Article:

Platts, Philip J., McClean, Colin J., Lovett, Jon C. et al. (2008) Predicting tree distributions in an East African biodiversity hotspot: model selection, data bias and envelope uncertainty. Ecological Modelling. pp. 121-134. ISSN: 0304-3800

https://doi.org/10.1016/j.ecolmodel.2008.06.028

Chapter 3 – Predicting Tree Distributions

Published in *Ecological Modelling* (Oct. 2008), Vol. 218, Issues 1-2, Pages 121-134

Predicting tree distributions in an East African biodiversity hotspot: model selection, data bias and envelope uncertainty

Philip J. Platts [a,*], Colin J. McClean [b], Jon C. Lovett [b], Rob Marchant [a]

[a] *The York Institute for Tropical Ecosystem Dynamics (KITE), Environment Department, University of York, Heslington, York YO10 5DD, UK.* [b] *Centre for Ecology Law and Policy, Environment Department, University of York, Heslington, York YO10 5DD, UK.*

** Corresponding author. Email: philip.platts@gmail.com*

Abstract

The Eastern Arc Mountains (EAMs) of Tanzania and Kenya support some of the most ancient tropical forest on Earth. The forests are a global priority for biodiversity conservation and provide vital resources to the Tanzanian population. Here, we make a first attempt to predict the spatial distribution of 40 EAM tree taxa (38 species), using generalised additive models, plot data and environmental predictor maps at 1 km resolution. The results of three modelling experiments are presented, investigating predictions obtained by (1) two different procedures for the stepwise selection of predictors, (2) down-weighting absence data, and (3) incorporating autocovariate terms to describe fine-scale spatial aggregation. In response to recent concerns regarding the extrapolation of model predictions beyond the restricted environmental range of training data, we also demonstrate a novel graphical tool for quantifying envelope uncertainty in restricted range niche-based models (envelope uncertainty maps). We find that even for taxa with very few documented occurrences useful estimates of distribution can be achieved. Initiating selection with a null model is found to be useful for explanatory purposes, whilst beginning with a full predictor set can over-fit the data. We show that a simple multimodel average of these two best-model predictions yields a superior compromise between generality and precision. Down-weighting absences shifts the balance of errors in favour of higher sensitivity, reducing the number of serious mistakes (i.e., falsely predicted absences); however, response functions are more complex, exacerbating uncertainty in larger models. Spatial autocovariates help describe fine-scale patterns of occurrence and significantly improve explained deviance, though if important environmental constraints are omitted then model stability and explanatory power can be compromised. We conclude that the best modelling practice is contingent both on the intentions of the analyst (explanation or prediction) and on the quality of distribution data. Generalised additive models have potential to provide valuable information for conservation in the EAMs, but methods must be carefully considered, particularly if occurrence data are scarce.

Keywords: Eastern Arc Mountains; tropical trees; generalised additive models; stepwise selection; model averaging; prevalence; spatial autocorrelation; extrapolation uncertainty.

Introduction

Research into the habitat requirements of species plays a fundamental role in planning for their future conservation, particularly if external pressures such as disturbance and climatic change threaten their persistence. Vegetation surveys provide point data for many taxa, but invariably survey sites are too sparse or spatially biased for species distributions to be estimated directly (Küper *et al.*, 2006). One solution is to model the likelihood of occurrence as a function of the local environment, using the available distribution data and environmental variables as predictors of habitat suitability. Species distribution models have been used previously for biodiversity analysis (Austin, 1999; Ferrier *et al.*, 2002b), improved sampling of rare and endangered species (Engler *et al.*, 2004; Guisan *et al.*, 2006), determination of reserve boundaries (Ferrier *et al.*, 2002a; Araújo *et al.*, 2004), historical reconstruction (Richards *et al.*, 2007) and assessment of climate change impacts (Thomas *et al.*, 2004; McClean *et al.*, 2005). All of these applications could prove extremely useful for the Eastern Arc Mountains of Tanzania and Kenya (EAMs; Lovett, 1985), one of the most important regions for conservation in the world (Olson and Dinerstein, 1998; Stattersfield *et al.*, 1998; Myers *et al.*, 2000), yet to our knowledge no regional-scale predictive model for tree distributions in this area has been published.

The EAMs are a particularly challenging environment to model, characterised by steep climatic gradients that must be portrayed at a high spatial resolution if the environmental tolerances of taxa are to be properly described. The study presented here uses generalised additive models (GAMs; Hastie and Tibshirani, 1990) to parameterise the responses of 38 large tree species (40 taxa, including subspecies and varieties) to a number of climatic and topographic gradients. GAMs are a semi-parametric class of regression model, chosen because of their ability to describe highly non-linear responses (Yee and Mitchell, 1991; Austin, 2007). The aim is to assess the potential of this data-driven tool for assisting research and conservation in the EAMs – the application of GAMs to small environmental datasets is increasingly common, but often due consideration is not given to pitfalls such as over-fitting.

As is common for studies of this nature, the distribution data available to us are not well suited to high-resolution raster-based regression analysis. Impediments to model performance may include mislocated or misidentified samples, low sample size and prevalence, and a biased or restricted distribution of occurrence data. In order to obtain robust estimates of species distributions, and for the benefit of other studies faced with similar challenges, we compare baseline model predictions with those that incorporate down-weighted absences (Maggini *et al.*, 2006) and spatial autocovariates (Augustin *et al.*, 1996). Given that predictions can be highly sensitive to the predictor sets used for modelling (e.g., Dormann *et al.*, 2008), we also calibrate and compare three different methods for model selection: two best-model stepwise procedures and one multimodel.

Model selection

The goal of selection is to construct as parsimonious a predictor set as possible whilst retaining sufficient information to predict the given distribution. A widely used procedure is to select predictors in a stepwise manner, beginning with either a null model (forward selection) or a full model (backward selection) and adding or removing predictors according to their impact on a global measure of model performance (Eberhardt, 2003). Marginal statistics can be biased by the inevitable collinearity amongst environmental predictors (Cohen *et al.*, 2003; Graham, 2003), and so the use of null hypothesis tests during selection is best avoided. Issues of multiple testing (Pearce and Ferrier, 2000a; Whittingham *et al.*, 2006) and arbitrary levels of statistical significance (Mickey and Greenland, 1989; Rushton *et al.*, 2004) further enforce this standpoint. Multimodel inference has been proposed as an alternative to best-model stepwise procedures. Anderson *et al.* (2000) for instance describe an approach called information-theoretic (IT), in which a number of good models are identified from an *a priori* set of hypotheses (predictor sets) and then compared using Akaike Information Criterion (AIC; Akaike, 1973), or combined in a model-average using Akaike weights. Although not strictly adhering to the IT philosophy of multimodel inference, many studies now adopt the use of AIC in stepwise procedures.

Data bias

With absences often far outweighing presences, particularly for rare and less well-known species, low sample prevalence is a common problem that can lead to misleading evaluations (Manel *et al.*, 2001; Engler *et al.*, 2004; McPherson *et al.*, 2004). A standardised prevalence can be achieved by applying weights to the absence data prior to parameterisation, as demonstrated by Maggini *et al.* (2006) in their modelling of Switzerland's forest communities. The technique was shown to perform well, improving both the accuracy and stability of predictions. Maggini *et al.* found that the application of weights increased the overall probabilities of occurrence, and also report that the balance of model fit may have been altered. It is the latter in which we see potential for improving our predictions: absence 'observations' are inherently unreliable (Anderson, 2003), and since misclassifications distort the modelled relationship between species and environment it follows that a strategic reduction in the dependence of models on absence data could be beneficial. Simulations based on use-availability data (resource selection function modelling; Johnson *et al.*, 2006) suggest that logistic regression is relatively robust to contamination rates of below 20% – a level that could well be exceeded in our data.

Another source of error is the tendency for nearby locations to be alike in terms of the communities they support, a trend known as spatial autocorrelation (SAC). If a regression model cannot explain fully the observed spatial clustering then its residuals exhibit spatial structure, violating the assumption that they should be independent and identically distributed. There are two reasons why this kind of error is common in niche models. First, predictors rarely contain sufficient information to describe fully the observed species aggregation (Guisan and Thuiller, 2005); missing pieces of the puzzle include dispersal patterns, competition/mutualism and disturbance. Second, ecologists are inclined to visit sites in more accessible locations and areas of particular interest, yielding a spatially clustered sampling distribution that may not be representative of SAC in the candidate predictors. Over recent years, the number of ecological studies to address SAC in models has increased, with a majority reporting significant improvements in model fit (Dormann, 2007b). Augustin *et al.* (1996) modelled deer populations using autologistic regression, a form of auto-model (Besag, 1974) that has since been applied to a variety of species distribution model (Miller *et al.*, 2007). In previous application to GAMs, this method for describing localised spatial dependence has performed well (e.g., Segurado and Araujo, 2004); however recent studies warn that autologistic (autocovariate) models may underestimate the environmental controls on a species distribution (Dormann, 2007a; Dormann *et al.*, 2007).

Envelope uncertainty

The breadth of niche-space spanned by the distribution data may not be sufficient to fully represent the study region, particularly if projecting models under climate change scenarios. This is a common problem in the estimation of species distributions, though there are few tools available for estimating the associated uncertainty in predictions (Pearson and Dawson, 2003; Thuiller *et al.*, 2004; Pearson *et al.*, 2006; Dormann, 2007c). For GAMs specifically, model uncertainty arises because response shapes are constructed using non-parametric smoothers – each smoother focuses on a specific portion of the data, and so the modelled response does not naturally extend past the limits of the training data. In essence, the problem is the same for all predictive models: that attempts to predict species occurrence beyond the documented niche-breadth are subject to high uncertainty, particularly if more than one environmental factor is under-represented (Thuiller *et al.*, 2004). At a time when extrapolations into unknown climate-space are increasingly in demand, the development of methodologies to address this issue has been identified as a priority for research (Araujo and Guisan, 2006). The solution we present is to accompany model predictions with envelope uncertainty maps (EUMs), which allow the analyst to identify geographical locations where the profile of environmental conditions at sample sites results in high model uncertainty.

Methods

Study region

The EAMs are part of the Eastern Afromontane Biodiversity Hotspot (Mittermeier *et al.*, 2004) and are defined as those ancient crystalline mountains under the direct climatic influence of the Indian Ocean (Lovett, 1990). Beginning in the Taita Hills of southern Kenya, they extend down through eastern Tanzania to the Udzungwa Mountains in the south (Fig. 3.1). The mountains are a chain of 13 disjoint blocs, isolated from the surrounding lowlands since the Miocene about 30 million years ago (Schlüter, 1997). Today they support 3300-5700 km^2 of moist tropical forest, though it has been estimated that this may be less than 30% of the original forest cover (Burgess *et al.*, 2007b). Much of the remaining area is protected by forest and nature reserves, national parks and community-based management, many covering critical water catchments; the EAMs are a source of drinking water and hydroelectric power for over half of Tanzania’s urban population. The archipelago-like distribution of mountain blocs promotes significantly higher range-size rarity than is found in some other high biodiversity tropical ecosystems (Taplin and Lovett, 2003; Burgess *et al.*, 2007a), rendering EAM flora particularly sensitive to further fragmentation. Species richness scores are high and the concentrations of endemism are exceptional (Burgess *et al.*, 2007b), though many hundreds of endemic plants and animals are threatened by extinction. Around 500 vascular plant species are putatively endemic, including over 80 tree species (Lovett *et al.*, 2006). Subject to significant anthropogenic pressure and harbouring such high biodiversity per unit area, the EAMs have been identified as one of Earth's most fragile biodiversity hotspots (Brooks *et al.*, 2002).

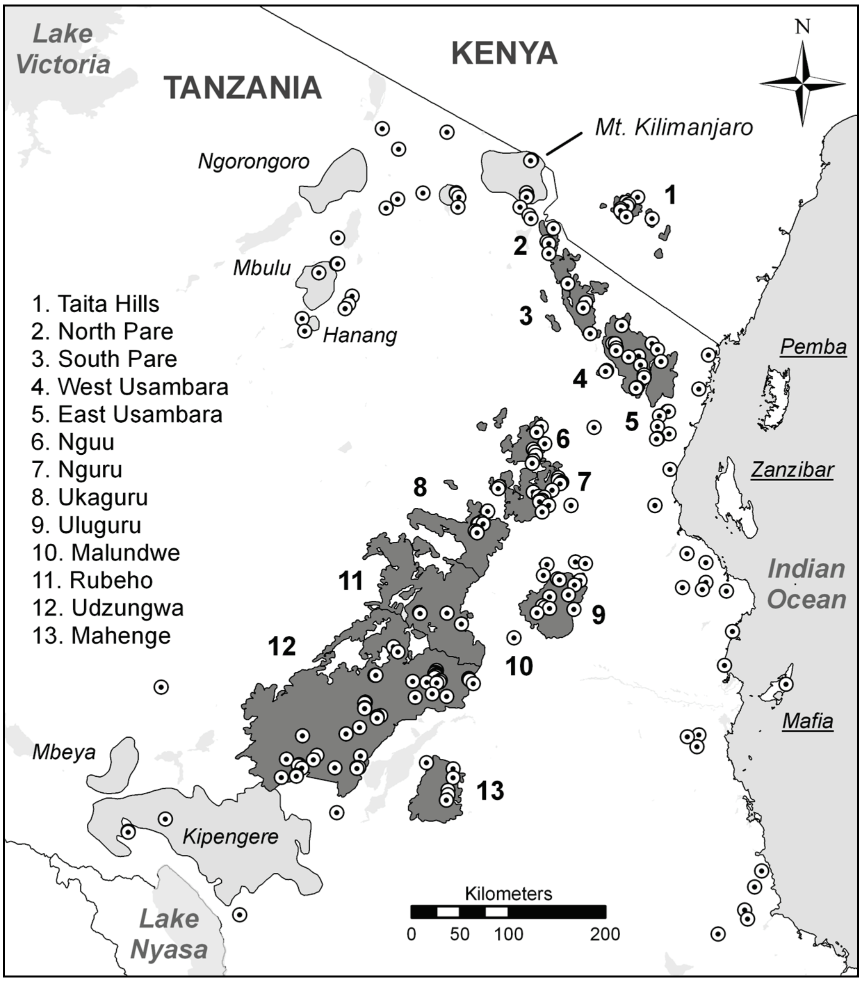


Figure 3.1. Map showing the 13 crystalline blocs that comprise the Eastern Arc Mountain chain. Encircled dots locate the 201 modelling points. Note the clustered distribution of samples – a classic problem in species distribution modelling. Region for model extrapolation was the full map extent: 32.5°E–40.5°E, 1.5°S–10.5°S.

