Diversity and Distributions, (Diversity Distrib.) (2007) **13**, 265–275

BIODIVERSITY RESEARCH

Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines

Jane Elith[1]* and John Leathwick[2]

[1]*School of Botany, The University of Melbourne, Parkville, Victoria, Australia 3010,* [2]*National Institute of Water and Atmospheric Research, PO Box 11115, Hamilton, New Zealand*

*Correspondence: Jane Elith, School of Botany, The University of Melbourne, Parkville, Victoria 3010, Australia. E-mail: j.elith@unimelb.edu.au

ABSTRACT

Current circumstances — that the majority of species distribution records exist as presence-only data (e.g. from museums and herbaria), and that there is an established need for predictions of species distributions — mean that scientists and conservation managers seek to develop robust methods for using these data. Such methods must, in particular, accommodate the difficulties caused by lack of reliable information about sites where species are absent. Here we test two approaches for overcoming these difficulties, analysing a range of data sets using the technique of multivariate adaptive regression splines (MARS). MARS is closely related to regression techniques such as generalized additive models (GAMs) that are commonly and successfully used in modelling species distributions, but has particular advantages in its analytical speed and the ease of transfer of analysis results to other computational environments such as a Geographic Information System. MARS also has the advantage that it can model multiple responses, meaning that it can combine information from a set of species to determine the dominant environmental drivers of variation in species composition. We use data from 226 species from six regions of the world, and demonstrate the use of MARS for distribution modelling using presence-only data. We test whether (1) the type of data used to represent absence or background and (2) the signal from multiple species affect predictive performance, by evaluating predictions at completely independent sites where genuine presence–absence data were recorded. Models developed with absences inferred from the total set of presence-only sites for a biological group, and using simultaneous analysis of multiple species to inform the choice of predictor variables, performed better than models in which species were analysed singly, or in which pseudo-absences were drawn randomly from the study area. The methods are fast, relatively simple to understand, and useful for situations where data are limited. A tutorial is included.

Keywords

MARS, presence-only, multiresponse, habitat model, pseudo-absence, community.

INTRODUCTION

Species, communities, and ecosystems are distributed across the Earth in interesting and complex patterns. These stimulate scientific research not only for their own sake (Whittaker, 1967; Levin, 1992; Graham *et al*., 2004a) but also because widespread clearing and other disturbances threaten the existence of many species and ecosystems (Margules & Pressey, 2000). Species distribution models are one of several methods for quantifying these patterns, relying on the association between a species' occurrence or abundance and environmental or geographical predictors (see reviews by Guisan & Zimmerman, 2000 and Araújo & Guisan, 2006). Distribution models (also called 'habitat models') now have an established place within conservation biology, where they inform survey design, strategic reserve placement, biosecurity risk assessment, and identification of suitable restoration sites (Funk & Richardson, 2002; Goolsby, 2004; Edwards *et al*., 2005).

While species data from planned surveys that describe species presence–absence or abundance are ideal for modelling distributions (Cawsey *et al*., 2002), records for most species of the world are in a 'presence-only' form. These are usually derived from ad-hoc compilations of observations that lack reliable records of species' absence (Dennis & Hardy, 1999; Ferrier *et al*., 2004; Graham *et al*., 2004b). Given both the prevalence of such data and the

DOI: 10.1111/j.1472-4642.2007.00340.x
www.blackwellpublishing.com/ddi

urgent need to manage and conserve biodiversity, there has been an increasing focus on their use in conservation planning — not only for estimating species' ranges, patterns in richness, and biodiversity hotspots, but also for making predictive maps of species' occurrence or richness (e.g. Robertson *et al.*, 2003; Brotons *et al.*, 2004; Hortal *et al.*, 2004; Phillips *et al.*, 2004). These maps are created at many different spatial scales that variously inform broad-scale issues including climate change and nationwide inventory (Araújo *et al.*, 2004), and finer-grained practices of land management and acquisition of land for reserves (Ferrier, 2002). Early modelling attempts used methods specifically designed to deal with presence-only data (e.g. climate envelopes, multivariate measures — Busby, 1991; Carpenter *et al.*, 1993), but more recent research has focused on adapting presence–absence methods such as logistic regression to the presence-only paradigm, because of their greater predictive ability (Elith *et al.*, 2006; Pearce & Boyce, 2006).

A range of issues are relevant to modelling presence-only records including identifying and ameliorating biases (Freitag *et al.*, 1998; Dennis & Hardy, 1999; Hortal & Lobo, 2005), understanding and dealing with errors in identification and spatial location (Wieczorek *et al.*, 2004; Frey, 2006), choosing appropriate grain sizes for analysis (McPherson *et al.*, 2006; Guisan *et al.*, 2007), selecting appropriate modelling methods (Elith *et al.*, 2006; Pearce & Boyce, 2006), and dealing with sparse records and lack of absences (Ferrier, 2002; Engler *et al.*, 2004). In this paper we focus on the interplay between a relatively new regression method for predicting species distributions, i.e. Multivariate Adaptive Regression Splines (MARS, Friedman, 1991; Hastie *et al.*, 1994, 2001; Leathwick *et al.*, 2005), and two methods for selecting absences. MARS is interesting because it gives comparable predictive performance to other nonlinear regression methods — such as generalized additive models (GAMs) — that have been shown to be useful for modelling ecological data, yet MARS is much faster than GAMS (Moisen & Frescino, 2002; Leathwick *et al.*, 2005). In addition, results are easily transferred into a Geographic Information System (GIS) for mapping. Perhaps most importantly, the data from multiple species can be used to inform model development for a target species, a feature achieved with a 'multiresponse' model (Hastie *et al.*, 1994), that is potentially useful for species with few records.

