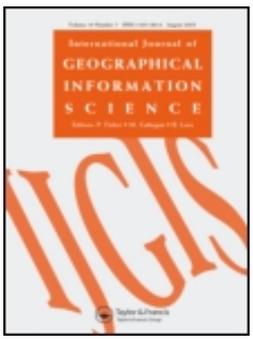

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Pseudo-absences, pseudo-models and pseudo-niches: pitfalls of model selection based on the area under the curve

Duncan Golicher , Andrew Ford , Luis Cayuela & Adrian Newton

To cite this article: Duncan Golicher , Andrew Ford , Luis Cayuela & Adrian Newton (2012) Pseudo-absences, pseudo-models and pseudo-niches: pitfalls of model selection based on the area under the curve, International Journal of Geographical Information Science, 26:11, 2049-2063, DOI: 10.1080/13658816.2012.719626

To link to this article: https://doi.org/10.1080/13658816.2012.719626

International Journal of Geographical Information Science
Vol. 26, No. 11, November 2012, 2049–2063

Taylor & Francis
Taylor & Francis Group

Pseudo-absences, pseudo-models and pseudo-niches: pitfalls of model selection based on the area under the curve

Duncan Golicher[a,b]*, Andrew Ford[b], Luis Cayuela[c] and Adrian Newton[b]

[a]*El Colegio de la Frontera Sur, Conservación de la Biodiversidad, San Cristóbal de las Casas, Mexico;* [b]*Applied Sciences, Bournemouth University, Bournemouth, UK;* [c]*Área de Biodiversidad y Conservación, Universidad Rey Juan Carlos, Madrid, Spain*

(*Received 29 November 2011; final version received 30 July 2012*)

The area under the curve (AUC) of the receiver operator characteristic (ROC) graph is regarded as an objective measure of the discrimination accuracy of predictive models. AUC scores calculated from background values, or pseudo-absences, have been proposed as a method of model selection for species distribution models (SDMs) fitted to presence-only data. However, the utility of AUC as a measure of model performance when data on confirmed absence are unavailable has not been fully investigated. We fitted SDMs using informative climatic variables for 2000 species of Mesoamerican trees. As a reference, we also built 'pseudo-models' using Gaussian random fields with no biological meaning. AUC correctly selected SDMs fitted to single environmental variables over 'pseudo-models' fitted to single random fields in almost all cases. However, when all seven variables were included in the models, AUC erroneously selected complex pseudo-models over complex climate models in 17% of the cases. The spatial distribution patterns predicted by the pseudo-models differed from the results derived from climate-based models, even when overall AUC scores were similar. Both model and pseudo-model AUC values increased when presence points were few and spatially aggregated. The results show that AUC calculated from presence-only data can be an unreliable guide for model selection. Pseudo-absences have ill-defined properties that challenge the interpretation of AUC values. Inference on multidimensional niche spaces should not be supported by AUC values calculated using pseudo-absences.

Keywords: species distribution modelling; model selection

Introduction

In recent years, predictive species distribution models (SDM) have become important tools for biogeography, conservation and research into the impact of climate change on biodiversity (Peterson and Kluza 2003, Araújo *et al*. 2004, Engler *et al*. 2004, Thuiller *et al*. 2005, Broenniman *et al*. 2006, Garcìa 2006, Stockwell 2006, Araújo and Rahbek 2007, Golicher *et al*. 2008). Where detailed knowledge of distribution patterns is available, as is the case for highly visible organisms in some temperate countries, the predictive ability of SDMs can be evaluated against known presences and absences. However, many organisms are cryptic. Furthermore, the biodiversity of a large part of the world's surface, particularly in the tropics, has not been systematically surveyed. The true spatial distribution of