Tree data

The tree database collates observations from 363 variable area plots visited between 1979 and 1994. Since some of our target species' ranges extend beyond the EAMs (e.g., Hemp, 2006), we included plots from other forested mountains such as Mt. Kilimanjaro, and also from the nearby coastal forests (Coastal Forests of Eastern Africa Biodiversity Hotspot; Mittermeier *et al.*, 2004). All plots share a common sampling method, whereby a focal point is chosen and the nearest 20 trees measuring at least 20 cm diameter at breast height are recorded. Lists of trees occurring outside the plots were also included in the database where available. Using field notes of location and altitude, we identified 201 distinct 30 arc-sec (920 m) grid cells across which the samples were collected (Fig. 3.1). The choice of scale is an important consideration for modelling; in the EAMs climatic conditions vary rapidly over short spatial scales, and so we used the finest cell size allowed by the field data. A coarser grid would aggregate more sites, reducing the impact of SAC and mislocation errors in the data, but critically for our study area might omit important changes in habitat across the altitudinal gradient. A species was recorded absent from a grid cell if there was no record of presence in either the plot data or the tree lists. These absences should be considered 'pseudo-absences' because the lists are not exhaustive and the 20-tree method is unlikely to capture all species present at a given site. A full list of the tree species modeled, including their sample prevalence, is presented in Appendix 3A. For further details of the field data we refer the reader to Lovett (1998).

Environmental predictor variables

Climatic and topographic predictor maps were used to estimate the environmental conditions at each site and to extrapolate predictions to the wider Eastern Arc region. Climate surfaces were obtained from the Centre for Resource and Environmental Studies, Australian National University (http://cres.anu.edu.au/). The grids are based on climate station data collected between 1920 and 1980, and provide estimates of mean monthly rainfall and mean daily temperature extremes at a spatial resolution of three arc minutes (5.5 km). To achieve consistency with the 30 arc-sec resolution of the tree data, we interpolated the surfaces using a distance-weighted average of the 16 nearest neighbours. Derived predictors were then calculated to better represent the climatic gradients directly affecting species distributions (Table 3.1). Absolute minimum temperature is as described by Prentice *et al.* (1992), potential evapotranspiration follows the Thornthwaite (1948) method, and annual moisture index is the ratio of mean annual precipitation to potential evapotranspiration. Our temperature-days variable is derived similarly to the growing degree-days measure commonly applied in temperate zones. Its inclusion here provides species-specific information on climatic suitability across all 12 months of the year. Since the phenologies of modelled taxa are not known, we bounded suitable conditions for growth using the presence records: for each species *i*, the upper bound was the maximum value of *tmeanw* across all sites where species *i* occurs; the lower bound was the minimum *tmeanc* (refer to Table 3.1 for abbreviations).

Table 3.1. Summary of environmental predictor variables. Climatic range is high because of Mt. Kilimanjaro (5895 m AMSL).

Predictor	Description	Units	Mean		Range	
			Sites	All cells	Sites	All cells
gradient*	Angle from horizontal	°	7.242	1.533	28.88	64.58
trasp*	Wetness/radiation index	–	0.6720	0.5366	1.000	1.000
trange*	Annual temperature range	°C	15.85	16.07	8.672	13.10
pptdry*	Precipitation driest month	mm	11.95	4.949	54.11	94.80
pptann^1	Mean annual precipitation	cm	107.4	87.14	121.1	194.1
pptwet^1	Precipitation wettest month	mm	229.1	186.4	389.3	437.0
tmean^2	Mean temperature	°C	21.20	22.71	14.01	36.44
tmeanw^2	Mean temp. warmest month	°C	23.23	24.69	14.03	36.45
tmeanc^2	Mean temp. coldest month	°C	18.44	20.23	14.75	36.23
tmaxw^2	Max. temp. warmest month	°C	28.83	30.46	13.58	38.50
tminc^2	Min. temp. coldest month	°C	13.15	14.52	17.56	33.75
tabsmin^2	Absolute minimum temp.	°C	4.486	7.228	22.45	51.94
tdays$^{(2)}$	Temperature-days	°-days	–	–	–	–
pet^2	Potential evapotranspiration	cm/year	105.4	115.1	99.77	223.9
ami^2	Annual moisture index	cm/year	1.093	0.7918	2.041	4.997

* = independent [abs(r) < 0.7]; 1 = first correlation group; 2 = second correlation group; () = species-specific

Topographic data were from the Shuttle Radar Topography Mission, U.S. Geological Survey (http://srtm.usgs.gov/). Two predictors were derived from the partial derivatives of elevation (Horn, 1981): gradient of the slope and a transformation of aspect (Table 3.1). The digital elevation model was supplied at a resolution of three arc-sec (92 m); derived predictors were rescaled to 30 arc-sec for compatibility with the tree data. In order to overcome the problems associated with using a circular predictor variable (i.e.,, 0° ≡ 360° on a compass) we used a cosine transformation of aspect to obtain a symmetric wetness/radiation index (Roberts and

Cooper, 1989). Plots of aspect against rainfall showed that on average slopes facing east-south-east receive the most rainfall during the dry season, when moisture carried by the trade winds is most critical, and so these slopes were allocated the highest wetness indices, and west-north-west facing slopes the lowest.

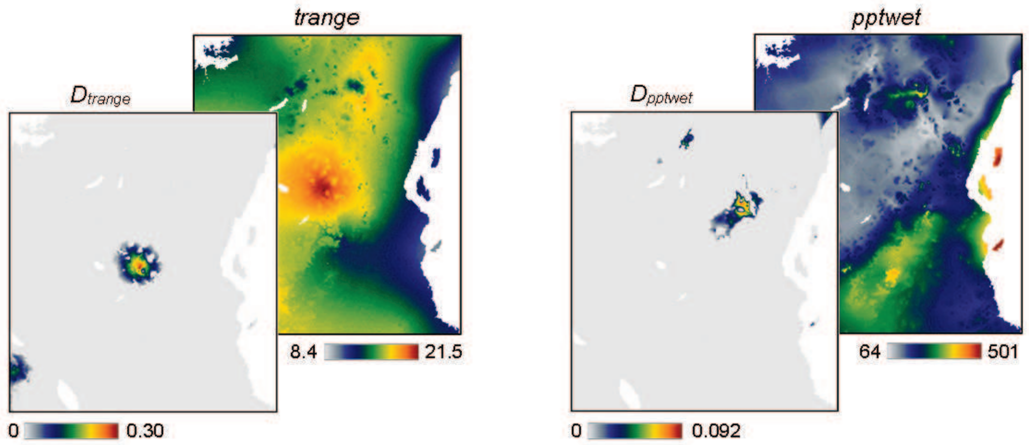


Figure 3.2. Environmental coverage of predictor variables. Background: annual temperature range (left) and precipitation during the wettest month (right). Foreground: proportional distance (D_X) of these predictors from the calibration envelope. Distance maps can be combined in a contribution-weighted average to yield envelope uncertainty maps (EUMs).

In order to evaluate how well the 201 modelling points captured the environmental range of our study region, the proportional 'distance' of each grid cell from the calibration envelope was calculated with respect to each predictor (Fig. 3.2). Envelope uncertainty maps (EUMs) estimate the associated model uncertainty using an average of these distance maps, weighted according to the relative contributions of predictors in a model. Cell i is given by

$$\mathrm{EUM}_i = \frac{\sum\limits_{X \in S} C_X D_{X_i}}{\sum\limits_{X \in S} C_X} ,$$

where S is the predictor set, C_X is the contribution of predictor X, and D_{Xi} is the proportional distance of X_i from the calibration envelope:

$$D_{X_i} = \frac{\max\left\{\hat{X}_{\min} - X_i, X_i - \hat{X}_{\max}, 0\right\}}{\hat{X}_{\max} - \hat{X}_{\min}},$$

where $\hat{X}$ denotes the calibration subset. In this paper we define predictor contributions to be the percentage drop in explained deviance when predictor X is removed from the final model. As a rule of thumb, Dormann (2007c) recommends that one should not extrapolate further than 1/10th of the parameter range (i.e., D_X should not exceed 0.1). Particular caution is therefore recommended for grid cells where the EUM > 0.1, since this indicates that at least one predictor has been extrapolated beyond the 1/10th level.

Statistics for calibration and evaluation

Model performance was assessed using the proportion of explained deviance (D^2), area under the receiver-operating characteristic curve (AUC; Green and Swets, 1974) and an associated measure of generalisation error (GE; see below). Predictions of occurrence were on a continuous scale, from zero to one. For direct comparison with the tree data, these were dichotomised by maximising the sum of sensitivity (proportion of presences correctly predicted) and specificity (proportion of absences correctly predicted), a method shown to perform well by Liu *et al*. (2005) in their comparative study. The AUC is a threshold-independent measure, incorporating both type I (false positive) and type II (false negative) error rates, and is largely unaffected by sample prevalence (McPherson *et al.*, 2004). In a recent critique of the AUC (Lobo *et al.*, 2008), the lack of spatial information and validity of symmetric error-weights are questioned – two weaknesses that could be mitigated by the use of EUMs and absence-weights, respectively. For testing the significance of differences between models we used the Wilcoxon rank sum (Mann–Whitney) test, a statistic closely related to the AUC (Pearce and Ferrier, 2000b).

Since occurrence data were too scarce to partition into independent sets for training and testing, we used cross-validation (CV; Stone, 1974) to assess generality. First the data were partitioned into ten disjoint subsets of roughly equal size. The model was then fitted to nine of them and assessed using the withheld fraction as pseudo-independent test data – this step was repeated ten times, each time omitting a different fraction of data. The entire procedure was repeated 20 times and results were averaged to give the final cross-validation index (Kohavi, 1995). To ensure that the subsets of data used to train and test models reflected the true sample prevalence, partitions were stratified such that prevalence was approximately equal between folds (Parker *et al.*, 2007; see also Appendix 3B).

For a particular model, the severity of generalisation error can be gauged by comparison of the cross-validated and resubstituted AUC (subscripted 'CV' and 'RS'), where resubstitution refers to the reuse of training data for testing. For a standardised comparison across all models, we used the following measure.

$$\mathrm{GE} = \frac{\mathrm{AUC}_{\mathrm{RS}} - \mathrm{AUC}_{\mathrm{CV}}}{\mathrm{AUC}_{\mathrm{RS}} - 0.5}\ , \quad \mathrm{AUC}_{\mathrm{CV}} < \mathrm{AUC}_{\mathrm{RS}}\ (> 0.5)$$

That is, the proportion of above chance AUC that is lost under cross-validation. Lower values are best: $\mathrm{GE} \approx 0$ indicates a very stable model, provided that extrapolation sites are within the environmental range of the training data; $\mathrm{GE} \approx 1$ warns that discriminatory ability at unvisited sites could be no better than that of a null model; $\mathrm{GE} > 1$ only when $\mathrm{AUC}_{\mathrm{CV}} < 0.5$ (worse than chance).

Statistical calculations were performed in R 2.3.1 (R-Development-Core-Team, 2009) using functions from the 'gam' and 'ROC' packages, together with custom R code. For spatial analyses we used GRASS GIS 6.0 (GRASS-Development-Team, 2009). The manipulation of map layers and calibration of models were automated using shell scripts and executed in Windows XP *via* a Linux emulation layer (http://cygwin.com/).

Modelling experiments

The 201 sites were located on predictor maps and the corresponding cell values were extracted for model calibration. GAMs were then fitted to the data using a logit link and binomial error term (Yee and Mitchell, 1991). Given that response shapes can vary greatly in natural systems, both between species (Pearce and Ferrier, 2000a) and with respect to different environmental gradients (Austin, 2002), we determined the effective number of parameters for smoothers (degrees of freedom, df) separately for each species-predictor pair, such that df in [1, 4] at intervals of 0.25. Where df = 1 we fitted parametric curves in order to reduce the uncertainty of extrapolating smooth functions (Hastie and Tibshirani, 1990); for higher degrees of freedom the smooth terms were retained. A schematic summary of the modelling procedure is shown in Fig. 3.3; details of the experiments are as follows.

Selecting predictors

All predictor pairs were tested for collinearity using Pear- son's correlation coefficient (r). If two predictors were highly collinear [$\mathrm{abs}(r) > 0.7$] then the one that yielded the highest univariate $\mathrm{AUC}_{\mathrm{CV}}$ was entered for selection. The motivation for this step was to allow predictors conveying subtly different information (e.g., *tminc* and *tabsmin*) to be available for all species, without excessive overlap in the data – highly correlated data are not parsimonious and may bias selection (Cohen *et al.*, 2003; Graham, 2003). Other studies have used factor analysis to similar effect, reducing the full predictor set to a smaller number of uncorrelated factors (e.g., Bakkenes *et al.*, 2002). We experimented with a range of thresholds before deciding on the appropriate level [abs(r)] at which predictors should be separated. Fixing the threshold at 0.7 was found to create three distinct subsets, such that predictors were either uncorrelated with all others or belonged to one of two mutually exclusive correlation groups (Table 3.1). To dampen sensitivity to weaker correlations, and those too non-linear to be detected by the Pearson coefficient, we cross-validated stepwise procedures (Hastie *et al.*, 2001; Maggini *et al.*, 2006) and avoided hypothesis tests in favour of global measures of model performance (Anderson *et al.*, 2000).

Two stepwise selection procedures were employed to further promote parsimony amongst solutions. The first, forward-backward selection (denoted 'FB'), began with an empty predictor set, sequentially added/removed variables according to the resultant change in AIC, and was complete when AIC ceased to improve. After each selection step the generality of predictions was assessed, and the final model was that which achieved the highest AUC_{CV}. The formula for AIC consists of two terms: the first evaluates model fit using a log-likelihood function; the second is a penalty term proportional to the number of predictors in the model. Its purpose here was to identify a set of candidate models from which the most robust could be selected by cross-validation.

The second method was backward-forward selection (denoted 'BF'). This time the procedure began with a full model and variables were removed/added according to BIC (Bayesian Information Criterion; Akaike, 1978). Again, the final model was determined by AUC_{CV}. BIC was preferred here because it penalises large models more heavily than AIC, encouraging the removal of noise variables and the selection of more parsimonious solutions. A simple multimodel solution (denoted 'MM') was achieved by weighting the two best-model predictions according to their respective above chance AUC_{CV} values ($AUC_{CV} - 0.5$), and taking the average.

Weighting absences

In baseline models presence records and absence records were treated with equal confidence, assuming no contamination of one class by the other. With weighted models we attempted a more realistic portrayal of the data by placing greater emphasis on observed presences (*P*) than on absences inferred from plot data (*A*). This was achieved by the weighting absence data by a factor of *P*/*A* (< 1 for all species), forcing a standardised prevalence of 0.5. The intention was to tilt the balance of errors away from false negatives and toward false positives (McPherson *et al.*, 2004). Such a shift is desirable because a presence observation necessitates suitable conditions for growth, whilst an absence record could be a consequence of the restricted sampling regime, or of ecological factors beyond the scope of the model (Anderson, 2003).