In a recent large international trial, MARS multiresponse[1] models performed particularly well for predicting occurrence patterns in independent data sets (Elith *et al.*, 2006). Here, we investigate what drives the improvement in performance over related single-species models. There are two possibilities. The first is choice of absences. Regression-based implementations of single-species models often use random pseudo-absences — i.e. random samples of the region in geographical space (e.g. Zaniewski *et al.*, 2002). However, in constructing a MARS multiresponse model with presence-only records, if a site is visited and a species is not recorded there, then analytically this site is treated as an 'inventory pseudo-absence' for that species. Although this approach to absence selection is uncommon (but see Lütolf *et al.*, 2006; Ferrier *et al.*, 2007), consideration of the nature of presence-only samples indicates that it may have strong practical advantages. Second, the improvement provided by MARS multiresponse models may be driven by effects on variable selection. Models that use simultaneously the signal from several to many species in selecting predictor variables may be more robust for prediction because data-rich species can help to inform models for data-poor species (Ferrier & Guisan, 2006). That is, relevant predictors are included because of their strong signal across all species, whereas that signal might be insufficient to trigger inclusion in single-species models (Leathwick *et al.*, 2005). Multiresponse models are effective for presence–absence data (Olden, 2003; Leathwick *et al.*, 2006) and conceptually are likely to be useful for presence-only data, in which some species may be poorly represented.

Our goal here is to assess the predictive performance of models for 226 species from six study areas, assessed with independent and separately collected presence–absence (PA) data sets in the same regions from which the presence-only data were compiled. This indicates whether the resulting predictive maps are accurate enough to be useful in conservation planning at regional scales, and gives insights into questions relevant to modelling presence-only data with regression-based methods.

METHODS

Data for modelling and evaluation

Presence-only data for model fitting were collated for 226 species from six regions of the world — birds and plants of the Australian Wet Tropics (AWT); birds of Ontario, Canada (CAN); plants, birds, mammals and reptiles of north-east New South Wales, Australia (NSW); plants of New Zealand (NZ); plants from five countries of South America (SA); and plants of Switzerland (SWI). Species data were mostly drawn from natural history collections, in which numbers of records per species varied from few (tens to hundreds of presence records) to many (tens of thousands of records; Table 1). MARS, in common with many other methods, requires data akin to absences for modelling species distributions, and for these we took a random sample of 10,000 sites to characterize the 'background'. These are our 'random pseudo-absences' and we compare them with 'inventory pseudo-absences' that are based only on sites that have been visited but at which the target species does not occur. The designation of the latter was partly dictated by the needs of the MARS multiresponse analysis, which requires a site by species matrix for fitting models. For this, we compiled the presence records for all species in a given biological group (e.g. 'plants' or 'birds'), and reduced them to the set of unique sites. These form the rows of the matrix, and each column represents a species, assigned a value of 'one' if the species was recorded at that site, and 'zero' otherwise. The zeros are therefore the 'inventory pseudo-absences'. The environmental data used for each region were selected for their relevance to the species being modelled, as selected by the data provider (Table 1, and see Elith *et al.*, 2006 for more detail).

[1]Previously we have called these 'community' models but now prefer the term 'multiresponse', as used in the statistical literature (Hastie *et al.*, 1994).

Table 1 Summary of data available for modelling and evaluation (adapted from Elith *et al.*, 2006).

	Species records			Predictor variables offered in MARS modelling	
Region	Biological groups (number of species)	PO: mean number (range); number of unique sites	PA: number sites; mean no. of pres	Broad class	Cell size (m) and Extent ($km^2 \times 10^6$)
Australian Wet Tropics (AWT)	Birds (20)	155 (32–265); 714	340; 97	5 climate 3 topography	80 m × 80 m 0.024
	Plants (20)	35 (9–74); 379	102; 30		
Ontario, Canada (CAN)	Birds (20)	255 (16–749); 3298	14571; 1282	2 climate 3 topography 1 distance	1 km × 1 km 1.088
New South Wales, Australia (NSW)	Birds (10)	162 (48–426); 1351	Mean 920; 74	4 climate 2 soil 1 moisture 3 topography 1 disturbance	100 m × 100 m 0.089
	Plants (29)	22 (2–69); 569	Mean 1333; 214		
	Mammals (7)	27 (6–49); 147	570; 76		
	Reptiles (8)	84 (34–168); 530	1008; 62		
New Zealand (NZ)	Plants (52)	59 (18–211); 2503	19120; 1801	7 climate 2 substrate 2 topography	100 m × 100 m 0.265
South America* (SA)	Plants (30)	74 (17–216); 1221	152; 12	8 climate	1 km × 1 km 14.654
Switzerland (SWI)	Plants (30)	1170 (36–5822); 11429	10013; 810	6 climate 2 substrate 2 topography 2 vegetation	100 m × 100 m 0.041

PO is the presence-only modelling data and PA, the presence–absence evaluation data.
*Five countries: continental Brazil, Ecuador, Colombia, Bolivia, and Peru.

Eleven to 13 predictors were supplied per region but these were reduced where necessary according to pairwise correlations (see later). Grid cell sizes ranged from about 100 m × 100 m (AWT, NSW, NZ, SWI) to 1000 m × 1000 m (CAN, SA), and these have been demonstrated elsewhere to be appropriate for modelling these data (Guisan *et al.*, 2007). For several regions previous research informed the development of ecologically relevant predictors (NSW, NZ, SWI). For other regions, variables typically used in distribution modelling were utilized, with emphasis on climatic data (CAN, SA, AWT).

Model fitting

MARS is a method of flexible nonparametric regression modelling (Friedman, 1991). It can model complex, nonlinear relationships between response and explanatory variables with similar levels of complexity to that of a GAM (Hastie, 1991). The MARS approach to fitting nonlinear functions is to fit linear segments — also called piecewise linear basis functions — to the data. MARS breaks the range of each predictor variable into subsets of the full range using 'knots', and allows the slope of the fitted linear segments between pairs of knots to vary while ensuring that the full fitted function is without breaks or sudden steps. In other words, a nonlinear MARS function consists of a series of connected straight line segments, rather than the smooth curve of a GAM. Model fitting is achieved with a very fast procedure that starts with forward steps that identify many knots, followed by a backward pruning routine to simplify the model. Additions and deletions are evaluated in terms of changes in residual squared errors using generalized cross-validation (GCV). As the available algorithms for MARS only accommodate normal error terms, we followed Friedman (1991) in adapting the model for presence–absence responses by fitting the MARS model, extracting its basis functions, and fitting these as predictor variables within a generalized linear model (GLM) with a binomial error distribution (Leathwick *et al.*, 2005). This ensured that the predictions were constrained between 0 and 1, but otherwise equated to a MARS model. All models were constructed using the free statistical software, R, version 2.1.1 (R Development Core Team, 2004), with the *mda* library and additional custom code by the authors (code and brief tutorial available online). Further statistical details are available in Friedman (1991), Hastie *et al.* (2001), and Leathwick *et al.* (2005). While MARS is also capable of automatically fitting interactions between predictors, we did not test that capability here. Our previous testing has failed to demonstrate that this provides any significant increase in predictive performance (Leathwick *et al.*, 2005; Elith *et al.*, 2006), though this feature may be useful in specific circumstances (Mark Lethbridge, pers. comm.).