*Corresponding author. Email: dgolicher@bournemouth.ac.uk

ISSN 1365-8816 print/ISSN 1362-3087 online

http://dx.doi.org/10.1080/13658816.2012.719626
http://www.tandfonline.com

most organisms remains unknown (Whittaker *et al*. 2005, Kuper *et al*. 2006). Data sets that provide confirmed evidence of absence are rarely available. Thus, species distributions are commonly inferred from the presence-only data that are provided by museum collections or herbaria (Graham *et al*. 2004). Despite this limited knowledge, reliable maps of realized and potential species distributions are required in order to inform conservation priority setting (Peterson and Kluza 2003, Moilanen *et al*. 2006) and evaluate extinction risk (Newton and Oldfield 2008). This has led to the development and application of computer software designed specifically for the purpose of producing spatially explicit predictions based on models of the relationship between species and environment fit using presence-only data (Stockwell and Noble 1992, Elith *et al*. 2006, Phillips *et al*. 2006).

SDMs are explicitly based on the niche concept (Peterson and Vieglais 2001). Early models used relatively simple rules based on climatic envelopes or elevational preference in order to suggest the species geographical range (Berry *et al*. 2002). Such models matched expert knowledge concerning the factors that limited species distributions. Choice of climatic envelope models is largely based on ecological understanding, rather than derived directly from the input data. In contrast, contemporary SDMs are data driven, relying heavily on modern machine learning techniques (Elith *et al*. 2006, Phillips *et al*. 2006, Elith and Leathwick 2007). Algorithms fit flexible curves and complex rules in order to find a model which maps closely onto the best available data. Models differing in the identity and number of input variables produce different spatial predictions. The process of producing a map that shows a species realised or potential distribution involves selecting appropriate input variables, choosing between a set of algorithms and finding threshold values in order to determine a species geographic range based on relative suitability. The choice of variables to include in a model may affect the relationships between the response and the predictors. Model selection therefore plays a vital role in the process of producing distribution maps.

Model selection is an important element of all statistical analysis (Burnham *et al*. 2001). Criteria derived from information theory, such as Akaike's Information Criteria (AIC) or Bayesian Information Criteria (BIC), are often used as formal measures of the evidence in favour of competing statistical models (Akaike 1974). However, in the case of presence-only modelling, these formal information criteria lose their theoretical justification. As there are no known absence points, models are fitted using 'pseudo-absences' or background points. There are no set rules for deciding on the number of pseudo-absence points that should be used when fitting and evaluating the models. Therefore, there are no justifiable methods for calculating the likelihood. The number of observations of one of the predicted classes is essentially arbitrary. Taking a number of pseudo-absences that matche the number of presence points does not resolve the issue, as there is no justification for assuming that the number of presence points is related to the number of true absence points in this manner. Imposing restrictions on the number of pseudo-absence (background) points will tend to reduce the amount of information available to the model regarding the range of environmental conditions. This causes poor performance against any criteria that assesses general discriminatory ability. Thus, much larger numbers of pseudo-absence (background) points than presence points are usually used when fitting models. As a result, conventional statistical diagnostics cannot be applied to models that have neither residuals nor measures of deviance. This makes model selection very challenging. Furthermore, as it is not possible to interpret model output as probability values, there are no obvious rules for setting thresholds based on a predicted proportion of occupied sites or pixels.

As a result of these challenges, receiver operator characteristics (ROCs) are commonly used in order to both select the 'best' models and suggest threshold values (Swets 1988,

Swets *et al*. 2000). ROC analysis has a long history of use in signal detection in order to depict the trade-offs between hits and false alarms (Egan 1975). The technique is commonly used in medical decision making. The ROC has a number of attractive properties which suggest that it can be successfully applied in the context of predictive species distribution modelling. Among these is the measure of prediction success that is provided by the area under the curve (AUC). ROC graphs are produced by plotting sensitivity as a function of commission error as some threshold value changes. The AUC provides a summary of the discrimination value of the model over all possible thresholds. ROC analysis allows the consequences of setting thresholds in order to divide the continuum of values provided by a classification process into binary outcomes. In the context of species distribution modelling, the binary outcomes are predictions of presence or absence at a given site. There are four possible outcomes. If an organism is truly present and it is predicted as present, it is counted as a true positive (TP); if the organism is classified as absent when truly present, it is a false negative (FN). If the organism is absent and it is classified as absent, it is a true negative (TN); if the organism is classified as present, it is a false positive (FP). FPs are also often called errors of commission, and FNs are called errors of omission. Several measures can then be derived. The TP rate or sensitivity is the proportion of TPs correctly classified as positive. The specificity is the proportion of TNs correctly classified as negative. The FP rate or commission error is the proportion of TNs incorrectly classified as positive.