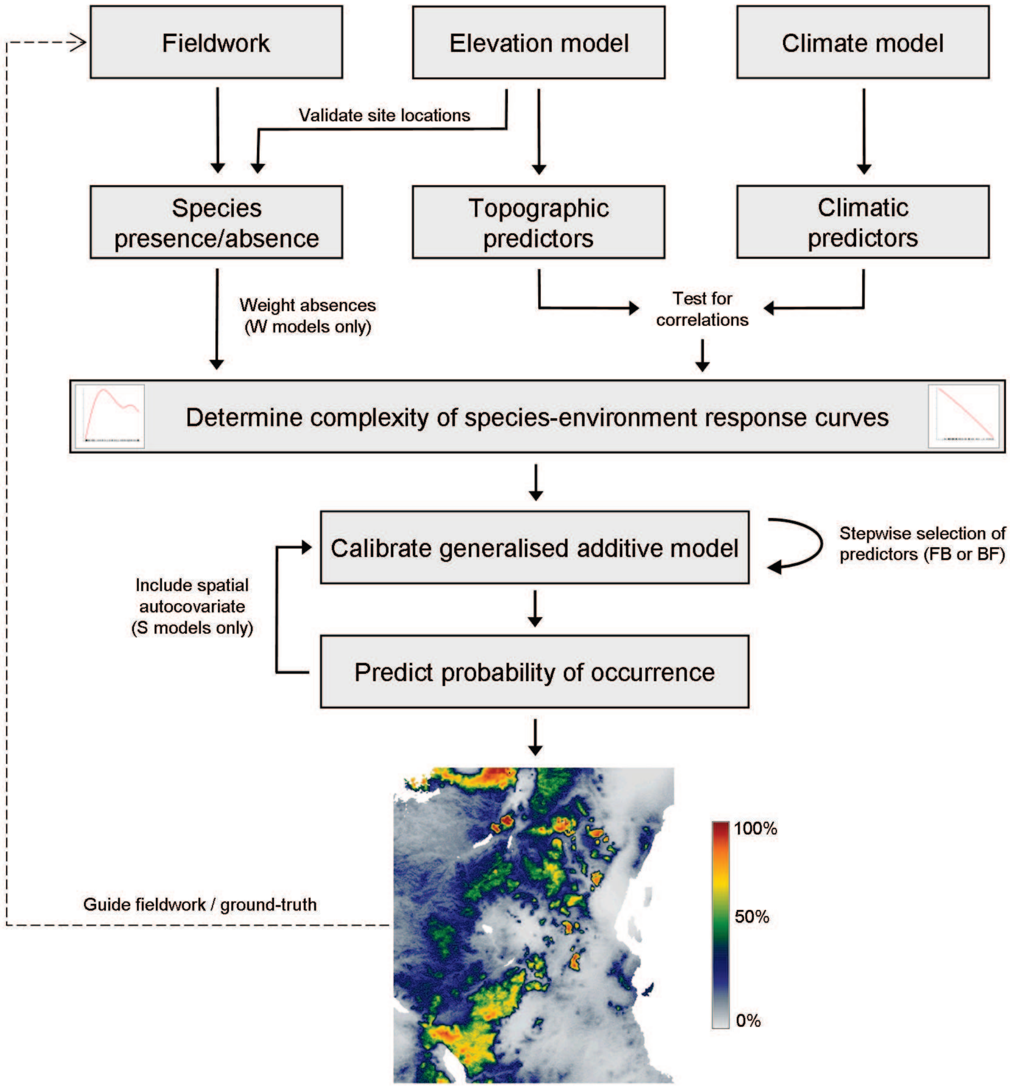


Figure 3.3. Schematic of the modelling procedure.

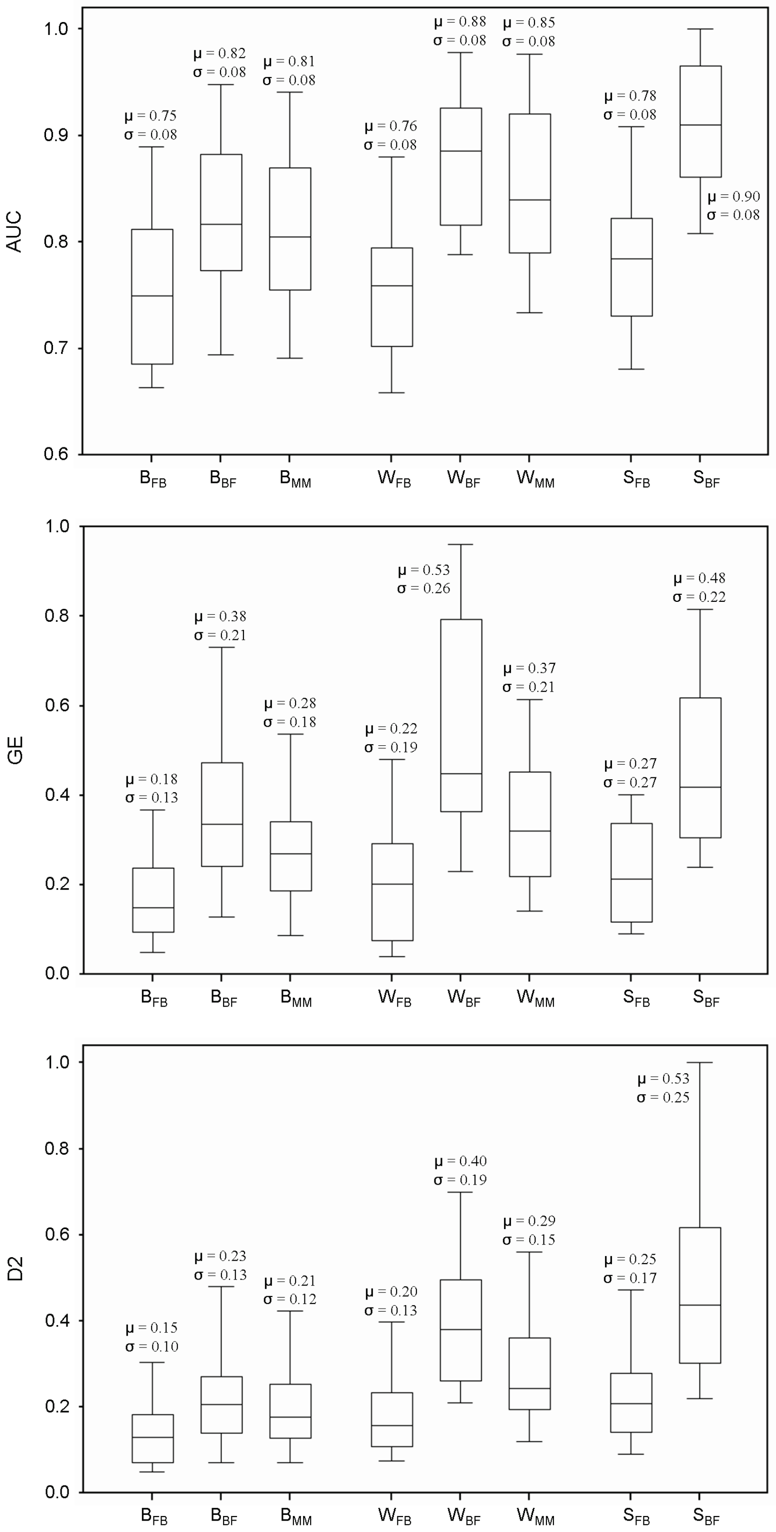


Figure 3.4. Boxplot comparisons of model performance in the different experiments. Box whiskers extend up to 150% of the interquartile range of each box. Upper: area under the receiver-operating characteristic curve (AUC). Middle: generalisation error (GE). Lower: proportion of deviance explained (D^2). Model type: B, baseline; W, weighted; S, spatial. Selection: FB, forward-backward; BF, backward-forward; MM, multimodel.

Spatial autocovariates

Autocovariate terms were used to describe fine-scale spatial clustering in species distributions. The first step was to obtain preliminary estimates of the distributions, for which we used weighted model predictions. Autocovariate terms were then derived such that each grid cell (i) was a distance-weighted average across a set of neighbours (k_i):

$$A_i = \frac{\sum_{j \in k_i} p_j d_{ij}^{-1}}{\sum_{j \in k_i} d_{ij}^{-1}}$$

where p_j is the probability of occurrence in neighbouring cell j, and d_{ij} is the Euclidean distance between i and j (Augustin *et al.*, 1996). Four autocovariates were calculated for each model, with neighbourhoods represented by squares of side 3, 5, 7 and 9 cells (2.8, 4.6, 6.4 and 8.3 km, respectively). The autocovariate included in the final model was that which led to the greatest improvement in explained deviance. Larger neighbourhoods were not included because seeds are typically heavy, limiting wind dispersal. Birds and mammals may carry fruits further, but successful establishment would be fragmented by the rapidly changing landscape and restricted environmental ranges of taxa. Spatial models were not constructed for multimodels because there was no formula to which to append the autocovariate.

Results

Baseline models (B)

Our interpretation of Swets (1988) analysis of the AUC measure is that for ecological studies a value in the range [0.7, 0.9) indicates a reasonable or good model and a value in the range [0.9, 1.0] indicates an excellent model, although any model with AUC > 0.5 should provide some discriminatory power. Following this classification for each of the 40 taxa, 27 forward-backward (FB) models and 36 backward-forward (BF) models produced reasonable, good or excellent predictions. For two species, *Syzygium cordatum* and *Tabernaemontana pachysiphon*, the FB solution was a null model (no predictors were added to the formula). For the same two species, BF returned non-trivial but highly unstable solutions (GE = 0.82 and 0.68, respectively). In general the FB method selected more parsimonious models with fewer predictors and better generalisation error. Conversely, BF models tended to be larger and better equipped to capture the observed deviance, though performance suffered under cross-validation (Table 3.2; Fig. 3.4). These two selection methods agreed for just five species. Baseline multimodels achieved a useful compromise, returning higher precision than FB models but with lower generalisation error than BF models. Moreover, for nearly half the taxa the multimodel produced higher AUC_{CV} than either of the best-model predictions.

Weighted models (W)

Applying weights to the absence data significantly altered the distribution of smoother complexity ($p < 0.01$): for baseline models the distribution was positively skewed, with simpler curves constructed for most predictors; for weighted models, the distribution was shifted in favour of more complex response shapes. This altered the predictor sets chosen by selection, and ultimately resulted in different spatial predictions (Table 3.2; Figs. 4 and 5). For FB selection, predictor sets chosen during weighted and baseline model calibration differed for 27 of the 40 taxa, with seven null models; for BF selection they differed for 26. Inspection of response curves showed that the change in smoother complexity had increased uncertainty, especially near the limits of the training data (Fig. 3.5).

Under FB selection the impact on the AUC was not significant, though other statistics revealed important differences: the mean proportion of errors that were false negatives decreased by 23% compared with baseline models (increased sensitivity), and the mean value of D^2 was significantly higher. Under BF selection, weighted models tended to fit the training data very well – all but one species (*S. cordatum*) achieved reasonable to excellent AUC and the mean value of D^2 was particularly high (Fig. 3.4). As for FB models the error distribution was also much improved, with a 46% reduction in the proportion of errors that were false negatives. BF models remained prone to over-fitting though, a problem that appears to have been exacerbated by the weights. Prediction error was again dampened by model averaging, with the weighted multimodel returning the highest mean AUC under cross-validation.

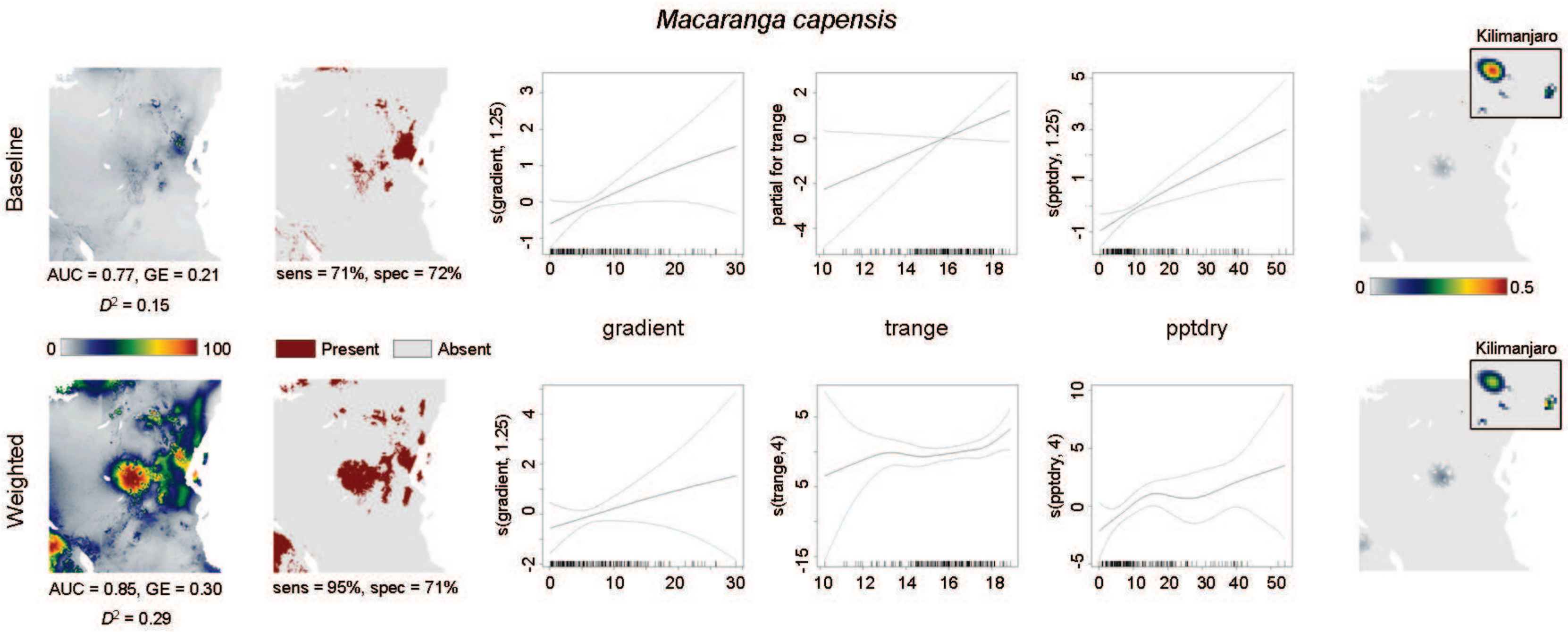


Figure 3.5. Baseline and weighted model predictions for *Macaranga capensis* (forward-backward selection). From left to right: probability of occurrence (%), presence-absence (maximising sum of sensitivity and specificity), response curves for selected predictors (including standard errors) and envelope uncertainty.

Table 3.2. Significance of differences between modelling experiments. With respect to the area under the receiver-operating characteristic curve (AUC), generalisation error (GE) and proportion of explained deviance (D^2). Model type: B, baseline; W, weighted; S, spatial. Selection method: FB, forward-backward; BF, backward-forward; MM, multimodel.