The MARS single-species models with random pseudo-absences were fitted with the 10,000 background absences weighted so that the total weight for presences equalled the total weight for absences. The weighting allowed the use of many pseudo-absences that sample the environmental space of the region thoroughly, while avoiding 'swamping' the model with so much absence data that trends in presence were hard to detect.

One feature of the R version of MARS that has been rarely investigated is its ability to use data from multiple species (but see Leathwick *et al.*, 2005, 2006). Multiresponse models are built and pruned in exactly the same way as a single-response MARS model, except that the residual squared errors are averaged across all response variables (here, all species), with individual basis functions selected that give the best average improvement in performance (Hastie *et al.*, 1994). The final multiresponse model uses a common set of basis functions for all species, but estimates a different set of coefficients for each species, so that the shapes of the fitted functions can differ between species. As explained above, the nature of a multiresponse model requires the use of inventory pseudo-absences, and therefore no multiresponse models were fitted with random pseudo-absences. We investigate here whether the multiresponse models have a better predictive performance than MARS models developed individually for each species on the same (inventory-based) data.

In summary, we ran three sets of models: (1) single-species MARS models fitted on presence records and using random pseudo-absences; (2) single-species MARS models fitted on presence records and inventory pseudo-absences, with the latter constructed using data for the biological groups listed in Table 1; and (3) multiresponse MARS models fitted on presence records and inventory pseudo-absences. Note that these multiresponse models differ from those presented in Elith *et al.* (2006) due to the latter's use of both inventory pseudo-absences as used here, and random pseudo-absences added to the site by species matrix to achieve consistency with other analyses in that study.

In all cases we reduced the candidate predictor variables to those with pairwise Pearson correlations of less than 0.85 (Elith *et al.*, 2006). Categorical predictors were excluded because the Windows R implementation of MARS does not allow for them, although subsequently we have written code to fit them. Plots of fitted functions and estimates of the contribution of each variable to model fit based on changes in deviance were produced using custom-written functions available online (and see Leathwick *et al.*, 2005, 2006).

Model evaluation

The predictive performance of models was evaluated with independent presence–absence (PA) data. We accessed the most accurate and well-planned collections available for each region, and these contained anywhere from a few hundred to tens of thousands of records from sites that were comprehensively surveyed (Table 1). To compare the predictions to the observations of presence and absence, we use two statistics: the area under the receiver operating characteristic curve, AUC (Hanley & McNeil, 1982; Fielding & Bell, 1997), and the deviance. AUC has been used extensively in evaluating species' distribution models, and measures the ability of a model to discriminate between sites where a species is present, vs. those where it is absent. A score of 0.5 indicates that a model has no discriminatory ability, while a score of 1 indicates that presences and absences are perfectly discriminated. AUC values can be interpreted as indicating the probability that, when a presence site and an absence site are drawn at random from the population, the first will have a higher predicted value than the second. It is a rank-based statistic — the prediction at the presence site can be higher than the prediction at the absence site by a small or a large amount, and the value of the statistic will be the same. Deviance complements AUC because it expresses the magnitude of the deviations of the fitted values from the observations, and was estimated as the mean deviance per observation. As it was calculated on independent evaluation data, we refer to it hereafter as 'predictive deviance'. Deviance has only a limited applicability for presence-only evaluation, because it is affected by the calibration of the models (i.e. how accurately predictions match the response, Pearce & Ferrier, 2000) and we do not expect presence-only models to be properly calibrated due to the lack of true absence data. It should therefore not be used independently of other measures for presence-only models because results could be misleading. Furthermore, it needs to be interpreted with reference to the calibration of the models. Both AUC and deviance deliberately use the full information in the predictions (i.e. using the predicted relative likelihoods that range from 0 to 1), rather than converting the predictions to a presence/absence estimate with a threshold (Liu *et al.*, 2005). This is based on our understanding that the full information is useful in conservation planning. Nevertheless, elsewhere related data have been analysed with kappa statistics derived from the confusion matrix (Elith *et al.*, 2006), and the results are consistent with those based on AUC.

Variation in AUC and predictive deviance across all models was summarized using Generalized Linear Mixed Models (GLMMs), with the measure of predictive performance (AUC or predictive deviance) as the response. *Model* (the three combinations of data and model construct) was fitted as a fixed effect, and *species* and an interaction between *model* and *region-group* were fitted as random effects, the interaction term allowing for differing performance of models across region-groups. GLMM analyses were performed using WINBUGS (Spiegelhalter *et al.*, 2003a), which fits a Bayesian model. We assumed uninformative priors for all parameters, resulting in a GLMM that is equivalent to one fitted using maximum likelihood. Comparisons of the three models were based on 50,000 Monte Carlo iterations after a burn-in period of 10,000. The performance of *model* was summarized as the median and 2.5% and 97.5% credible intervals of the posterior distributions. The importance of each term in the GLMM was assessed by change in the Deviance Information Criterion (DIC, Spiegelhalter *et al.*, 2003b) for the full GLMM compared with subsets where each term was excluded from the GLMM. The DIC is the Bayesian equivalent of Akaike's Information Criterion, and rules of thumb suggest that changes in DIC of more than 10 units indicate that the excluded term had an important effect (Burnham & Anderson, 2002; McCarthy & Masters, 2005).

Table 2 Predictive performance for the three combinations of model type and data.

Model/Data	Single/random	Single/inventory	Community/inventory
Predictive deviance (per observation)	1.701 (1.357, 2.069)	0.935 (0.580, 1.300)	0.841 (0.480, 1.200)
AUC	0.684 (0.650, 0.717)	0.724 (0.691, 0.757)	0.737 (0.702, 0.771)

Values are modelled medians from the Generalized Linear Mixed Model, with lower and upper credible intervals (2.5% and 97.5%) in brackets.

