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

Philip J. Platts\textsuperscript{a,}\*, Colin J. McClean\textsuperscript{b}, Jon C. Lovett\textsuperscript{b}, Rob Marchant\textsuperscript{a}

\textsuperscript{a} The York Institute for Tropical Ecosystem Dynamics, Environment Department, University of York, Heslington, York YO10 5DD, UK
\textsuperscript{b} Centre for Ecology Law and Policy, Environment Department, University of York, Heslington, York YO10 5DD, UK

ARTICLE INFO

Article history:
Received 30 January 2008
Received in revised form 6 June 2008
Accepted 13 June 2008
Published on line 28 July 2008

Keywords:
Eastern Arc Mountains
Tropical trees
Generalised additive models
Stepwise selection
Model averaging
Prevalence
Spatial autocorrelation
Extrapolation uncertainty

ABSTRACT

The Eastern Arc Mountains (EAMs) of Tanzania and Kenya support some of the most ancient tropical rainforest 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 species, using generalised additive models, plot data and environmental predictor maps at sub 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 an autocovariate term 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 species 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, while 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 (parsimony). 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. Full results and details of all species models are supplied in an online Appendix.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Research into the habitat requirements of species plays a fundamental role in planning for their future conservation, particularly if their persistence is threatened by external pressures such as disturbance and climatic change. 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., 2002a), improved sampling of rare and endangered species (Engler et al., 2004; Guisan et al., 2005), determination of reserve boundaries (Ferrier et al., 2002b; 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 (Stattersfield et al., 1998; Myers et al., 2000; Olson and Dinerstein, 2002), 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 40 large tree species 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 response shapes (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., 2007a), we also calibrate and compare three different methods for model selection: two best-model stepwise procedures and one multimodel.

1.1. 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; Akaike, 1973, 1974), 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 Akaile weights. Although not strictly adhering to the IT philosophy of multimodel inference, many studies now adopt the use of AIC in stepwise procedures.

1.2. 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; e.g., 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 fully explain 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 fully describe the observed aggregation (Guisan and Thuiller, 2005), missing pieces of the puzzle include dispersal patterns, competition/mutualism and disturbance; second, ecologists are inclined to choose sample sites in more accessible locations and areas of particular interest, yielding a non-random distribution of sites that can confound SAC in models. 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, 2007a). 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 Araújo, 2004); however recent studies warn that autologistic (autocovariate) models may underestimate the environmental controls on a species distribution (Dormann, 2007b; Dormann et al., 2007b).

### 1.3. Envelope uncertainty

The breadth of niche-space spanned by the distribution data may not be sufficient to fully represent the study region, particularly if we consider projecting our 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 (Araújo and Guisan, 2006). The solution we present is that model predictions be accompanied by envelope uncertainty maps (EUMs), allowing the analyst to identify geographical locations where the profile of environmental conditions at sample sites has resulted in high model uncertainty.

### 2. Methods

#### 2.1. Study region

The EAMs are part of the Eastern Afromontane biodiversity hotspot (Mittermeier et al., 2004) and are defined as

---

![Map showing the 13 crystalline blocs of the Eastern Arc Mountains. Encircled dots locate the 201 modelling points. Note the biased distribution of sample sites, a classic problem in environmental datasets. Region for extrapolation is the full map extent: 32.5°E–40.5°E, 1.5°S–10.5°S. Background map is a 30 arc second (920 m) cosine transformation of slope aspect (trasp).](image-url)

---
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. 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 ca. 3300 km² of closed-canopy tropical rainforest, though it has been estimated that this may be less than 30% of the original forest cover (Burgess et al., 2007a). Much of the remaining area is protected by community-based management schemes, national parks and forest reserves, some 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., 2007b), rendering EAM flora particularly sensitive to further fragmentation. Species richness scores are high and the concentrations of endemism are exceptional (see Burgess et al., 2007a), though many hundreds of endemic plants and animals are threatened by extinction. At least 800 vascular plant species are endemic, of which around 80 are trees (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., 2007).

2.2. 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 second (920 m) grid cells across which the samples were collected (Fig. 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 must 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 40 tree species modelled including their sample prevalence is presented in the online Appendix. For further details of the field data we refer the reader to Lovett (1998).