	AUC			GE			D^2		
Baseline selection									
	FB	BF	MM	FB	BF	MM	FB	BF	MM
FB	–	***	***	–	***	***	–	***	***
BF	–	–	ns	–	–	**	–	–	ns
Forward-backward models									
	B	W	S	B	W	S	B	W	S
B	–	ns	ns	–	ns	**	–	*	***
W	–	–	ns	–	–	ns	–	–	ns
Backward-forward models									
	B	W	S	B	W	S	B	W	S
B	–	***	***	–	***	**	–	***	***
W	–	–	ns	–	–	ns	–	–	*
Multimodels									
	B	W		B	W		B	W	
B	–	*		–	**		–	***	

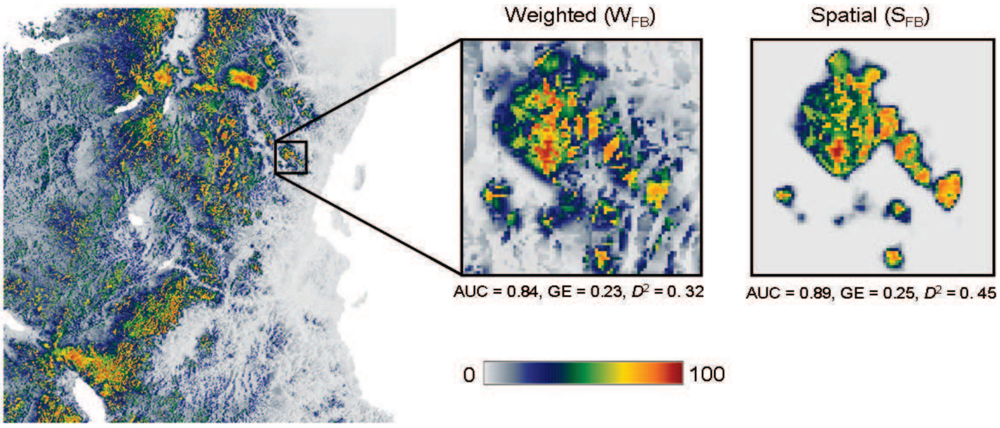
***, $p \leq 0.01$; **, $p \leq 0.05$; *, $p \leq 0.1$; ns, not significant (Wilcoxon rank sum tests, two-sided)

Spatial models (S)

The inclusion of a spatial autocovariate increased the pro- portion of explained deviance in all cases. Spatial models were significantly better at correctly predicting presences and absences (Table 3.3), and for BF selection they were also more stable (Fig. 3.4). Model size was typically larger in BF models and so climatic and topographic constraints were better represented alongside the autocovariate: the mean collective contribution of environmental predictors was 11% in FB models and 24% in BF models; the mean contribution of the autocovariate was 20% and 21%, respectively.

The neighbourhood size chosen for the autocovariate varied between species and no particular scale was superior overall (Appendix 3C). Visual inspection of the predicted distributions showed the environmental preferences of taxa to be more clearly delineated in spatial models: fine- scale patchiness in weighted model predictions had been smoothed, revealing what appear to be more realistic spatial patterns (e.g., Fig. 3.6a). There were some cases however where the incorporation of an autocovariate led to over-fitting. The spatial model for *Syzygium micklethwaitii* yielded a prediction with perfect discriminatory ability but high generalisation error (Fig. 3.6b). In this example the non-spatial model is more useful for inference since it retains a realistic gradient of suitability.

(a) *Neoboutonia macrocalyx*



(b) *Syzygium micklethwaitii*

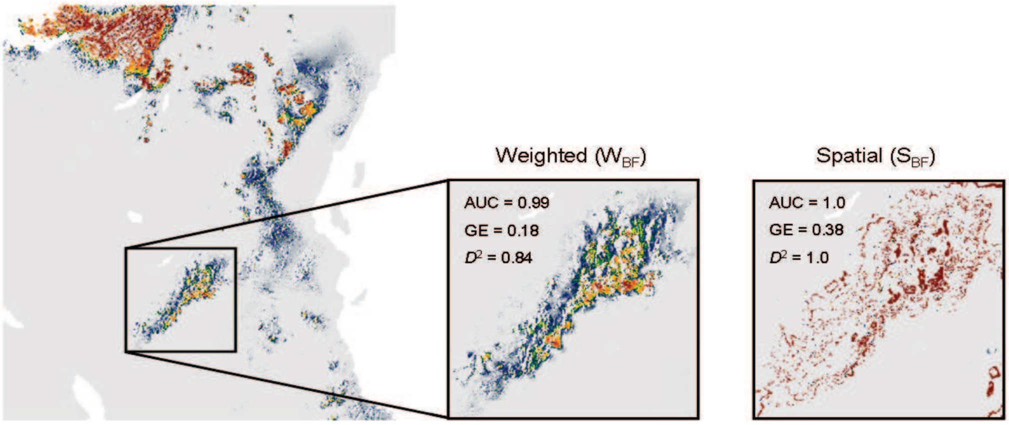


Figure 3.6. Weighted and spatial predictions for (a) *Neoboutonia macrocalyx*, focussing on the Usambara Mountains, and (b) *Syzygium micklethwaitii*, focussing on the Udzungwa Mountains. Scale bar shows probability of occurrence (%).

Table 3.3. Performance of predictors in non-spatial (baseline and weighted) models. Percentage selection rate in group, mean contribution to model, mean contribution of other covariates when selected. Predictor contributions are defined as the percentage drop in explained deviance when removed from the final model, standard errors in parentheses.

Predictor	Selection rate	Contribution (s.e.)	Covariate contribution
Independent			
gradient	–	19.61 (1.44)	18.34
trasp	–	18.75 (1.46)	18.50
trange	–	19.66 (1.47)	18.54
pptdry	–	19.85 (1.45)	19.58
mean	–	19.47 (1.46)	18.74
Group 1			
pptann	49.41	22.30 (2.52)	19.09
pptwet	50.59	18.05 (2.10)	18.14
mean	–	20.18 (2.31)	18.62
Group 2			
tmean	6.35	33.83 (6.12)	27.59
tmeanw	5.56	25.31 (4.88)	21.21
tmeanc	3.17	14.60 (5.16)	28.57
tmaxw	7.14	23.68 (5.28)	20.00
tminc	15.87	18.28 (2.00)	24.10
tabsmin	8.73	18.53 (3.10)	26.19
tdays	7.94	36.09 (6.01)	26.32
pet	26.98	19.50 (2.39)	26.56
ami	18.25	16.15 (2.39)	22.12
mean	–	22.89 (4.15)	24.74

Total times selected: Independent, 343; Group 1, 85; Group 2, 126

Predictors and envelope uncertainty

Both topographic variables were independent of correlation groups, as were *trange* and *pptdry*. These four predictors were the most frequently selected for inclusion in the final model, and each contributed similarly to model performance. The most popular predictor overall was *pptdry*. In the first correlation group, *pptann* and *pptwet* were chosen a similar number of times, though *pptann* was marginally better at explaining deviance. In the second correlation group *tmeanc* contributed the least to D^2 and was the least frequently selected, often dropped in favour of similar measures such as *tminc* and *tabsmin*. Predictors that describe moisture availability, such as *pet* and *ami*, were often included but their mean contributions were below average. The most successful descriptor of deviance was the temperature-days variable, which when removed from the final models resulted in a mean drop in D^2 of 36% (Table 3.3).

The environmental range of survey sites was generally good, with the exception that climatic predictors lacked coverage near the summit of Mt. Kilimanjaro (Table 3.1; Fig. 3.5). The only notable shortfalls were for *trange* and *pptwet*: both were within the calibration envelope for most of the study region, but *trange* was up to 30% beyond the envelope near the Maasai Steppe, and *pptwet* was up to 9.2% beyond the envelope for a small area south of the Pare Mountains (Fig. 3.2).

Discussion

The potential of GAMs to estimate the distribution of EAM trees shows promise, with a number of models achieving a high level of predictive success. However, it is clear from our experiments that the distributions predicted are highly sensitive to the modelling method employed. Selection procedures frequently disagreed, produced different spatial predictions, and yet often returned similar validation scores. These findings illustrate the importance of understanding the biases imposed by the selection procedure in use, and of not relying solely on validation scores as evidence of good model performance – consideration should also be given to the chosen predictor set and spatial patterns predicted. Whittingham *et al*. (2006) advise against the use of stepwise procedures, arguing that there is rarely a true 'best model' for selection to identify and that different predictor sets are likely to explain the response equally well. This conjuncture is supported by our experiments, though we suggest that studies with access to more extensive distribution data are likely to find greater agreement between selection methods.

Whilst forward-backward models often lacked precision, particularly if neither topographic predictor was selected, they invariably produced stable predictions using minimal predictor sets, and are therefore likely to be more useful than backward-forward methods for inferring causal relation- ships. Backward-forward selection described the data well but retained too many predictors to avoid over-fitting. Given the disagreement between selection procedures, there is a good argument for favouring expert opinion over computer selection (but see:

Pearce *et al.*, 2001; Seoane *et al.*, 2005). Regardless, automated procedures remain necessary when deriving models for a large number of taxa, particularly if their ecologies are not well known. The multimodel, averaging forward-backward and backward-forward predictions according to their relative cross-validated performance, identified a superior trade-off between generality and precision that in many cases outperformed both conventional selection procedures (higher AUC_{CV}). The weakness of this method is the need to compute two sets of predictions, increasing computation time. However, if models are lacking, either in fit or stability, we think it prudent to investigate other selection options as a matter of course, in which case the computation of a model average would be trivial. Other kinds of multimodel have also produced favourable results (Anderson *et al.*, 2000; Johnson and Omland, 2004; Hartley *et al.*, 2006; Dormann *et al.*, 2008), and appear to be generally superior to best-model approaches for predictive purposes.

Both the performance and reliability of models were correlated with sample prevalence, such that low prevalence led to more discriminative but less stable models. This may reflect the fact that restricted range tree species in the EAMs typically cover a narrow altitudinal range (Lovett, 1996; Lovett *et al.*, 2001), making their climatic preferences easier for models to capture but rendering them highly sensitive to errors in the distribution data. The relationship was found to be stronger for weighted models than for baseline models due to the higher levels over-fitting. In contrast to our results, Maggini *et al.* (2006) found that weighting absences improved model performance without impairing stability. A probable reason for the difference in our results is the comparatively high instability of our baseline models: the more robust baseline predictions tended to remain stable in the weighted experiment, whereas those with high GE suffered from further over-fitting. We did observe a useful shift in the error distribution, such that weighted model predictions were much less likely to contain false negative errors. During recent field expeditions to Nguu and North Pare (Fig. 3.1), we found that the higher sensitivity of weighted models gave a better indication of the actual forest distributions, especially for spatial models. Further ground-truthing of this result is encouraged.

In the spatial experiment we aimed to improve weighted model predictions by including an autocovariate to account for fine-scale spatial clustering. In agreement with previous applications of this technique (e.g., Augustin *et al.*, 1996; Segurado and Araújo, 2004), spatial models fitted the training data more accurately than non-spatial models and were superior for describing fine-scale patterns in distribution. Where model size was large (five or six predictors) spatial models also had lower generalisation error. For smaller model sizes though the contribution of autocovariates in explaining deviance was around twice that of environmental predictors, which may be a cause for concern given recent suggestions that autocovariates can lead to biased predictions (Dormann, 2007a; Dormann *et al.*, 2007). One should certainly be sceptical of extrapolations into different points in time (e.g., historical reconstruction or climate change studies) since spatial dependencies could well be different (Guisan and Thuiller, 2005). The degree to which predictions of this kind can be truly representative of the actual distributions will always be uncertain, because we cannot be sure to what extent a species realised ecological niche is restricted by its environmental tolerance and to what extent by competition/mutualism with other species (Pulliam, 2000). Community interactions are expected to play an important role in such an ancient ecosystem, though the relevant spatial scales are not well understood. In New Zealand's old-growth forests, attempts have been made to model competition between tree species using logistic regression: Leathwick and Austin (2001) found that including the presence/absence of one species as a covariate alongside climatic constraints could improve the predicted distribution of another. This study was based on community compositional data for just two competing species; here we are concerned with a web of interactions involving tens, possibly hundreds of taxa, presenting a far greater challenge for modelling.

The optimum neighbourhood size for a particular tree often differed according to the selection procedure used to obtain the initial prediction, and so did not provide much insight into the processes underpinning SAC. This might simply reflect the high variability in predictor sets chosen by selection and the associated omission of different autocorrelates (Lichstein *et al.*, 2002). It may also be the case that no single range-specific autocovariate could carry sufficient information to identify the true scales at which aggregation occurs (van Teeffelen and Ovaskainen, 2007). Dispersal is one factor known to drive spatial patterns, but this mechanism is under-researched in the EAMs and few empirical data are currently available for parameterising/validating models. We also draw attention to the fact that clustering was assumed by autocovariates to be roughly isotropic, i.e.,, apparent in equal measure in all directions (neighbourhoods were approximated by squares). However, spatial patterns may actually be elongated in some areas as a result of elevational migration. It was further assumed that SAC operates similarly in different regions, which is unlikely to be the case given the high topographic heterogeneity of the study area. The possibility that regression parameters could reflect local rather than global trends has been investigated by Fotheringham *et al.* (2002) and appears to work well in some settings (e.g., Brunsdon *et al.*, 2007), though we have reservations as to the suitability of geographically weighted regression for our dataset (*cf.* Austin, 2007).

Dichotomising probabilities of occurrence using the sensitivity-specificity threshold, we compared the different areas of occupancy forecast by models (Appendix 3D). On average, the number of cells predicted present was similar across the baseline, weighted and spatial experiments, despite often large differences in occurrence probabilities (e.g., Fig. 3.5). In the selection experiment, forward-backward models were the least well constrained by environmental variables, resulting in the greatest areas of predicted occupancy. Interestingly, the number of cells predicted present by the multimodels was similar to that forecast by backward-forward models. Multimodels contained the most information and also gave the highest AUC under cross-validation, and so we are inclined to trust the magnitude of backward-forward area pre- dictions more than those yielded by the forward-backward method.

The frequency with which moisture related variables were selected by models is not surprising given that the EAMs are under the direct climatic influence of the Indian Ocean (Lovett, 1990; Marchant *et al.*, 2007). The gradient of the slope, in addition to indicating moisture availability *via* its relationship with run-off, also helps to distinguish montane habitats from the surrounding lowlands. The wetness/radiation index performed well, reflecting the importance of moisture carried by trade winds during the drier months. Response curves constructed for climatic predictors were not calibrated with respect to the environmental extremes found near the summit of Mt. Kilimanjaro, and so predictions of occurrence in these grid cells are subject to high uncertainty. The sample sites were otherwise found to cover a wide breadth of environmental conditions and were generally representative of the Eastern Arc region. Where extrapolations beyond the envelope occurred, the worst affected models were those that relied heavily on the predictor contributions of annual temperature range and rainfall during the wettest month. The shortfall in these predictors had the greatest impact on forward-backward models, where model size was smaller. For *Macaranga capensis* we found that the highest occurrence probabilities were obtained by extrapolating beyond the range of the training data. The weighted model in particular predicts that this pioneer tree, usually associated with submontane and riverine forests (Lovett *et al.*, 2006), should also be suited to the Maasai Steppe, a lowland savannah habitat. The EUM confirms that the grid cells with the highest envelope uncertainty correspond precisely with the region deemed most favourable by the model. Here inference can only be made after inspecting response curves beyond the limits of the training data. Because EUMs pinpoint the locations where a model may be weakest, we suggest they might also be useful in targeting field sampling in a way that most improves data quality.