RESULTS

Comparative performance of models — broad trends

Results clearly indicate that all three modelling methods are capable of producing predictions that accord well with independent data. The overall mean of AUC across all models was 0.72, but for approximately 47% of these models, the AUC exceeded 0.75. Individual models constructed using random pseudo-absences provided the lowest predictive performance (Table 2). The use of inventory pseudo-absences gave a substantial gain in predictive performance, both in their ability to rank sites well (AUC) and in their ability to predict patterns of occurrence in the evaluation data as indicated by deviance. While there was a consistent trend for multiresponse models to improve prediction further, the changes relative to those for change in pseudo-absences were smaller, and the credible intervals showed more overlap (Table 2).

Comparative performance of models between region-groups

Closer examination of the comparative performance of models indicated that the effect of changing model type and data construct varied between regions and biological groups, as evidenced by the trends in predictive performance (Fig. 1) and the magnitude of the interaction effect in the GLMM (Table 3). We present the results for the distinct biological groups –10 in all — because in the two regions with more than one group, the groups behaved differently. The change in data construct from random pseudo-absences to inventory pseudo-absences resulted in an increase in mean AUC for eight groups (Fig. 1a), and a decrease in two (NZ and AWT plants). The pattern was similar (but opposite direction, as expected, Fig. 1b) for predictive deviance. While there was a consistent continuation of the trends in predictive deviance when moving from single to multiresponse models for all groups, the changes in AUC were slightly more variable (Fig. 1). Five regions were consistent with the overall trend — i.e. mean AUC increased with use of a multiresponse model, though in many cases the increase was relatively small. The exception — NSW — showed varying behaviour among groups. For one NSW group, plants, the mean AUC increased from 0.691 to 0.732 with use of multiresponse models, whereas for the other three groups (mammals, birds, and reptiles) AUC decreased slightly (Fig. 1a).

Table 3 Results from a Generalized Linear Mixed Model analysis of the importance of factors affecting predictive performance of models fitted to presence-only data as measured by AUC.

Model	DIC	ΔDIC
Full model:		
AUC ~ method + method*group + species	−1395	
Without method	−1298	97
Without interaction (method*group)	−1341	54
Without species	−750	645

Changes of greater than 10 in the Deviance Information Criterion (DIC) are indicated as important.

Examination of the geographical distribution of random pseudo-absences for the different regions indicated the potential for *random pseudo-absences* and *inventory pseudo-absences* to

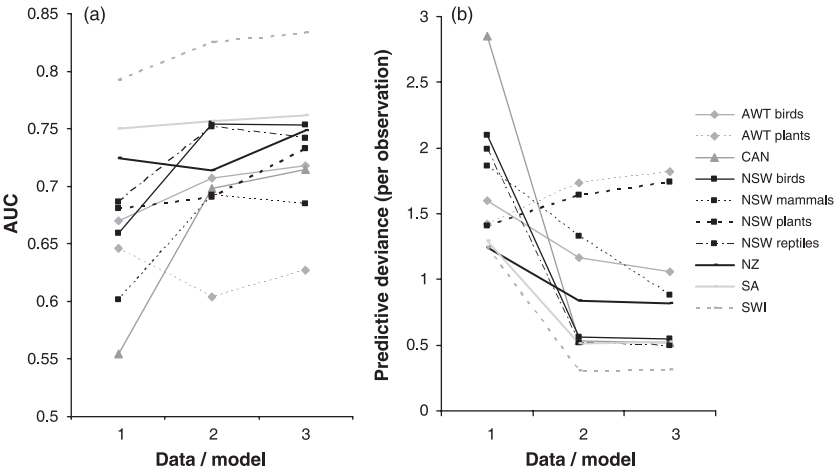

Figure 1 Changes in predictive performance as measured by AUC (a) and predictive deviance for 10 region-groups (b), for three data/model constructs. (1 = single-species model, random pseudo-absences, 2 = single-species model, inventory pseudo-absences, 3 = multiresponse model, inventory pseudo-absences). Data set codes as in main text.

Figure 2 Distribution of community sites (grey) and evaluation sites (black) for NSW plants (left) and New Zealand.

Table 4 Numbers of variables and basis functions fitted in MARS (multivariate adaptive regression splines) models [means and standard errors (SE)], compared with their predictive performance.

	Number of variables ± SE			No. of coefficients ± SE			Frequencies for AUC results	
	Model 1*	Model 2*	Model 3*	Model 1*	Model 2*	Model 3*	Mod 2 > Mod 1	Mod 3 > Mod 2
AWT birds	5.8 ± 0.9	4.7 ± 1.2	4 ± 0	12.0 ± 2.1	8.3 ± 2.1	9 ± 0	0.70	0.50
AWT plants	1.9 ± 0.9	2.2 ± 0.9	1 ± 0	3.7 ± 1.5	3.7 ± 1.2	2 ± 0	0.30	0.65
CAN	3.9 ± 1.0	3.6 ± 1.0	3 ± 0	10.9 ± 3.3	8.0 ± 2.8	9 ± 0	0.70	0.60
NSW birds	6.5 ± 1.4	5.5 ± 1.4	5 ± 0	11.5 ± 3.5	9.7 ± 2.6	10 ± 0	0.86	0.43
NSW mammals	4.3 ± 2.1	4.9 ± 1.6	2 ± 0	8.2 ± 3.5	7.7 ± 2.3	4 ± 0	1.00	0.30
NSW plants	2.4 ± 1.2	3.3 ± 1.4	3 ± 0	4.4 ± 2.5	5.7 ± 2.5	7 ± 0	0.55	0.72
NSW reptiles	6.4 ± 1.8	4.9 ± 0.8	3 ± 0	11 ± 3.5	7.8 ± 2.3	5 ± 0	0.63	0.25
NZ	4.0 ± 1.8	4.7 ± 1.8	6 ± 0	7.1 ± 3.5	8.0 ± 3.1	9 ± 0	0.44	0.73
SA	3.4 ± 1.3	4.6 ± 1.6	6 ± 0	7.2 ± 2.9	8.9 ± 2.8	9 ± 0	0.50	0.53
SWI	8.7 ± 1.5	8.0 ± 1.8	11 ± 0	17.9 ± 2.7	15.1 ± 3.5	23 ± 0	0.83	0.83

*Model 1 = single/random; model 2 = single/inventory; model 3 = multiresponse/inventory.

sample different parts of the overall geographical/environmental space. For example, the plant sites for NSW were relatively well distributed over the whole region (left panel, Fig. 2) and mimicked a random sample, whereas those for NZ left large tracts of dry, lowland land, much of which is now cleared, very sparsely sampled, and the difference between these and a random sample was more marked. Comparable figures for all regions and groups are available online (see Figure S1 in Supplementary Material).

