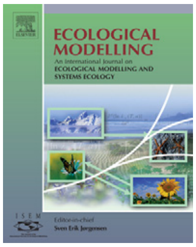
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



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, 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*

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.

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

[*] *Corresponding author*. Tel.: +44 1904 434 780; fax: +44 1904 432 998.
E-mail address: pp509@york.ac.uk (P.J. Platts).

0304-3800/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2008.06.028

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 Akaike 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