Conclusions

GAMs can provide useful information for conservation in the EAMs, even when the frequency of documented occurrence is low. Indeed it was the modelling method employed and the quality not quantity of distribution data that mattered most. However, there were a number of instances where over-fitting seriously compromised the generality of predictions, and we recommend that the application of GAMs to small datasets be approached with care. If over-fitting cannot be avoided, then the parametric terms of generalised linear models should be considered in preference to data-driven smoothers. With respect to the different methodologies investigated, our main observations are as follows.

1. Forward-backward selection is less discriminative than backward-forward selection, but is the more useful of the two for explanatory purposes. Backward-forward selection retains more ecologically relevant detail but can suffer from high prediction error. Multimodels provide a useful compromise, and are arguably the best choice for predictive purposes.
2. Models calibrated with weighted absence data are superior in terms of overall accuracy and have better sensitivity, though they can be especially vulnerable to over-fitting if the distribution data are not well described by environmental predictors.
3. Including a spatial autocovariate improves model fit and better represents spatial clustering in predictions; the stability of models may however suffer if environmental constraints are inadequately represented.
4. Envelope uncertainty maps display important information that should be taken into account when drawing inference from predictions, especially if a model is to be extrapolated into novel parameter space.

This work involved the parameterisation of environmental response functions for 40 taxa of large tree, targeted for modelling because of historical patterns occurrence (Mumbi *et al.*, 2008), endemism (Lovett *et al.*, 2006) and conservation interest (http://iucnredlist.org/). It is hoped that further analysis of response shapes will add to our understanding of their habitat preferences, and specifically the degree to which environmental controls restrict their distributions.

Acknowledgements

This research was funded by the Marie-Curie Excellence programme of the European 6th Framework under contract MEXT-CT-2004-517098 to R. Marchant. We are grateful to A. Ahrends, M. Bulling, J. Finch and A. Jump for commenting on previous versions of this manuscript, and for the helpful feedback of two anonymous reviewers.

Author contributions

The study was conceived by P.J.P., who programmed and implemented the modelling experiments, analysed the results and prepared the manuscript. Plot data was provided by J.C.L. The work was supervised by R.M. and C.J.M.

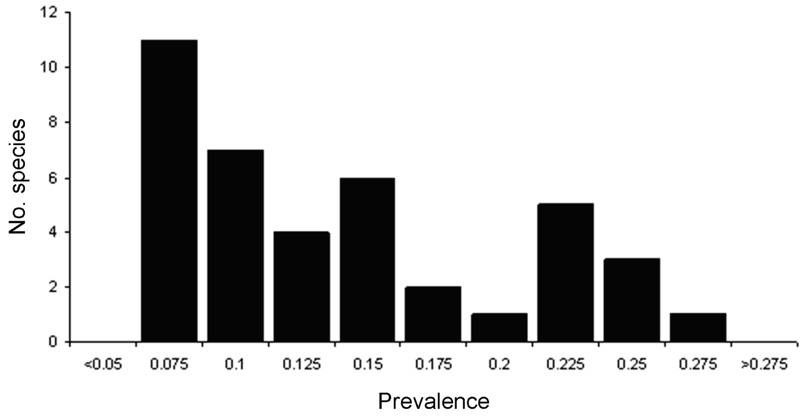
Appendix 3A. Tree species modelled

Species code	Family	Species/taxon name	Presences (prevalence)
sp1	Anacardiaceae	*Sorindeia madagascariensis* Thouars ex DC.	41 (0.20)
sp2	Annonaceae	*Lettowianthus stellatus* Diels	11 (0.05)
sp3	Apocynaceae	*Tabernaemontana pachysiphon* Stapf	21 (0.10)
sp4	Aquifoliaceae	*Ilex mitis* (L.) Radlk.	18 (0.09)
sp5	Araliaceae	*Polyscias stuhlmannii* Harms	15 (0.07)
sp6	Araliaceae	*Polyscias fulva* (Hiern) Harms	41 (0.20)
sp7	Celastraceae	*Maytenus acuminata* (L.f.) Loes.	29 (0.14)
sp8	Chrysobalanaceae	*Parinari excelsa* Sabine	41 (0.20)
sp9	Combretaceae	*Terminalia sambesiaca* Engl. & Diels	15 (0.07)
sp10	Euphorbiaceae	*Croton macrostachyus* Hochst. ex Delile	17 (0.08)
sp11	Euphorbiaceae	*Drypetes gerrardii* Hutch.	28 (0.14)
sp12	Euphorbiaceae	*Drypetes natalensis* (Harv.) Hutch.	11 (0.05)
sp13	Euphorbiaceae	*Drypetes usambarica* (Pax) Hutch.	16 (0.08)
sp14	Euphorbiaceae	*Macaranga capensis* (Baill.) Benth. ex Sim var. *capensis*	21 (0.10)
sp15	Euphorbiaceae	*Macaranga capensis* (Baill.) Benth. ex Sim var. *kilimandscharica* (Pax) Friis & M. G. Gilbert	41 (0.20)
sp16	Euphorbiaceae	*Neoboutonia macrocalyx* Pax	17 (0.08)
sp17	Lauraceae	*Ocotea usambarensis* Engl.	41 (0.20)
sp18	Leguminosae (Fabaceae): Mimosoideae	*Newtonia buchananii* (Baker f.) G. C. C. Gilbert & Boutique	48 (0.24)
sp19	Loganiaceae	*Anthocleista grandiflora* Gilg	21 (0.10)
sp20	Loganiaceae	*Nuxia congesta* R. Br. ex Fresen.	40 (0.20)
sp21	Meliaceae	*Trichilia emetica* Vahl.	10 (0.05)
sp22	Monimiaceae	*Xymalos monospora* (Harv.) Warb.	46 (0.23)
sp23	Myricaceae	*Morella salicifolia* (Hochst. ex A. Rich.) Verdc. & Polhill	28 (0.14)
sp24	Myrsinaceae	*Myrsine melanophloeos* (L.) R.Br.	51 (0.25)
sp25	Myrtaceae	*Syzygium guineense* (Willd.) DC. subsp. *afromontanum* F. White	47 (0.23)
sp26	Myrtaceae	*Syzygium micklethwaitii* Verdc.	14 (0.07)
sp27	Myrtaceae	*Syzygium cordatum* Hochst.	13 (0.06)
sp28	Oleaceae	*Olea capensis* (L.) subsp. *macrocarpa* (C. H. Wright) I. Verd.	29 (0.14)
sp29	Oleaceae	*Olea europea* (L.) subsp. *cuspidata* (Wall. ex G. Don) Cif.	12 (0.06)
sp30	Palmae (Arecaceae)	*Phoenix reclinata* Jacq.	27 (0.13)
sp31	Podocarpaceae	*Afrocarpus falcatus* (Thunb.) C. N. Page	12 (0.06)
sp32	Podocarpaceae	*Podocarpus latifolius* (Thunb.) R.Br. ex Mirb.	35 (0.17)
sp33	Rosaceae	*Hagenia abyssinica* J.F. Gmel.	18 (0.09)
sp34	Rosaceae	*Prunus africana* (Hook.f.) Kalkman	29 (0.14)
sp35	Sapindaceae	*Dodonea viscosa* Jacq.	14 (0.07)
sp36	Sapotaceae	*Chrysophyllum gorungosanum* Engl.	33 (0.16)
sp37	Sterculiaceae	*Dombeya torrida* (J.F. Gmel.) Bamps	18 (0.09)
sp38	Sterculiaceae	*Leptonychia usambarensis* K. Schum.	19 (0.09)
sp39	Ulmaceae	*Celtis africana* Burm.f.	11 (0.05)
sp40	Ulmaceae	*Trema orientalis* (L.) Blume	21 (0.10)

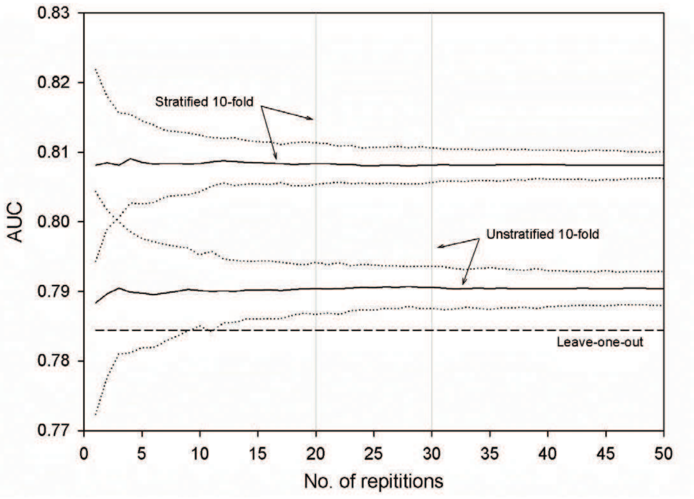
Nomenclature follows Lovett *et al*. (2006).

Appendix 3B. Stratified cross-validation

Given the low number of presences for many taxa and the use of cross-validation during model calibration, it was important to ensure that the subsets of data used to train and test models reflected the true sample prevalence. Stratifying cross-validation such that prevalence was consistent between partitions reduced both variance and pessimistic bias in the estimation of the AUC. The ten-fold method was preferred to leave-one-out cross-validation (LOOCV) because it provides a more robust measure of model stability, particularly when repeated multiple times. Our tests with LOOCV showed that for baseline models it was also marginally more pessimistic than ten-fold, despite of the greater proportion of data available for training (Appendix 3D). For weighted models the values of AUC_{CV} and AUC_{LOOCV} were closer, suggesting that down-weighting absences might reduce stratification bias.



Above. Histogram of prevalence across the 201 sample locations. The lowest prevalence permitted by our modelling procedure was 0.05 (10 presences), because stratified ten-fold cross-validation requires that each fold contain at least one presence record.



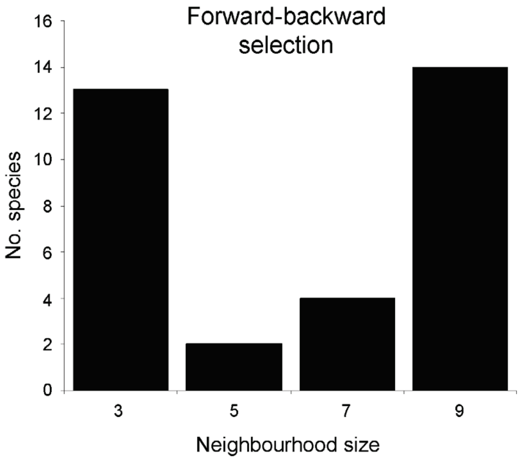
Left. Comparison of cross-validation methods, using *Drypetes natalensis* as an example. How many repetitions are required for consistent results? Averaged over 100 runs for stratified and unstratified cross-validation, results show mean AUC values ± ten standard errors. Dashed line shows the result from leave-one-out cross-validation (*N*-fold). Vertical lines indicate the number of repetitions required for s.e. < 0.0003.

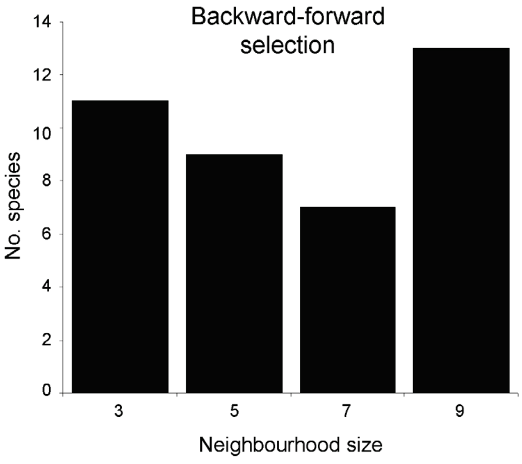
Appendix 3C. Neighbourhood size for spatial autocovariates

Neighbourhood sizes that minimised explained deviance in the spatial experiment (width of neighbourhood square, number of cells). Refer to Appendix 3A for species names.

Species code	Neighbourhood size		Species code	Neighbourhood size	
	Forward-backward	Backward-forward		Forward-backward	Backward-forward
sp1	5	9	sp21	9	9
sp2	3	3	sp22	9	9
sp3	5	9	sp23	9	3
sp4	–	5	sp24	3	3
sp5	9	7	sp25	3	5
sp6	3	5	sp26	–	3
sp7	7	9	sp27	9	3
sp8	9	9	sp28	7	5
sp9	3	9	sp29	–	3
sp10	9	5	sp30	3	9
sp11	9	5	sp31	3	7
sp12	–	7	sp32	3	3
sp13	3	3	sp33	–	5
sp14	3	5	sp34	7	9
sp15	9	9	sp35	9	7
sp16	9	9	sp36	–	7
sp17	7	3	sp37	3	7
sp18	9	5	sp38	–	9
sp19	9	7	sp39	3	3
sp20	3	3	sp40	9	9

Forward-backward: mean = 6.16, median = 7
Backward-forward: mean = 6.10, median = 6





Appendix 3D. Species-specific results

Results from modelling experiments, detailed by species (see Appendix 3 for species names). From left to right: area under the receiver-operating characteristic curve (AUC_{RS}, resubstituted training data; AUC_{CV}, stratified 10-fold cross-validation; AUC_{LOOCV}, leave-one-out cross-validation); generalisation error (GE); explained deviance (D^2); presence/absence cut-point; sensitivity and specificity at this cut-point; number of grid cells predicted suitable for the species.

