. Testing a predictive landscape model of favorable gray wolf habitat: logistic regression models vs. new field data. *Ecol. Appl.* 9: 37-44.

Moore, D.M., Lee, B.G. & Davey, S.M. 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system. *Environ. Manage.* 15: 59-71.

Muñoz, J. 1999. A revision of *Grimmia* (Musci, Grimmiaceae) in the Americas. 1: Latin America. *Ann. Mo. Bot. Gard.* 86: 118-191.

Muñoz, J. & Pando, F. 2000. A world synopsis of the genus *Grimmia* (Musci, Grimmiaceae). *Monogr. Syst. Bot. Mo. Bot. Gard.* 83: 1-133.

Narumalani, S., Jensen, J.R., Althausen, J.D., Burkhalter, S. & Mackey, H.E. 1997. Aquatic macrophyte modeling using GIS and logistic multiple regression. *Photogramm. Engin. Remote Sens.* 63: 41-49.

Rydgren, K., Økland, R.H. & Økland, T. 2003. Species response curves along environmental gradients. A case study from SE Norwegian swamp forests. *J. Veg. Sci.* 14: 869-880.

van de Rijt, C.W.C.J., Hazelhoff, L. & Blom, C.W.P.M. 1996. Vegetation zonation in a former tidal area: A vegetation-type response model based on DCA and logistic regression using GIS. *J. Veg. Sci.* 7: 505-518.

Yee, T.W. & Mitchell, N.D. 1991. Generalized additive models in plant ecology. *J. Veg. Sci.* 2: 587-602.

Received 8 October 2002;
Accepted 17 November 2003.
Co-ordinating Editor: R.H. Økland.