2.3. 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 second 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

---

Table 1 – Summary of environmental predictor variables

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description</th>
<th>Units</th>
<th>Sites Mean</th>
<th>All cells Mean</th>
<th>Sites Range</th>
<th>All cells Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>gradient*</td>
<td>Angle from horizontal</td>
<td>°</td>
<td>7.242</td>
<td>1.533</td>
<td>28.88</td>
<td>64.58</td>
</tr>
<tr>
<td>trasp*</td>
<td>Wetness/radiation index</td>
<td></td>
<td>0.6720</td>
<td>0.5366</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>transe*</td>
<td>Annual temperature range</td>
<td>°C</td>
<td>15.85</td>
<td>16.07</td>
<td>8.672</td>
<td>13.10</td>
</tr>
<tr>
<td>pptdry*</td>
<td>Precipitation driest month</td>
<td>mm</td>
<td>11.95</td>
<td>4.949</td>
<td>54.11</td>
<td>94.80</td>
</tr>
<tr>
<td>pptann¹</td>
<td>Mean annual precipitation</td>
<td>cm</td>
<td>107.4</td>
<td>87.14</td>
<td>121.1</td>
<td>194.1</td>
</tr>
<tr>
<td>pptwet¹</td>
<td>Precipitation wettest month</td>
<td>mm</td>
<td>229.1</td>
<td>186.4</td>
<td>389.3</td>
<td>437.0</td>
</tr>
<tr>
<td>tmean²</td>
<td>Mean temperature</td>
<td>°C</td>
<td>21.20</td>
<td>22.71</td>
<td>14.01</td>
<td>36.44</td>
</tr>
<tr>
<td>tmeanw²</td>
<td>Mean temp. warmest month</td>
<td>°C</td>
<td>23.23</td>
<td>24.69</td>
<td>14.03</td>
<td>36.45</td>
</tr>
<tr>
<td>tmeanc²</td>
<td>Mean temp. coldest month</td>
<td>°C</td>
<td>18.44</td>
<td>20.23</td>
<td>14.75</td>
<td>36.23</td>
</tr>
<tr>
<td>tmmax²</td>
<td>Max. temp. warmest month</td>
<td>°C</td>
<td>28.83</td>
<td>30.46</td>
<td>13.58</td>
<td>38.50</td>
</tr>
<tr>
<td>tmmin²</td>
<td>Min. temp. coldest month</td>
<td>°C</td>
<td>13.15</td>
<td>14.52</td>
<td>17.56</td>
<td>33.75</td>
</tr>
<tr>
<td>tabsmin²</td>
<td>Absolute minimum temp.</td>
<td>°C</td>
<td>4.486</td>
<td>7.228</td>
<td>22.45</td>
<td>51.94</td>
</tr>
<tr>
<td>tdays⁵</td>
<td>Temperature-days</td>
<td>–days</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>pet²</td>
<td>Potential evapotranspiration</td>
<td>cm/year</td>
<td>105.4</td>
<td>115.1</td>
<td>99.77</td>
<td>223.9</td>
</tr>
<tr>
<td>ami²</td>
<td>Annual moisture index</td>
<td>cm/year</td>
<td>1.093</td>
<td>0.7918</td>
<td>2.041</td>
<td>4.997</td>
</tr>
</tbody>
</table>

¹, independent [abs(r) < 0.7]; ², first correlation group; ², second correlation group; (), species-specific. Climatic range is high because of Mt. Kilimanjaro (5895 m AMSL).
the climatic gradients directly affecting species distributions (Table 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 tmeanw (refer to Table 1 for abbreviations).

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 1; Fig. 1). The digital elevation model was supplied at a resolution of three arc seconds (92 m); derived predictors were rescaled to 30 arc seconds 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.

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. 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

\[ \text{EUM}_i = \sum_{X \in S} C_X D_{X_i}, \]

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

\[ D_{X_i} = \frac{\max(X_{\min} - X_i, X_i - X_{\max}, 0)}{X_{\max} - X_{\min}}, \]

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.