In analysing variation in the predictive gains with progression from *single-species models* to *multiresponse models* (both on inventory-based absences), we found varying degrees of relationship between performance improvement and properties of the modelling or evaluation data. The changes in AUC were most interesting in their variation between region-groups. There was no consistent trend in predictive gain with number of modelling records or prevalence of the species as measured in the PA data (Fig. 3). However, there was a strong positive correlation (Pearson $r = 0.76$) between the number of species in a group and the frequency with which the AUC for the multiresponse model was higher than the AUC for the comparable single-species model (Fig. 4) — i.e. the advantages of using multiresponse models increased with increasing numbers of species in a biological group. This was partially reflected in the complexity of the selected models — biggest gains from use of multiresponse models generally occurred where a larger number of variables and basis functions were selected (compare the mean number of variables with the last column in Table 4). In contrast, there was no clear trend in model complexity with change in data construct (i.e.

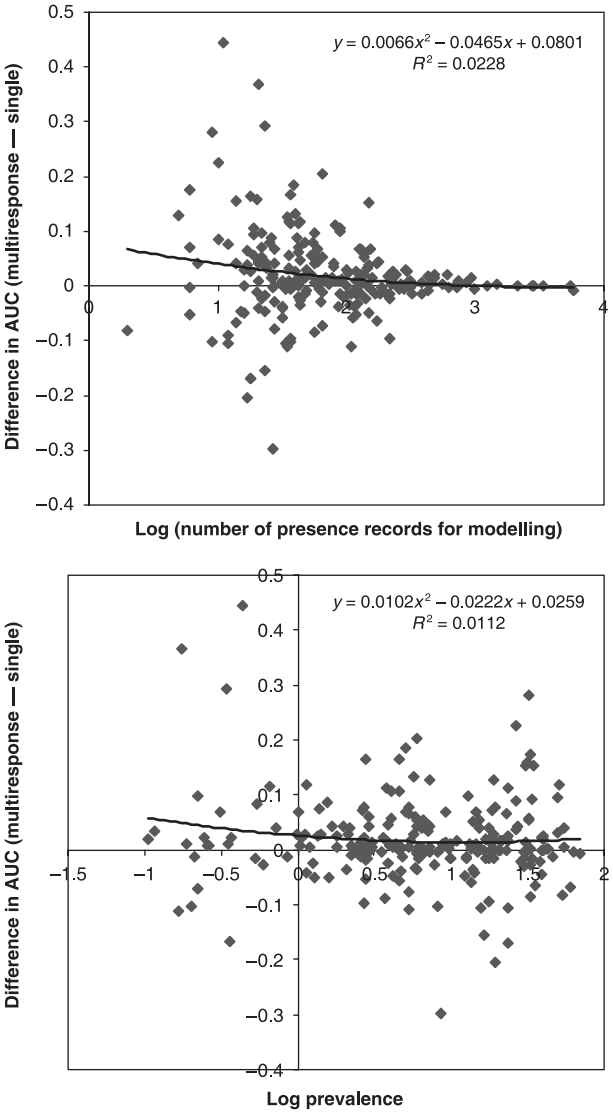

Figure 3 Relationship between multiresponse advantage (as measured by difference in AUC) and characteristics of the modelling and evaluation data. Prevalence was estimated from frequencies in the evaluation data.

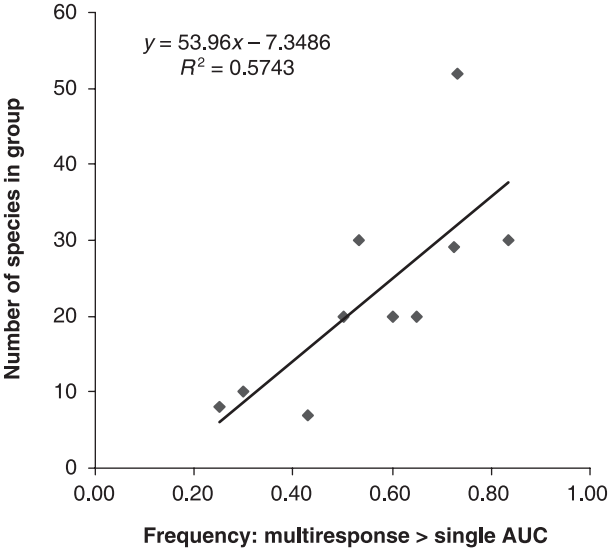

Figure 4 Relationship between number of species in a group and multiresponse advantage as measured by the frequency with which the multiresponse AUC was higher than the single-species AUC within that group. All models fitted on inventory pseudo-absences.

from single-species models based on random pseudo-absences to single-species models based on inventory pseudo-absences) (Table 4). We investigate the variables selected and provide detailed examples for two species in the Supplementary Material (see Appendix S1).

DISCUSSION

Predictive performance with presence-only data

Our results indicate that models fitted with presence-only data can be sufficiently accurate to be useful. The best models used inventory pseudo-absences and multiresponse models, and gave a mean predictive AUC of 0.74, sufficient for a useful contribution to conservation planning applications (Pearce & Ferrier, 2000). This is the most reliable quantitative test we can offer of model ability to produce predictive maps that indicate the current distribution of the species. The marked variation in results across regions and groups, and among species, provides a reminder that the fitting of reliable models is not always possible, and it is important to evaluate the output as rigorously and comprehensively as possible. Differences between regions and species are discussed elsewhere (Elith *et al.*, 2006).

Comparison of data constructs

Inventory pseudo-absences strongly outperformed the random pseudo-absences that are more traditionally used in analyses of presence-only data. Several explanations are possible. In conceptual terms, any set of constructed absence records provides a contrast against which the presences are modelled. If the presence records are restricted to only a part of the environmental or geographical space in a region, the most sensible option may be to also restrict the universe from which absences are selected to the same space. This prevents the selection of absences from combinations of environment where the species has never been searched for.

This restriction of the searched space may represent one of two things. First, presence records might exist within a subset of the space because other parts of that space have been severely modified (e.g. by clearing) or do not support relevant habitat. For example, the NZ plant presence records (Fig. 2) were drawn from a constrained geographical and environmental subset, because the vast majority of dry lowland landscapes have been largely cleared and now support agriculture.