Baseline models: forward-backward selection

Species code	AUC$_{RS}$	AUC$_{CV}$	AUC$_{LOOCV}$	GE	D^2	Cut-point	Sensitivity	Specificity	Suitable cells
sp1	0.74	0.67	0.66	0.29	0.12	0.11	0.81	0.60	21394
sp2	0.72	0.65	0.63	0.30	0.07	0.05	0.82	0.61	421148
sp3	0.67	0.63	0.61	0.24	0.06	0.16	0.79	0.51	223438
sp4	0.68	0.61	0.58	0.40	0.06	0.09	0.71	0.64	654270
sp5	0.69	0.67	0.64	0.10	0.07	0.08	0.71	0.59	681076
sp6	0.67	0.64	0.61	0.20	0.06	0.08	0.83	0.51	250372
sp7	0.80	0.75	0.75	0.17	0.18	0.13	0.86	0.61	86333
sp8	0.84	0.81	0.78	0.09	0.20	0.09	0.91	0.81	170296
sp9	0.71	0.58	0.55	0.62	0.11	0.10	0.44	0.89	5435
sp10	0.92	0.86	0.86	0.15	0.43	0.14	0.89	0.92	293153
sp11	0.68	0.65	0.62	0.16	0.05	0.11	0.56	0.73	52242
sp12	0.81	0.70	0.68	0.36	0.17	0.10	0.79	0.70	62036
sp13	0.88	0.83	0.82	0.15	0.25	0.05	0.91	0.80	502100
sp14	0.77	0.71	0.70	0.21	0.15	0.10	0.71	0.72	34261
sp15	0.75	0.74	0.73	0.03	0.14	0.10	0.98	0.39	568572
sp16	0.66	0.65	0.63	0.08	0.07	0.10	0.93	0.37	565894
sp17	0.76	0.71	0.70	0.16	0.13	0.10	0.96	0.47	336861
sp18	0.69	0.68	0.67	0.05	0.09	0.22	0.84	0.50	425298
sp19	0.83	0.80	0.79	0.10	0.21	0.07	0.94	0.67	132737
sp20	0.75	0.70	0.69	0.21	0.14	0.22	0.85	0.61	32667
sp21	0.78	0.74	0.74	0.12	0.17	0.24	0.80	0.71	171969
sp22	0.82	0.79	0.79	0.09	0.25	0.16	0.88	0.63	94454
sp23	0.76	0.72	0.72	0.14	0.16	0.25	0.55	0.90	205934
sp24	0.89	0.88	0.88	0.03	0.30	0.15	0.83	0.91	295372
sp25	0.71	0.67	0.66	0.20	0.11	0.28	0.54	0.86	10426
sp26	0.71	0.61	0.58	0.46	0.07	0.15	0.78	0.61	439504
sp27	0.67	0.63	0.59	0.24	0.05	0.05	0.83	0.47	345107
sp28	0.78	0.71	0.70	0.25	0.17	0.20	0.74	0.67	157091
sp29	0.91	0.86	0.86	0.13	0.43	0.22	0.80	0.92	27642
sp30	0.78	0.73	0.72	0.17	0.13	0.30	0.56	0.89	217405
sp31	0.82	0.78	0.77	0.13	0.23	0.11	0.90	0.61	448853
sp32	0.75	0.74	0.73	0.03	0.13	0.18	0.83	0.60	550821
sp33	–	–	–	–	–	–	–	–	–
sp34	0.65	0.64	0.63	0.08	0.05	0.20	0.77	0.48	342629
sp35	0.89	0.84	0.84	0.12	0.33	0.09	0.86	0.82	24942
sp36	–	–	–	–	–	–	–	–	–
sp37	0.74	0.72	0.70	0.11	0.11	0.05	1.00	0.45	680240
sp38	0.60	0.57	0.52	0.32	0.02	0.09	0.81	0.36	273575
sp39	0.65	0.64	0.64	0.05	0.07	0.10	0.40	0.93	107119
sp40	0.72	0.70	0.69	0.10	0.11	0.17	0.89	0.47	355242
mean	0.75	0.71	0.70	0.18	0.15	0.14	0.79	0.66	270208
median	0.75	0.71	0.69	0.15	0.13	0.11	0.82	0.62	236905

Baseline models: backward-forward selection

Species code	AUC_{RS}	AUC_{CV}	AUC_{LOOCV}	GE	D^2	Cut-point	Sensitivity	Specificity	Suitable cells
sp1	0.84	0.71	0.70	0.39	0.22	0.08	0.90	0.63	43052
sp2	0.72	0.65	0.63	0.30	0.07	0.05	0.82	0.61	421148
sp3	0.69	0.63	0.61	0.32	0.07	0.18	0.73	0.61	178928
sp4	0.80	0.59	0.58	0.69	0.15	0.11	0.71	0.78	403058
sp5	0.69	0.67	0.64	0.13	0.07	0.08	0.71	0.59	681076
sp6	0.78	0.58	0.53	0.74	0.14	0.10	0.89	0.62	365584
sp7	0.84	0.73	0.72	0.34	0.23	0.15	0.82	0.72	63670
sp8	0.84	0.81	0.78	0.09	0.20	0.09	0.91	0.81	170296
sp9	0.81	0.53	0.51	0.90	0.22	0.06	0.88	0.59	145052
sp10	0.95	0.85	0.85	0.23	0.51	0.17	0.89	0.94	288678
sp11	0.92	0.79	0.78	0.31	0.37	0.12	0.94	0.80	51715
sp12	0.88	0.69	0.68	0.51	0.27	0.10	0.95	0.72	80009
sp13	0.88	0.82	0.81	0.16	0.26	0.05	0.91	0.77	472874
sp14	0.88	0.74	0.72	0.38	0.30	0.14	0.81	0.84	47054
sp15	0.78	0.69	0.68	0.33	0.19	0.17	0.88	0.57	274456
sp16	0.76	0.62	0.61	0.53	0.13	0.18	0.66	0.76	46086
sp17	0.82	0.71	0.70	0.34	0.20	0.17	0.79	0.73	104347
sp18	0.78	0.69	0.68	0.32	0.20	0.28	0.78	0.64	143382
sp19	0.90	0.79	0.79	0.26	0.31	0.11	0.94	0.79	56596
sp20	0.80	0.69	0.69	0.37	0.19	0.21	0.85	0.65	70541
sp21	0.82	0.73	0.72	0.28	0.23	0.24	0.83	0.72	94813
sp22	0.89	0.79	0.79	0.25	0.36	0.18	0.93	0.76	85960
sp23	0.82	0.73	0.73	0.28	0.24	0.31	0.55	0.94	138977
sp24	0.95	0.83	0.83	0.26	0.49	0.22	0.92	0.96	265119
sp25	0.78	0.66	0.66	0.42	0.18	0.24	0.66	0.79	25125
sp26	0.83	0.66	0.66	0.51	0.21	0.13	0.85	0.68	99850
sp27	0.95	0.75	0.74	0.44	0.43	0.16	0.92	0.91	15136
sp28	0.84	0.69	0.69	0.44	0.24	0.14	0.94	0.60	223287
sp29	0.97	0.86	0.87	0.24	0.59	0.10	1.00	0.87	33842
sp30	0.82	0.70	0.69	0.36	0.20	0.23	0.85	0.68	251412
sp31	0.85	0.77	0.77	0.23	0.27	0.11	0.93	0.63	448802
sp32	0.81	0.74	0.73	0.23	0.21	0.14	0.93	0.59	301476
sp33	0.76	0.55	0.52	0.82	0.13	0.06	0.85	0.59	200723
sp34	0.65	0.64	0.63	0.09	0.05	0.20	0.77	0.48	342629
sp35	0.97	0.90	0.90	0.14	0.55	0.12	0.93	0.90	53470
sp36	0.73	0.57	0.56	0.68	0.13	0.11	0.71	0.68	248870
sp37	0.81	0.66	0.65	0.48	0.18	0.06	0.93	0.58	369912
sp38	0.77	0.55	0.53	0.83	0.13	0.12	0.86	0.73	186184
sp39	0.65	0.65	0.64	0.04	0.07	0.10	0.40	0.93	107119
sp40	0.76	0.67	0.67	0.34	0.15	0.29	0.65	0.77	94322
mean	0.82	0.70	0.69	0.38	0.23	0.15	0.83	0.72	192366
median	0.82	0.69	0.69	0.34	0.21	0.14	0.86	0.72	144217

Baseline multimodels

Species code	AUC_{RS}	AUC_{CV}	AUC_{LOOCV}	GE	D^2	Cut-point	Sensitivity	Specificity	Suitable cells
sp1	0.81	0.70	0.69	0.36	0.19	0.12	0.76	0.73	17777
sp2	0.72	0.65	0.63	0.31	0.07	0.05	0.82	0.61	421148
sp3	0.68	0.63	0.61	0.28	0.07	0.15	0.82	0.49	258707
sp4	0.76	0.61	0.57	0.60	0.11	0.09	0.82	0.66	629437
sp5	0.69	0.67	0.64	0.09	0.07	0.08	0.71	0.59	681076
sp6	0.74	0.61	0.58	0.53	0.09	0.08	0.94	0.56	353643
sp7	0.83	0.74	0.73	0.27	0.21	0.14	0.86	0.68	70896
sp8	0.84	0.81	0.78	0.09	0.20	0.09	0.91	0.81	170296
sp9	0.76	0.55	0.52	0.83	0.15	0.07	0.63	0.74	83035
sp10	0.94	0.86	0.86	0.19	0.48	0.15	0.89	0.93	303473
sp11	0.92	0.79	0.78	0.31	0.30	0.12	0.89	0.84	40495
sp12	0.87	0.70	0.70	0.45	0.23	0.12	0.89	0.78	53927
sp13	0.88	0.83	0.81	0.12	0.26	0.05	0.91	0.78	490520
sp14	0.85	0.74	0.72	0.32	0.25	0.15	0.71	0.83	18404
sp15	0.77	0.72	0.71	0.19	0.16	0.17	0.85	0.56	294318
sp16	0.73	0.64	0.62	0.37	0.11	0.16	0.79	0.60	102968
sp17	0.80	0.72	0.70	0.27	0.18	0.11	0.96	0.56	230123
sp18	0.76	0.69	0.68	0.26	0.16	0.24	0.88	0.55	272413
sp19	0.88	0.81	0.79	0.19	0.27	0.09	0.94	0.75	76974
sp20	0.80	0.71	0.71	0.30	0.17	0.28	0.75	0.76	18647
sp21	0.81	0.74	0.73	0.20	0.21	0.25	0.83	0.75	118974
sp22	0.88	0.80	0.80	0.20	0.32	0.20	0.88	0.75	65364
sp23	0.80	0.73	0.72	0.23	0.22	0.25	0.59	0.89	176967
sp24	0.91	0.88	0.88	0.07	0.42	0.14	0.92	0.92	332387
sp25	0.76	0.67	0.67	0.34	0.15	0.27	0.56	0.82	11078
sp26	0.81	0.67	0.66	0.45	0.17	0.13	0.85	0.66	161996
sp27	0.94	0.79	0.77	0.34	0.36	0.14	0.92	0.94	13563
sp28	0.82	0.71	0.71	0.35	0.21	0.19	0.80	0.69	154268
sp29	0.97	0.87	0.88	0.20	0.54	0.18	0.93	0.90	27309
sp30	0.82	0.72	0.72	0.31	0.18	0.28	0.73	0.83	184573
sp31	0.84	0.78	0.78	0.17	0.25	0.11	0.93	0.60	450491
sp32	0.79	0.75	0.74	0.14	0.18	0.22	0.73	0.73	280462
sp33	–	–	–	–	–	–	–	–	–
sp34	0.65	0.64	0.63	0.08	0.05	0.20	0.77	0.48	342629
sp35	0.95	0.88	0.88	0.16	0.48	0.11	0.86	0.89	41431
sp36	–	–	–	–	–	–	–	–	–
sp37	0.78	0.70	0.69	0.30	0.15	0.04	1.00	0.43	680447
sp38	0.72	0.56	0.52	0.74	0.07	0.10	0.86	0.54	228553
sp39	0.65	0.64	0.64	0.05	0.07	0.10	0.40	0.93	107119
sp40	0.75	0.70	0.69	0.21	0.13	0.18	0.89	0.53	308557
mean	0.81	0.72	0.71	0.28	0.21	0.15	0.82	0.71	217749
median	0.81	0.71	0.71	0.27	0.18	0.14	0.86	0.73	173632

Weighted models: forward-backward selection

Species code	AUC_{RS}	AUC_{CV}	AUC_{LOOCV}	GE	D^2	Cut-point	Sensitivity	Specificity	Suitable cells
sp1	0.74	0.65	0.63	0.37	0.14	0.51	0.81	0.56	45546
sp2	0.65	0.61	0.61	0.26	0.08	0.49	0.82	0.52	751681
sp3	0.64	0.61	0.60	0.22	0.05	0.43	0.88	0.37	457182
sp4	–	–	–	–	–	–	–	–	–
sp5	0.70	0.68	0.68	0.10	0.12	0.53	0.79	0.56	694878
sp6	0.77	0.67	0.66	0.39	0.20	0.52	0.89	0.58	387197
sp7	0.77	0.73	0.73	0.14	0.15	0.45	0.82	0.66	83594
sp8	0.81	0.79	0.79	0.05	0.25	0.44	0.91	0.69	351505
sp9	0.71	0.59	0.58	0.54	0.12	0.56	0.44	0.89	5440
sp10	0.90	0.85	0.85	0.13	0.44	0.49	0.89	0.79	407486
sp11	0.69	0.66	0.66	0.15	0.07	0.34	1.00	0.32	372303
sp12	–	–	–	–	–	–	–	–	–
sp13	0.94	0.83	0.82	0.27	0.58	0.51	1.00	0.83	229045
sp14	0.85	0.74	0.75	0.30	0.29	0.46	0.95	0.71	90049
sp15	0.75	0.74	0.74	0.03	0.16	0.37	0.88	0.51	338540
sp16	0.73	0.66	0.66	0.28	0.12	0.35	0.93	0.42	277180
sp17	0.82	0.73	0.73	0.29	0.27	0.41	0.93	0.61	194677
sp18	0.69	0.68	0.68	0.07	0.10	0.44	0.84	0.50	461690
sp19	0.84	0.77	0.77	0.23	0.32	0.51	0.82	0.74	97072
sp20	0.77	0.70	0.70	0.25	0.17	0.47	0.85	0.63	32773
sp21	0.77	0.75	0.75	0.08	0.22	0.47	0.88	0.58	215701
sp22	0.76	0.73	0.72	0.12	0.20	0.51	0.61	0.77	91447
sp23	0.72	0.71	0.71	0.06	0.11	0.48	0.72	0.67	147475
sp24	0.94	0.81	0.81	0.30	0.59	0.48	1.00	0.83	434070
sp25	0.69	0.65	0.64	0.22	0.09	0.51	0.59	0.78	17084
sp26	–	–	–	–	–	–	–	–	–
sp27	0.73	0.58	0.57	0.65	0.16	0.54	0.75	0.63	308626
sp28	0.77	0.71	0.71	0.20	0.20	0.43	0.89	0.55	257171
sp29	–	–	–	–	–	–	–	–	–
sp30	0.78	0.74	0.73	0.15	0.16	0.56	0.71	0.73	324872
sp31	0.77	0.76	0.75	0.04	0.15	0.34	0.93	0.51	388590
sp32	0.75	0.74	0.74	0.03	0.14	0.45	0.85	0.59	557024
sp33	–	–	–	–	–	–	–	–	–
sp34	0.65	0.64	0.64	0.08	0.06	0.45	0.77	0.50	329889
sp35	0.78	0.76	0.76	0.07	0.20	0.63	0.71	0.78	57192
sp36	–	–	–	–	–	–	–	–	–
sp37	0.84	0.74	0.74	0.29	0.33	0.59	1.00	0.66	378559
sp38	–	–	–	–	–	–	–	–	–
sp39	0.67	0.52	0.52	0.91	0.10	0.61	0.40	0.93	113265
sp40	0.71	0.70	0.70	0.04	0.11	0.47	0.80	0.57	316507
mean	0.76	0.70	0.70	0.22	0.20	0.48	0.82	0.64	279252
median	0.76	0.71	0.71	0.20	0.16	0.48	0.85	0.63	308626