2.4. 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 10 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; online Appendix).

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.

\[ \text{GE} = \frac{\text{AUC}_{\text{RS}} - \text{AUC}_{\text{CV}}}{\text{AUC}_{\text{RS}} - 0.5}. \]

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

Statistical calculations were performed in R 2.3.1 (http://r-project.org/) using functions from the ‘gam’ and ‘ROC’ packages, together with custom R code. For spatial analyses we used GRASS GIS 6.0 (http://grass.itc.it/). 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/).
Fig. 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).

2.5. 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, d.f.) separately for each species-predictor pair, such that d.f. ∈ [1,4] at intervals of 0.25. Where d.f. = 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; details of the experiments are as follows.

2.5.1. Selecting predictors
All predictor pairs were tested for collinearity using Pearson’s correlation coefficient (r). If two predictors were highly collinear [abs(r) > 0.7] then the one that yielded the highest univariate AUCCV 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 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 AUCCV. 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 AUCCV. 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 multi-model solution (denoted ‘MM’) was achieved by weighting the two best-model predictions according to their respective above chance AUCCV values (AUCCV − 0.5), and taking the average. 2.5.2. 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 PA/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, while an absence record could be a consequence of the restricted sampling regime, or of ecological factors beyond the scope of the model (Anderson, 2003).

2.5.3. 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):

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

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.

3. Results
3.1. 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 AUCCV > 0.5 should provide some discriminatory power. Following this classification for each of the 40 species, 27 forward–backward (FB) models and 36 backward–forward (BF) models produced reasonable, good or excellent predictions. For two species, Syzygium cordatum and Tabernamontana 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
Table 2 – Significance of differences between modelling experiments

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>GE</th>
<th>$D^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline selection</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB</td>
<td>***</td>
<td>**</td>
<td>***</td>
</tr>
<tr>
<td>BF</td>
<td>ns</td>
<td>*</td>
<td>ns</td>
</tr>
<tr>
<td>Forward–backward models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>ns</td>
<td></td>
<td>ns</td>
</tr>
<tr>
<td>S</td>
<td>ns</td>
<td></td>
<td>ns</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forward–backward models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>***</td>
<td>**</td>
<td>***</td>
</tr>
<tr>
<td>W</td>
<td>ns</td>
<td>*</td>
<td>ns</td>
</tr>
<tr>
<td>S</td>
<td>ns</td>
<td></td>
<td>ns</td>
</tr>
<tr>
<td>Multimodels</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>W</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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. 4). As for FB models the error distribution was also much improved, with a 46% reduction in the propor-

3.2. 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 2; Figs. 4 and 5). For FB selection, predictor sets chosen during weighted and baseline model calibration differed for 27 of the 40 species, 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. 5).

Under FB selection the impact on the AUC was not significant, though other statistics revealed important differences:
tion 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.

3.3. Spatial models (S)

The inclusion of a spatial autocovariate increased the proportion of explained deviance in all cases. Spatial models were significantly better at correctly predicting presences and absences (Table 3), and for BF selection they were also more stable (Fig. 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 (see online Appendix). 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. 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. 6b). In this example the non-spatial model is more useful for inference since it retains a realistic gradient of suitability.

3.4. 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 tmean contributed the least to $D^2$ and was the least frequently selected, often dropped in favour of similar measures such as tminc and tbsmin. 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).

The environmental range of survey sites was generally good, with the exception that climatic predictors lacked coverage near the summit of Mt. Kilimanjaro (Table 1; Fig. 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. 2).

4. 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 can be 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.

While 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 relationships. 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 AUCCV). 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., 2007a), 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. 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, 2007b; Dormann et al., 2007b). 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 (see online Appendix). 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. 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 predictions 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 gaps in 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, i.e., the Maasai Steppe. 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.

5. 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 has involved the parameterisation of environmental response functions for 40 species 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 a thorough analysis of response shapes will add to our understanding of their habitat preferences, and specifically the degree to which environmental controls restrict their distributions. Those interested in the particular tree species modelled can find full model details, including regression parameters, in the online Appendix.

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.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2008.06.028.

References


Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ecolmodel.2008.06.028.