Second, collection of records might be biased to more accessible areas, meaning that important parts of potentially good habitat

remain unsampled — e.g. CAN data were severely biased to southern parts of the sampling area, and species' ranges were only partially represented (Figure S1 in Supplementary Material). In this case inventory pseudo-absences were still appropriate because they remained within the sampled portion of the region, but, as a consequence of this bias, predictions outside the sampled space may be unreliable. In our study most of the evaluation data sets had reasonably complete environmental and geographical coverage (Elith *et al.*, 2006; see Figure S1), enabling prediction to the full region to be adequately tested. The one exception was the CAN evaluation data, which had similar biases to the modelling data, and did not give a comprehensive test of predictive performance outside the areas sampled by the presence-only data. The move from random to inventory pseudo-absences had the most marked effect in CAN, but this was a test largely restricted to that part of the region sampled by the presence-only data. Clearly the risk in using any type of pseudo-absence is that the models will not extrapolate to all areas in the region, but this is true wherever substantial parts of the environmental space are unsampled (see Araújo *et al.*, 2005; Randin *et al.*, 2006; for discussion of extrapolating predictions). This suggests that where samples are severely biased, areas falling well outside the sampled space should be masked out to prevent the use of unreliable predictions. Alternatively, some indication of the reliability of predictions should be provided either by analysis of variation in sample intensity or by mapping environmental distance between a site of interest and the closest available sampling (e.g. Leathwick, 2001).

The interplay between the modelling and evaluation data also affects other, possibly anomalous, results in this study. For example, the AWT plants showed opposite trends to most groups, with random pseudo-absences giving better predictive performance than inventory pseudo-absences. The AWT presence-only data (and therefore the inventory pseudo-absences) were biased to the northern part of the region, whereas the evaluation data were a small sample, well distributed across the entire geographical region (see Figure S1 in Supplementary Material). None of the models would be well informed about good habitat in the south because there were few presence-only sites there (i.e. the region appears not to be covered comprehensively with herbarium collections), so all are likely to lack some realism — the extent to which this was a problem depended on the geographical distribution of the important environmental drivers of distribution.

The NSW data provided an opportunity to compare biological groups within one region. Inventory pseudo-absence models were superior for the three fauna groups, but held no advantage for plants. Collection and survey effort for plants tended to be distributed evenly across the landscape, while fauna surveys were biased towards larger patches of forest. The biased distribution of fauna sites was not reflected in a random sample of absences, and a more marked improvement in model performance occurred when the bias was accounted for by the use of inventory pseudo-absences. In NSW the evaluation data tended to follow the same forest-biased distribution as the modelling data and so the prediction to less frequently sampled areas was not extensively tested.

Comparison of model constructs

A change in model construct, from single-species to community models, gave variable results, with overlap of the credible intervals for both AUC and deviance (Fig. 1 and Table 2). The multiresponse models were useful in some regions — for example, in NZ they gave a mean increase in AUC from 0.71 to 0.75 and a small decrease in mean predictive deviance (0.84–0.82). Six of the 11 candidate variables were selected in the multiresponse model (Table 4 and Table S1 in Supplementary Material), whereas fewer tended to be fitted in the single-species models (mean 4.7 variables). Presumably, there was enough consensus in the broad trends across the species for a single set of variables to suit most species. Modelling all species simultaneously in the multiresponse model may improve the stability of variable selection, which is especially important where there are limited data to parameterize the model (NZ had an average of 59 records per species in the presence-only data, with a range of 18–211).

The advantage of multiresponse models was strongly correlated with the number of species in the biological group, i.e. multiresponse models are likely to be particularly advantageous for large collections of data with many species in a group. A further advantage is the reduced time taken when fitting and making predictions from only one multiresponse model.

Further considerations

Our results can be linked to those from the growing literature on the measurement of biases in presence-only data (Freitag *et al.*, 1998; Griffiths *et al.*, 1999; Reddy & Dávalos, 2003) and how to design surveys to provide better coverage when existing data are inadequate (Dennis & Hardy, 1999; Hortal & Lobo, 2005; Ferrier *et al.*, in press). Geographical biases are most commonly measured, and these may be directly relevant to species models where distributions are significantly affected by factors operating primarily in geographical space, e.g. historical disturbances, or physical barriers to dispersal (Leathwick, 1998), or where they are at a scale that causes spatial autocorrelation in the records (Legendre, 1993). However, most models are constructed in environmental space, so biases in the *environmental* distribution of records will be most critical for distribution models. These might, of course, coincide with geographical biases (Kadmon *et al.*, 2004). The impact of biases on environmental space will vary according to what response is being modelled. For regression models of species' occurrence (i.e. response = presence or absence), survey effort at a site will produce a reliable presence record unless there are locational and taxonomic uncertainties. Selection of absences is most problematic, as demonstrated by our results here. Analyses of survey biases in environmental space could help to identify the best sampled areas, and be used to guide placement or weighting of pseudo-absences. Where species richness is the modelled response, biases are likely to be more critical, and analyses of survey effort (e.g. Soberón & Llorente, 1993; Reddy & Dávalos, 2003) should inform inclusion or weighting of records and to define unsampled areas.

While we have demonstrated useful predictive performance from presence-only species data, it is important to also recognize their limitations. Because presence-only data give no reliable information on the prevalence (frequency of occurrence) of species in the region, and we have used them in a standard regression model, predictions are not probabilities of occurrence but indicate relative likelihoods of species presence. This means that the predictions can inform ranking of sites with respect to an individual species, but cannot be used to make statements about relative differences in occurrence between species. It is likely that the results for change in data construct are related to this issue. The models based on random absence samples have no information at all on prevalence, and in this application we simply weighted the random samples so their total weight equalled that of the presence records. This will give poor calibration — i.e. the predicted probabilities will not accurately match the true probabilities, and this will be reflected in the predictive deviance but not in the rank-based AUC. Calibration of the models will change with the change in data construct unless weights are used in the same way. Using inventory pseudo-absences is likely to provide some information on prevalence, because few records of the species among all sites may give some indication of its true prevalence in the landscape, though this could be confounded by sampling biases towards rare and unusual species, and away from those that are difficult to detect or collect. Nevertheless, there appears to be enough useful information in these data to improve the calibration of the models, resulting in a generally lower predictive deviance.