Weighted models: backward-forward selection

Species code	AUC_{RS}	AUC_{CV}	AUC_{LOOCV}	GE	D^2	Cut-point	Sensitivity	Specificity	Suitable cells
sp1	0.88	0.70	0.69	0.47	0.37	0.39	1.00	0.68	73568
sp2	0.96	0.65	0.64	0.67	0.60	0.58	1.00	0.88	246162
sp3	0.76	0.55	0.55	0.82	0.17	0.47	0.79	0.63	90106
sp4	0.86	0.51	0.50	0.97	0.32	0.54	0.88	0.79	303555
sp5	0.90	0.52	0.51	0.95	0.39	0.46	0.93	0.75	282147
sp6	0.90	0.58	0.58	0.81	0.39	0.54	0.94	0.78	382911
sp7	0.86	0.72	0.71	0.40	0.33	0.47	0.86	0.74	66309
sp8	0.91	0.56	0.54	0.86	0.50	0.64	0.91	0.87	104118
sp9	0.90	0.54	0.52	0.91	0.42	0.50	1.00	0.78	110231
sp10	0.97	0.79	0.79	0.37	0.64	0.63	0.94	0.93	282084
sp11	0.94	0.77	0.78	0.39	0.52	0.56	0.94	0.87	29023
sp12	0.90	0.70	0.69	0.50	0.43	0.45	1.00	0.72	67646
sp13	0.97	0.61	0.59	0.76	0.68	0.63	1.00	0.91	166910
sp14	0.92	0.78	0.79	0.33	0.48	0.63	0.90	0.87	54375
sp15	0.82	0.70	0.70	0.38	0.26	0.52	0.78	0.71	136965
sp16	0.82	0.63	0.63	0.58	0.26	0.40	0.97	0.54	124153
sp17	0.85	0.72	0.72	0.38	0.36	0.43	1.00	0.66	116149
sp18	0.82	0.70	0.71	0.36	0.26	0.62	0.67	0.83	64097
sp19	0.93	0.74	0.74	0.44	0.56	0.57	1.00	0.85	39553
sp20	0.82	0.70	0.70	0.38	0.23	0.48	0.85	0.71	61985
sp21	0.81	0.74	0.74	0.23	0.27	0.46	0.93	0.64	181472
sp22	0.89	0.78	0.79	0.28	0.40	0.53	0.83	0.84	61078
sp23	0.86	0.70	0.70	0.45	0.32	0.44	0.90	0.70	401593
sp24	0.98	0.77	0.76	0.45	0.79	0.60	1.00	0.94	354581
sp25	0.81	0.66	0.65	0.49	0.23	0.36	0.95	0.51	121408
sp26	0.89	0.65	0.66	0.61	0.41	0.46	1.00	0.75	125538
sp27	0.98	0.69	0.68	0.60	0.70	0.59	1.00	0.92	18822
sp28	0.85	0.70	0.71	0.43	0.32	0.44	0.94	0.68	165059
sp29	0.99	0.87	0.88	0.25	0.81	0.72	1.00	0.97	33751
sp30	0.83	0.70	0.71	0.39	0.26	0.55	0.78	0.78	166891
sp31	0.87	0.75	0.75	0.32	0.36	0.39	1.00	0.64	519086
sp32	0.80	0.73	0.74	0.23	0.23	0.38	0.95	0.57	303805
sp33	0.92	0.50	0.49	0.99	0.48	0.55	1.00	0.83	169248
sp34	0.65	0.64	0.64	0.10	0.06	0.46	0.77	0.50	329887
sp35	0.99	0.90	0.90	0.18	0.84	0.77	1.00	0.97	23081
sp36	0.84	0.55	0.54	0.85	0.27	0.58	0.71	0.82	195811
sp37	0.92	0.64	0.63	0.67	0.49	0.58	1.00	0.81	169939
sp38	0.79	0.50	0.49	0.99	0.21	0.42	1.00	0.55	180849
sp39	0.93	0.52	0.48	0.96	0.49	0.55	1.00	0.84	164666
sp40	0.72	0.70	0.70	0.10	0.12	0.40	0.89	0.45	380214
mean	0.88	0.67	0.67	0.53	0.40	0.52	0.93	0.76	171721
median	0.89	0.70	0.70	0.45	0.38	0.53	0.95	0.78	150816

Weighted multimodels

Species code	AUC_{RS}	AUC_{CV}	AUC_{LOOCV}	GE	D^2	Cut-point	Sensitivity	Specificity	Suitable cells
sp1	0.87	0.70	0.70	0.46	0.28	0.37	1.00	0.62	46384
sp2	0.95	0.68	0.66	0.61	0.40	0.64	0.91	0.93	206072
sp3	0.71	0.59	0.59	0.57	0.09	0.52	0.73	0.62	185036
sp4	–	–	–	–	–	–	–	–	–
sp5	0.75	0.67	0.67	0.32	0.14	0.54	0.79	0.65	647126
sp6	0.86	0.64	0.63	0.62	0.24	0.59	0.78	0.83	302319
sp7	0.83	0.73	0.73	0.31	0.24	0.48	0.82	0.74	61834
sp8	0.84	0.78	0.78	0.19	0.27	0.47	0.91	0.75	292247
sp9	0.87	0.55	0.53	0.87	0.24	0.44	0.94	0.76	145776
sp10	0.95	0.85	0.85	0.22	0.52	0.59	0.89	0.91	337941
sp11	0.93	0.78	0.78	0.36	0.38	0.56	0.89	0.87	26316
sp12	–	–	–	–	–	–	–	–	–
sp13	0.96	0.80	0.80	0.34	0.57	0.55	1.00	0.87	193690
sp14	0.90	0.78	0.79	0.30	0.38	0.54	0.90	0.84	56897
sp15	0.79	0.73	0.73	0.22	0.20	0.40	0.88	0.58	292191
sp16	0.79	0.66	0.66	0.45	0.18	0.47	0.90	0.58	131372
sp17	0.85	0.73	0.73	0.34	0.27	0.41	0.96	0.65	143825
sp18	0.79	0.71	0.71	0.28	0.19	0.48	0.88	0.59	247869
sp19	0.80	0.68	0.68	0.41	0.30	0.47	0.82	0.67	81663
sp20	0.81	0.71	0.71	0.31	0.20	0.47	0.88	0.65	38198
sp21	0.80	0.75	0.75	0.17	0.22	0.52	0.83	0.70	142424
sp22	0.88	0.78	0.78	0.27	0.33	0.46	0.88	0.73	94834
sp23	0.82	0.72	0.72	0.32	0.23	0.36	0.97	0.56	473737
sp24	0.97	0.81	0.80	0.35	0.66	0.56	1.00	0.91	377097
sp25	0.79	0.66	0.66	0.43	0.18	0.41	0.83	0.60	47630
sp26	–	–	–	–	–	–	–	–	–
sp27	0.98	0.72	0.72	0.53	0.55	0.56	1.00	0.91	17233
sp28	0.83	0.72	0.72	0.33	0.24	0.45	0.91	0.66	180995
sp29	–	–	–	–	–	–	–	–	–
sp30	0.83	0.73	0.73	0.32	0.20	0.54	0.83	0.74	211851
sp31	0.84	0.77	0.77	0.20	0.25	0.37	1.00	0.60	496825
sp32	0.78	0.75	0.75	0.13	0.19	0.51	0.76	0.73	271233
sp33	–	–	–	–	–	–	–	–	–
sp34	0.65	0.64	0.64	0.07	0.06	0.46	0.74	0.51	315043
sp35	0.99	0.88	0.89	0.22	0.64	0.59	1.00	0.97	26794
sp36	–	–	–	–	–	–	–	–	–
sp37	0.91	0.69	0.68	0.53	0.34	0.60	1.00	0.81	204731
sp38	–	–	–	–	–	–	–	–	–
sp39	0.92	0.48	0.45	1.06	0.32	0.48	1.00	0.81	202659
sp40	0.71	0.70	0.70	0.08	0.11	0.45	0.85	0.51	315015
mean	0.85	0.71	0.71	0.37	0.29	0.49	0.89	0.72	206511
median	0.84	0.72	0.72	0.32	0.24	0.48	0.89	0.73	193690

Spatial models: forward-backward selection

Species code	AUC_{RS}	AUC_{CV}	AUC_{LOOCV}	GE	D^2	Cut-point	Sensitivity	Specificity	Suitable cells
sp1	0.73	0.64	0.63	0.37	0.15	0.45	0.90	0.51	47155
sp2	0.66	0.60	0.59	0.35	0.08	0.44	0.91	0.44	799290
sp3	0.68	0.61	0.61	0.38	0.09	0.53	0.61	0.65	267164
sp4	–	–	–	–	–	–	–	–	–
sp5	0.71	0.63	0.62	0.41	0.14	0.45	0.86	0.51	704839
sp6	0.76	0.68	0.68	0.32	0.20	0.51	0.89	0.57	402397
sp7	0.80	0.74	0.73	0.21	0.24	0.57	0.79	0.75	64565
sp8	0.82	0.78	0.79	0.12	0.26	0.39	0.91	0.66	361760
sp9	0.81	0.69	0.69	0.39	0.23	0.46	0.81	0.71	602218
sp10	0.92	0.88	0.88	0.09	0.49	0.54	0.89	0.85	382879
sp11	0.68	0.63	0.62	0.30	0.11	0.41	0.94	0.45	198565
sp12	–	–	–	–	–	–	–	–	–
sp13	0.95	0.86	0.87	0.21	0.70	0.70	1.00	0.91	117676
sp14	0.89	0.81	0.82	0.19	0.41	0.65	0.86	0.86	55147
sp15	0.77	0.75	0.75	0.07	0.18	0.66	0.59	0.81	75668
sp16	0.80	0.72	0.72	0.27	0.24	0.56	0.76	0.73	183307
sp17	0.85	0.72	0.73	0.36	0.32	0.53	0.89	0.69	140616
sp18	0.74	0.71	0.71	0.11	0.15	0.41	0.90	0.47	528320
sp19	0.89	0.80	0.80	0.25	0.45	0.42	1.00	0.66	136213
sp20	0.80	0.70	0.70	0.32	0.22	0.49	0.83	0.64	27964
sp21	0.78	0.76	0.75	0.10	0.22	0.54	0.78	0.70	148039
sp22	0.83	0.79	0.79	0.12	0.29	0.55	0.68	0.83	317731
sp23	0.73	0.70	0.70	0.13	0.12	0.46	0.72	0.64	171792
sp24	0.98	0.87	0.85	0.24	0.82	0.81	1.00	0.95	164066
sp25	0.70	0.64	0.64	0.31	0.10	0.51	0.63	0.75	14221
sp26	–	–	–	–	–	–	–	–	–
sp27	0.75	0.64	0.63	0.45	0.19	0.51	0.83	0.60	444505
sp28	0.81	0.72	0.71	0.31	0.27	0.52	0.86	0.65	204149
sp29	–	–	–	–	–	–	–	–	–
sp30	0.79	0.74	0.74	0.18	0.22	0.56	0.68	0.78	292962
sp31	0.77	0.74	0.74	0.09	0.18	0.42	0.83	0.60	260912
sp32	0.75	0.74	0.73	0.05	0.15	0.43	0.83	0.61	543616
sp33	–	–	–	–	–	–	–	–	–
sp34	0.65	0.63	0.63	0.16	0.07	0.48	0.79	0.49	320161
sp35	0.80	0.76	0.76	0.13	0.21	0.42	0.86	0.63	212122
sp36	–	–	–	–	–	–	–	–	–
sp37	0.85	0.81	0.81	0.12	0.39	0.53	0.93	0.73	361628
sp38	–	–	–	–	–	–	–	–	–
sp39	0.74	0.35	0.31	1.65	0.16	0.41	0.90	0.54	420617
sp40	0.71	0.69	0.68	0.11	0.12	0.39	0.89	0.48	481060
mean	0.78	0.71	0.71	0.27	0.25	0.51	0.83	0.66	286464
median	0.78	0.72	0.72	0.21	0.21	0.51	0.86	0.65	260912

Spatial models: backward-forward selection

Species code	AUC_{RS}	AUC_{CV}	AUC_{LOOCV}	GE	D^2	Cut-point	Sensitivity	Specificity	Suitable cells
sp1	0.94	0.82	0.81	0.28	0.62	0.55	1.00	0.82	38252
sp2	1.00	0.60	0.62	0.80	1.00	0.01	1.00	1.00	91235
sp3	0.77	0.59	0.59	0.66	0.20	0.41	0.91	0.57	172857
sp4	0.89	0.54	0.54	0.90	0.39	0.56	0.88	0.79	414473
sp5	0.91	0.53	0.52	0.93	0.44	0.50	0.93	0.82	240624
sp6	0.94	0.71	0.72	0.52	0.57	0.56	0.94	0.84	346245
sp7	0.88	0.75	0.75	0.35	0.38	0.39	0.93	0.72	71253
sp8	0.97	0.64	0.63	0.71	0.76	0.51	1.00	0.96	55450
sp9	0.93	0.58	0.57	0.82	0.51	0.53	1.00	0.80	128832
sp10	0.98	0.85	0.85	0.27	0.76	0.48	1.00	0.91	295523
sp11	0.95	0.79	0.79	0.36	0.61	0.57	0.94	0.91	26828
sp12	0.99	0.71	0.71	0.56	0.78	0.66	1.00	0.96	32757
sp13	1.00	0.64	0.64	0.72	1.00	0.01	1.00	1.00	125498
sp14	0.94	0.81	0.81	0.28	0.59	0.38	1.00	0.81	134333
sp15	0.84	0.74	0.74	0.30	0.30	0.52	0.83	0.75	141835
sp16	0.90	0.74	0.74	0.41	0.43	0.48	0.97	0.74	124958
sp17	0.86	0.74	0.75	0.33	0.38	0.43	0.96	0.70	100539
sp18	0.89	0.72	0.73	0.43	0.40	0.52	0.84	0.79	155040
sp19	0.92	0.71	0.72	0.50	0.59	0.45	1.00	0.84	48672
sp20	0.83	0.72	0.72	0.33	0.26	0.41	0.92	0.64	89729
sp21	0.83	0.76	0.76	0.21	0.30	0.44	0.93	0.63	181975
sp22	0.90	0.79	0.80	0.27	0.43	0.53	0.88	0.83	65657
sp23	0.90	0.71	0.71	0.46	0.43	0.52	0.86	0.78	376909
sp24	1.00	0.73	0.73	0.55	1.00	0.01	1.00	1.00	161628
sp25	0.85	0.67	0.67	0.51	0.29	0.40	0.93	0.66	116575
sp26	0.91	0.70	0.70	0.51	0.48	0.50	0.96	0.80	111526
sp27	0.99	0.73	0.74	0.52	0.82	0.77	1.00	0.96	12280
sp28	0.91	0.76	0.76	0.36	0.48	0.49	0.97	0.72	173739
sp29	1.00	0.83	0.85	0.33	0.98	0.76	1.00	1.00	31599
sp30	0.86	0.76	0.77	0.28	0.34	0.49	0.85	0.77	161457
sp31	0.87	0.74	0.74	0.35	0.40	0.46	0.97	0.69	456731
sp32	0.81	0.73	0.74	0.24	0.24	0.35	0.95	0.56	273061
sp33	0.95	0.60	0.58	0.79	0.62	0.69	1.00	0.91	177411
sp34	0.65	0.63	0.62	0.18	0.07	0.49	0.77	0.49	329138
sp35	1.00	0.81	0.81	0.38	1.00	0.01	1.00	1.00	95358
sp36	0.91	0.70	0.69	0.53	0.47	0.41	1.00	0.76	212241
sp37	0.93	0.75	0.75	0.40	0.59	0.57	1.00	0.85	132383
sp38	0.81	0.52	0.52	0.95	0.27	0.50	0.90	0.64	111054
sp39	0.98	0.68	0.66	0.64	0.76	0.66	1.00	0.94	100337
sp40	0.72	0.69	0.69	0.14	0.12	0.38	0.96	0.45	347581
mean	0.90	0.71	0.71	0.48	0.53	0.46	0.95	0.80	161589
median	0.91	0.72	0.72	0.42	0.47	0.49	0.96	0.80	130608

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium of Information Theory* (Ed, Csaki, F. and Petrov, B. N.) Akademiai Kiado, Budapest, Hungary.