A more statistically rigorous approach to modelling with pseudo-absences would be to make adjustments to the regression method, rather than using techniques designed for true presence-absence data. Case-control regression that takes account of contaminated zeros is one example of this (Keating & Cherry, 2004; Pearce & Boyce, 2006), and relevant software for dealing with typical ecological applications is in development (Simon Barry, Gill Ward, pers.comm.).

Final comments

Testing of MARS models over 226 species has demonstrated that data from multiple species are useful both for the signal in the species data and for inventory information. The advantage of multiresponse models was variable across regions, and was strongly correlated with the number of species in the biological group. The particular trade-offs of multiresponse methods under a range of circumstances require further research. Our results confirm the sensitivity of model outcomes to the methods used for selection of absences. Random pseudo-absences sample the background environmental space in proportion to its frequency in geographical space. Inventory pseudo-absences focus on the sampled space, and avoid placement of absences in unsampled areas. Other strategies have been suggested (Pearce & Boyce, 2006), including making models of sampling intensity and using these as inclusion probabilities for absence samples (Zaniewski *et al.*, 2002), and sampling in areas known or estimated to be unsuitable for the species (Engler *et al.*, 2004; Lobo *et al.*, 2006). These are motivated by different philosophies about samples and models, and different purposes for the modelled output. A more comprehensive theoretical and practical comparison of strategies for selecting pseudo-absence and of their implications for a range of modelling methods is clearly warranted.

ACKNOWLEDGEMENTS

This research was initiated in a working group at the National Center for Ecological Analysis and Synthesis (NCEAS), Santa Barbara, USA: 'Testing Alternative Methodologies for Modelling Species' Ecological Niches and Predicting Geographic Distributions', conceived of and led by Town Peterson and Craig Moritz. The participants there contributed to the broader study design, and in particular Simon Ferrier suggested that we think about how to select pseudo-absences. Data were kindly supplied by many individuals and institutions including Andrew Ford and Caroline Bruce (CSIRO, Atherton), Karen Richardson (University of Queensland) and Stephen Williams (James Cook University) (AWT data); Falk Huettman (University of Alaska, Fairbanks), Mark and George Peck (Royal Ontario Museum), Mike Cadman (Canadian Wildlife Service of Environment Canada) (CAN data); Simon Ferrier (Department of Environment and Conservation, NSW) (NSW data); Jake Overton (Landcare Research, New Zealand) (NZ data); Bette Loiselle, Lucia Lohmann, Robert Magill and Trisha Consiglio (University of Missouri, St Louis) (SA data); Antoine Guisan (University of Lausanne), Nicolas Zimmerman, T. Wohlgemuth, and U. Braendi (WSL Switzerland) (SWI data). Catherine Graham (Stonybrook University, NY) helped to prepare the data. Trevor Hastie (Stanford University) stimulated our interest in MARS models. Simon Ferrier helped interpret species responses in NSW and provided invaluable insight into the use of deviance as a measure of predictive performance. Comments from Brendan Wintle, Joaquín Hortal and an anonymous reviewer substantially improved the manuscript. Jane Elith was funded by ARC grants DP0209303 and LP0348897, and the Australian Centre of Excellence for Risk Analysis. Finally, we thank the Universidad Internacional de Andalucía, sede Antonio Machado, Baeza, Spain, for organizing the workshop 'Predictive modelling of species distribution — New tools for the XXI century', from which this special issue emerged.

REFERENCES

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

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

Araújo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. (2005) Validation of species-climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Brotons, L., Thuiller, W., Araujo, M.B. & Hirzel, A.H. (2004) Presence–absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437–448.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and inference: a practical information-theoretic approach*, 2nd edn. Springer-Verlag, New York.

Busby, J.R. (1991) BIOCLIM — a bioclimate analysis and prediction system. *Nature conservation: cost effective biological surveys and data analysis* (ed. by C.R. Margules and M.P. Austin), pp. 64–68. CSIRO, Canberra, Australia.

Carpenter, G., Gillison, A.N. & Winter, J. (1993) DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, **2**, 667–680.

Cawsey, E.M., Austin, M.P. & Baker, B.L. (2002) Regional vegetation mapping in Australia: a case study in the practical use of statistical modelling. *Biodiversity and Conservation*, **11**, 2239–2274.

Dennis, R.L.H. & Hardy, P.B. (1999) Targeting squares for survey: predicting species richness and incidence of species for a butterfly atlas. *Global Ecology and Biogeography*, **8**, 443–454.

Edwards, T.C., Cutler, D.R., Zimmerman, N.E., Geiser, L. & Alegria, J. (2005) Model-based stratifications for enhancing the detection of rare ecological events. *Ecology*, **86**, 1081–1090.

Elith, J., Graham, C.H., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K.S., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S. & Zimmermann, N.E. (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Engler, R., Guisan, A. & Rechsteiner, L. (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, **41**, 263–274.

Ferrier, S. (2002) Mapping spatial pattern in biodiversity for regional conservation planning: where to from here? *Systematic Biology*, **51**, 331–363.

Ferrier, S. & Guisan, A. (2006) Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology*, **43**, 393–404.

Ferrier, S., Manion, G., Elith, J. & Richardson, K. (2007) Using generalised dissimilarity modelling to analyse and predict patterns of beta-diversity in regional biodiversity assessment. *Diversity and Distributions*.

Ferrier, S., Powell, G.V.N., Richardson, K.S., Manion, G *et al.* (2004) Mapping more of terrestrial biodiversity for global conservation assessment. *Bioscience*, **54**, 1101–1109.

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

Freitag, S., Hobson, C., Biggs, H.C. & Van Jaarsveld, A.S. (1998) Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. *Animal Conservation*, **1**, 119–127.

Frey, J. (2006) Inferring species distributions in the absence of occurrence records: An example considering wolverine (*Gulo gulo*) and Canada lynx (*Lynx canadensis*) in New Mexico. *Biological Conservation*, **130**, 16–24.

Friedman, J.H. (1991) Mutlivariate adaptive regression splines (with discussion). *Annals of Statistics*, **19**, 1–141.

Funk, V. & Richardson, K. (2002) Systematic data in biodiversity studies: use it or lose it. *Systematic Biology*, **51**, 303–316.

Goolsby, J.A. (2004) Potential distribution of the invasive old world climbing, fern, *Lygodium microphyllum* in north and south America. *Natural Areas Journal*, **24**, 351–353.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004b) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, **19**, 497–503.

Graham, C.H., Ron, S.R., Santos, J.C., Schneider, C.J. & Moritz, C. (2004a) Integrating phylogenetics and environmental niche models to explore speciation mechanisms in dendrobatid frogs. *Evolution*, **58**, 1781–1793.

Griffiths, G.H., Eversham, B.C. & Roy, D.B. (1999) Integrating species and habitat data for nature conservation in Great Britain: data sources and methods. *Global Ecology and Biogeography*, **8**, 329–345.

Guisan, A. & Zimmerman, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Guisan, A., Graham, C., Elith, J., Huettmann, F. & NCEAS modelling Group (2007) Sensitivity of predictive species distribution models to change in grain size: insights from an international experiment across five continents. *Diversity and Distributions*.

Hanley, J.A. & McNeil, B.J. (1982) The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hastie, T. (1991) Generalized additive models. *Statistical models in S* (ed. by J.M. Chambers and T.J. Hastie), pp. 249–308. Wadsworth and Brooks/Cole Advanced Books and Software, California.

Hastie, T., Tibshirani, R. & Buja, A. (1994) Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 1255–1270.

Hastie, T., Tibshirani, R. & Friedman, J.H. (2001) *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag, New York.

Hortal, J., Garcia-Pereira, P. & Garcia-Barros, E. (2004) Butterfly species richness in mainland Portugal: predictive models of geographic distribution patterns. *Ecography*, **27**, 68–82.

Hortal, J. & Lobo, J.M. (2005) An ED-based protocol for optimal sampling of biodiversity. *Biodiversity and Conservation*, **14**, 2913–2947.

Kadmon, R., Farber, O. & Danin, A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

Keating, K.A. & Cherry, S. (2004) Use and interpretation of logistic regression in habitat selection studies. *Journal of Wildlife Management*, **68**, 774–789.

Leathwick, J.R. (1998) Are New Zealand's *Nothofagus* species in equilibrium with their environment? *Journal of Vegetation Science*, **9**, 719–732.

Leathwick, J.R. (2001) New Zealand's potential forest pattern as predicted from current species-environment relationships. *New Zealand Journal of Botany*, **39**, 447–464.

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

Leathwick, J.R., Rowe, D., Richardson, J., Elith, J. & Hastie, T. (2005) Using multivariate adaptive regression splines to predict the distributions of New Zealand's freshwater diadromous fish. *Freshwater Biology*, **50**, 2034–2052.

Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.

Levin, S.A. (1992) The problem of pattern and scale in ecology. *Ecology*, **73**, 1943–1967.

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

Lobo, J.M., Verdú, J.R. & Numa, C. (2006) Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Diversity and Distributions*, **12**, 179–188.

Lütolf, M., Kienast, F. & Guisan, A. (2006) The ghost of past species occurrence: improving species distribution models for presence-only data. *Journal of Applied Ecology*, **43**, 802–815.

Margules, C.R. & Pressey, R.L. (2000) Systematic conservation planning. *Nature (London)*, **405**.

McCarthy, M.A. & Masters, P. (2005) Profiting from prior information in Bayesian analyses of ecological data. *Journal of Applied Ecology*, **42**, 1012–1019.

McPherson, J.M., Jetz, W. & Rogers, D.J. (2006) Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions-possibilities and 2 limitations. *Ecological Modelling*, **192**, 499–522. 3.

Moisen, G.G. & Frescino, T.S. (2002) Comparing five modeling techniques for predicting forest characteristics. *Ecological Modelling*, **157**, 209–225.

Olden, J.D. (2003) A species-specific approach to modelling biological communities and its potential for conservation. *Conservation Biology*, **17**, 854–863.

Pearce, J.L. & Boyce, M.S. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43**, 405–412.

Pearce, J. & Ferrier, S. (2000) Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological Modelling*, **133**, 225–245.

Phillips, S.J., Dudik, M. & Schapire, R.E. (2004) A maximum entropy approach to species distribution modeling. (ed. by C.E. Brodley), *Proceedings of the 21st International Conference on Machine Learning, Banff, Alberta, Canada, 2004*. pp. 472–486, URL http://www.aicml.cs.ualberta.ca/_banff04/icml/pages/accepted.htm

R Development Core Team (2004) *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Randin, C.F., Dirnböck, T., Dullinger, S., Zimmerman, N.E., Zappa, M. & Guisan, A. (2006) Are niche-based species distribution models transferable in space? *Journal of Biogeography*, **33**, 1689–1703.

Reddy, S. & Dávalos, L.M. (2003) Geographical sampling bias and its implications for conservation priorities in Asia. *Journal of Biogeography*, **30**, 1719–1727.

Robertson, M.P., Peter, C.I., Villet, M.H. & Ripley, B.S. (2003) Comparing models for predicting species' potential distributions: a case study using correlative and mechanistic predictive modelling techniques. *Ecological Modelling*, **164**, 153–167.

Soberón, J.M. & Llorente, J.B. (1993) The use of species accumulation functions for the prediction of species richness. *Conservation Biology*, **7**, 480–488.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2003a) *Winbugs user manual*, Version 1.4. MRC Biostatistics Unit, Cambridge, UK.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & van der Linde, A. (2003b) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **64**, 583–639.

Whittaker, R.H. (1967) Gradient analysis of vegetation. *Biological Review*, **49**, 207–264.

Wieczorek, J.R., Guo, Q. & Hijmans, R.J. (2004) The point-radius method for georeferencing point localities and calculating associated uncertainty. *International Journal of Geographic Information Science*, **18**, 745–767.

Zaniewski, A.E., Lehmann, A. & Overton, J.M. (2002) Predicting species distribution using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, **157**, 261–280.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article:

Figure S1 Geographical distributions of sites for biological groups, from presence-only data (red) and evaluation sites (blue).

Figure S2 Fitted functions for the NSW tree, *Corymbia intermedia*.

Figure S3 Fitted functions for the NZ tree, *Prumnopitys ferruginea*.

Appendix S1 Variables selected and two example species for MARS models.

Appendix S2 Tutorial.

Table S1 Sharing of predictor values in MARS models.

Table S2 Changes in deviance explained (in modelling data) after dropping a variable.

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/10.1111/j.1472-4642.2007.00340.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.