Akaike, H. (1978) A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, **30**, 9-14.

Anderson, D. R., Burnham, K. P. and Thompson, W. L. (2000) Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management*, **64**, 912-923.

Anderson, R. P. (2003) Real vs. artefactual absences in species distributions: tests for Oryzomys albigularis (Rodentia : Muridae) in Venezuela. *Journal of Biogeography*, **30**, 591-605.

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

Araújo, M. B., Cabeza, M., Thuiller, W., Hannah, L. and 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.

Augustin, N. H., Mugglestone, M. A. and Buckland, S. T. (1996) An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, **33**, 339-347.

Austin, M. (2007) Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, **200**, 1-19.

Austin, M. P. (1999) The potential contribution of vegetation ecology to biodiversity research. *Ecography*, **22**, 465-484.

Austin, M. P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological Modelling*, **157**, 101-118.

Bakkenes, M., Alkemade, J. R. M., Ihle, F., Leemans, R. and Latour, J. B. (2002) Assessing effects of forecasted climate change on the diversity and distribution of European higher plants for 2050. *Global Change Biology*, **8**, 390-407.

Besag, J. (1974) Spatial interaction and statistical-analysis of lattice systems. *Journal of the Royal Statistical Society Series B-Methodological*, **36**, 192-236.

Brooks, T. M., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., Rylands, A. B., Konstant, W. R., Flick, P., Pilgrim, J., Oldfield, S., Magin, G. and Hilton-Taylor, C. (2002) Habitat loss and extinction in the hotspots of biodiversity. *Conservation Biology*, **16**, 909-923.

Brunsdon, C., Fotheringham, S. and Charlton, M. (2007) Geographically weighted discriminant analysis. *Geographical Analysis*, **39**, 376-396.

Burgess, N. D., Balmford, A., Cordeiro, N. J., Fjeldså, J., Kuper, W., Rahbek, C., Sanderson, E. W., Scharlemann, J. P. W., Sommer, J. H. and Williams, P. H. (2007a) Correlations among species distributions, human density and human infrastructure across the high biodiversity tropical mountains of Africa. *Biological Conservation*, **134**, 164-177.

Burgess, N. D., Butynski, T. M., Cordeiro, N. J., Doggart, N. H., Fjeldså, J., Howell, K. M., Kilahama, F. B., Loader, S. P., Lovett, J. C., Mbilinyi, B., Menegon, M., Moyer, D. C., Nashanda, E., Perkin, A., Rovero, F., Stanley, W. T. and Stuart, S. N. (2007b) The biological importance of the Eastern Arc Mountains of Tanzania and Kenya. *Biological Conservation*, **134**, 209-231.

Cohen, J., Cohen, P., West, S. G. and Aiken, L. S. (2003) *Applied Multiple Regression/Correlation Analysis For The Behavioural Sciences*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA.

Dormann, C. F. (2007a) Assessing the validity of autologistic regression. *Ecological Modelling*, **207**, 234-242.

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

Dormann, C. F. (2007c) Promising the future? Global change projections of species distributions. *Basic and Applied Ecology*, **8**, 387-397.

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kuhn, I., Ohlemuller, R., Peres-Neto, P. R., Reineking, B., Schroder, B., Schurr, F. M. and Wilson, R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609-628.

Dormann, C. F., Schweiger, O., Arens, P., Augenstein, I., Aviron, S., Bailey, D., Baudry, J., Billeter, R., Bugter, R., Bukacek, R., Burel, F., Cerny, M., De Cock, R., De Blust, G., DeFilippi, R., Diekotter, T., Dirksen, J., Durka, W., Edwards, P. J., Frenzel, M., Hamersky, R., Hendrickx, F., Herzog, F., Klotz, S., Koolstra, B., Lausch, A., Le Coeur, D., Liira, J., Maelfait, J. P., Opdam, P., Roubalova, M., Schermann-Legionnet, A., Schermann, N., Schmidt, T., Smulders, M. J. M., Speelmans, M., Simova, P., Verboom, J., van Wingerden, W. and Zobel, M. (2008) Prediction uncertainty of environmental change effects on temperate European biodiversity. *Ecology Letters*, **11**, 235-244.

Eberhardt, L. L. (2003) What should we do about hypothesis testing? *Journal of Wildlife Management*, **67**, 241-247.

Engler, R., Guisan, A. and 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., Drielsma, M., Manion, G. and Watson, G. (2002a) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. II. Community-level modelling. *Biodiversity and Conservation*, **11**, 2309-2338.

Ferrier, S., Watson, G., Pearce, J. and Drielsma, M. (2002b) Extended statistical approaches to modelling spatial pattern in biodiversity in northeast New South Wales. I. Species-level modelling. *Biodiversity and Conservation*, **11**, 2275-2307.

Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2002) *Geographical Weighted Regression: The Analysis of Spatially Relationships*, Wiley, Chichester, UK.

Graham, M. H. (2003) Confronting multicollinearity in ecological multiple regression. *Ecology*, **84**, 2809-2815.

GRASS-Development-Team (2009) Geographic Resources Analysis Support System (GRASS), GNU General Public License (http://grass.osgeo.org/).

Green, D. M. and Swets, J. A. (1974) *Signal Detection Theory and Psychophysics*, Krieger, New York, USA.

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

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

Hartley, S., Harris, R. and Lester, P. J. (2006) Quantifying uncertainty in the potential distribution of an invasive species: climate and the Argentine ant. *Ecology Letters*, **9**, 1068-1079.

Hastie, T. J., Friedman, J. H. and Tibshirani, R. J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York, USA.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*, Chapman and Hall, London, UK.

Hemp, A. (2006) Vegetation of Kilimanjaro: hidden endemics and missing bamboo. *African Journal of Ecology*, **44**, 305-328.

Horn, B. K. P. (1981) Hill shading and reflectance map. *Proceedings of the IEEE*, **69**, 14-47.

Johnson, C. J., Nielsen, S. E., Merrill, E. H., McDonald, T. L. and Boyce, M. S. (2006) Resource selection functions based on use-availability data: Theoretical motivation and evaluation methods. *Journal of Wildlife Management*, **70**, 347-357.

Johnson, J. B. and Omland, K. S. (2004) Model selection in ecology and evolution. *Trends in Ecology & Evolution*, **19**, 101-108.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*.

Küper, W., Sommer, J. H., Lovett, J. C. and Barthlott, W. (2006) Deficiency in African plant distribution data - missing pieces of the puzzle. *Botanical Journal of the Linnean Society*, **150**, 355-368.

Leathwick, J. R. and Austin, M. P. (2001) Competitive interactions between tree species in New Zealand's old-growth indigenous forests. *Ecology*, **82**, 2560-2573.

Lichstein, J. W., Simons, T. R., Shriner, S. A. and Franzreb, K. E. (2002) Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs*, **72**, 445-463.

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

Lobo, J. M., Jimenez-Valverde, A. and Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145-151.

Lovett, J. C. (1985) Moist forsts of eastern Tanzania. *Swara*, **8**, 8-9.

Lovett, J. C. (1990) Classification and status of the moist forests of Tanzania. *Notifications from the Institute for General Botany Hamburg*, **23a**, 287–300.

Lovett, J. C. (1996) Elevational and latitudinal changes in tree associations and diversity in the Eastern Arc mountains of Tanzania. *Journal of Tropical Ecology*, **12**, 629-650.

Lovett, J. C. (1998) Importance of the Eastern Arc Mountains for vascular plants. *Journal of East African Natural History*, **87**, 37-58.

Lovett, J. C., Clarke, G. P., Moore, R. and Morrey, G. H. (2001) Elevational distribution of restricted range forest tree taxa in eastern Tanzania. *Biodiversity and Conservation*, **10**, 541-550.

Lovett, J. C., Ruffo, C. K., Gereau, R. E. and Taplin, J. R. D. (2006) *Field guide to the moist forest trees of Tanzania*, The Society for Environmental Exploration, UK and the University of Dar es Salaam, The United Republic of Tanzania.

Maggini, R., Lehmann, A., Zimmermann, N. E. and Guisan, A. (2006) Improving generalized regression analysis for the spatial prediction of forest communities. *Journal of Biogeography*, **33**, 1729-1749.

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

Marchant, R., Mumbi, C., Behera, S. and Yamagata, T. (2007) The Indian Ocean dipole - the unsung driver of climatic variability in East Africa. *African Journal of Ecology*, **45**, 4-16.

McClean, C. J., Lovett, J. C., Kuper, W., Hannah, L., Sommer, J. H., Barthlott, W., Termansen, M., Smith, G. E., Tokamine, S. and Taplin, J. R. D. (2005) African plant diversity and climate change. *Annals of the Missouri Botanical Garden*, **92**, 139-152.

McPherson, J. M., Jetz, W. and Rogers, D. J. (2004) The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, **41**, 811-823.

Mickey, R. M. and Greenland, S. (1989) The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology*, **129**, 125-137.

Miller, J., Franklin, J. and Aspinall, R. (2007) Incorporating spatial dependence in predictive vegetation models. *Ecological Modelling*, **202**, 225-242.

Mittermeier, R. A., Robles Gil., P., Hoffmann, M., Pilgrim, J., Brooks, T., Mittermeier, C. G., Lamoreux, J. and da Fonseca, G. A. B. (2004) *Hotspots Revisited*, CEMEX, Mexico.

Mumbi, C. T., Marchant, R., Hooghiemstra, H. and Wooller, M. J. (2008) Late Quaternary vegetation reconstruction from the Eastern Arc Mountains, Tanzania. *Quaternary Research*, **69**, 326-341.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. and Kent, J. (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853-858.

Olson, D. M. and Dinerstein, E. (1998) The global 200: A representation approach to conserving the Earth's most biologically valuable ecoregions. *Conservation Biology*, **12**, 502-515.

Parker, B. J., Guenter, S. and Bedo, J. (2007) Stratification bias in low signal microarray studies. *BMC Bioinformatics*, **8**, 16.

Pearce, J. and Ferrier, S. (2000a) An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, **128**, 127-147.

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

Pearce, J. L., Cherry, K., Drielsma, M., Ferrier, S. and Whish, G. (2001) Incorporating expert opinion and fine-scale vegetation mapping into statistical models of faunal distribution. *Journal of Applied Ecology*, **38**, 412-424.

Pearson, R. G. and Dawson, T. P. (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, **12**, 361-371.

Pearson, R. G., Thuiller, W., Araujo, M. B., Martinez-Meyer, E., Brotons, L., McClean, C., Miles, L., Segurado, P., Dawson, T. P. and Lees, D. C. (2006) Model-based uncertainty in species range prediction. *Journal of Biogeography*, **33**, 1704-1711.

Prentice, I. C., Cramer, W., Harrison, S. P., Leemans, R., Monserud, R. A. and Solomon, A. M. (1992) A global biome model based on plant physiology and dominance, soil properties and climate. *Journal of Biogeography*, **19**, 117-134.

Pulliam, H. R. (2000) On the relationship between niche and distribution. *Ecology Letters*, **3**, 349-361.

R-Development-Core-Team (2009) *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria (http://www.R-project.org/).

Richards, C. L., Carstens, B. C. and Knowles, L. L. (2007) Distribution modelling and statistical phylogeography: an integrative framework for generating and testing alternative biogeographical hypotheses. *Journal of Biogeography*, **34**, 1833-1845.

Roberts, D. W. and Cooper, S. V. (1989) Concepts and techniques of vegetation mapping. In *Land Classifications Based on Vegetation: Applications for Resource Management* (Eds, Ferguson, D., Morgan, P. and Johnson, F. D.) USDA For. Serv. Gen. Tech. Rep. INT-257, Ogden, Utah, USA, pp. 90-96.

Rushton, S. P., Ormerod, S. J. and Kerby, G. (2004) New paradigms for modelling species distributions? *Journal of Applied Ecology*, **41**, 193-200.

Schlüter, T. (1997) Geology of East Africa. In *Contributions to the Regional Geography of the Earth* (translated from German), Borntraeger, Berlin, Germany.

Segurado, P. and Araujo, M. B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555-1568.

Segurado, P. and Araújo, M. B. (2004) An evaluation of methods for modelling species distributions. *Journal of Biogeography*, **31**, 1555-1568.

Seoane, J., Bustamante, J. and Diaz-Delgado, R. (2005) Effect of expert opinion on the predictive ability of environmental models of bird distribution. *Conservation Biology*, **19**, 512-522.

Stattersfield, A. J., Crosby, M. J., Long, A. J. and Wege., D. C. (1998) *Endemic Bird Areas of the World: Priorities for Biodiversity Conservation*, BirdLife International, Cambridge, UK.

Stone, M. (1974) Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **36**, 111-147.

Swets, J. A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285-1293.

Taplin, J. R. D. and Lovett, J. C. (2003) Can we predict centres of plant species richness and rarity from environmental variables in sub-Saharan Africa? *Botanical Journal of the Linnean Society*, **142**, 187-197.

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., de Siqueira, M. F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Peterson, A. T., Phillips, O. L. and Williams, S. E. (2004) Extinction risk from climate change. *Nature*, **427**, 145-148.

Thornthwaite, C. W. (1948) An Approach toward a Rational Classification of Climate. *Geographical Review*, **38**, 55-94.

Thuiller, W., Brotons, L., Araújo, M. B. and Lavorel, S. (2004) Effects of restricting environmental range of data to project current and future species distributions. *Ecography*, **27**, 165-172.

van Teeffelen, A. J. A. and Ovaskainen, O. (2007) Can the cause of aggregation be inferred from species distributions? *Oikos*, **116**, 4-16.

Whittingham, M. J., Stephens, P. A., Bradbury, R. B. and Freckleton, R. P. (2006) Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, **75**, 1182-1189.

Yee, T. W. and Mitchell, N. D. (1991) Generalized additive models in plant ecology. *Journal of Vegetation Science*, **2**, 587-602.