ROC graphs are produced by plotting sensitivity (TP) as a function of commission error (FP) as some threshold value changes. The AUC provides a summary of the discrimination value of the model over all possible thresholds. The AUC is equivalent to the probability that the classifier will rank a randomly chosen TP instance (presence) higher than a randomly chosen TN instance (absence). It has been shown to be equivalent to the non-parametric Wilcoxon test of ranks (Hanley and McNeil 1982).

When presence-only data are used, it appears that ROC curves are not applicable. There is no source of negative instances with which to measure specificity. However, Phillips *et al*. (2006) affirm that the problem can be avoided by considering the classification problem to be one of distinguishing presences from random (i.e. background or pseudo-absences). The AUC that is obtained is a measure of discrimination rather than prediction. It is included in the diagnostic component of the SDM software 'MAXENT' (Phillips *et al*. 2006). MAXENT has been suggested as the preferred contemporary modelling framework when small samples are available (Hernandez *et al*. 2006, Papes and Gaubert 2007, Pearson *et al*. 2007).

The AUC is quantified using a subset of the input points that are withheld when fitting the model. Models with the highest AUC values are then usually selected over models with lower AUC values. The underlying basis of this approach has been criticised on theoretical grounds (Lobo *et al*. 2008, Peterson *et al*. 2008). AUC analysis using background points of pseudo-absences is clearly difficult to interpret. There is a danger that AUC could provide a justification for extremely complex models that cannot be interpreted in the context of a species ecological niche. Unlike model evaluation based on information theory, AUC analysis does not include any penalty for the number of parameters used. Models of greatly varying complexities are all evaluated using essentially the same criterion. If AUC analysis suggests that a model has a good discriminatory performance, it also implies that the set of variables used in the model also have good predictive and explanatory properties. However, when presence-only data are used, the model's performance will have been judged only on the ability to discriminate between two ill-defined classes of observations. Model performance has not been subjected to a rigorous statistical analysis that would provide the formal measure of a model's explanatory ability. The role of each variable in

producing the spatial predictions may be unclear. Furthermore, if no independent data set is available, any ability to discriminate within the data used for model fitting cannot be equated with an ability to predict truly new data. Splitting the available data into training and validation sets only partially addresses this issue. Decisions regarding the manner in which the data are divided are arbitrary and the data sets are not truly independent.

In order to test whether these concerns are valid in a practical setting, we tested the properties of AUC analysis using a large set of herbarium-derived presence points. Models were fitted to both real and simulated environmental variables. If the set of variables used as predictors are the result of a purely random process with no possible implications for species distributions, then any fitted models will also have no biological meaning. These could be referred to as 'pseudo-models'. Pseudo-models may still have convincing discrimination value as measured by AUC. If so, they may be selected over more meaningful models that would be much better predictors of new data. Interpretation of 'pseudo-models' in terms of niche space would therefore be misleading and could suggest 'pseudo-niches'.

Methods

Presence-only data were obtained for 2995 species of Mesoamerican trees and shrubs from the Missouri Botanical Garden's TROPICOS database. Only specimens with recorded geographical coordinates were used. Records that had clearly been rounded to the nearest degree were discarded. We used 2000 of these species in the models with a maximum number of occurrence points of 1239 and a minimum of 8. The total number of points for all species used was 131,780. Despite the apparent size of this database, most species collections have been obtained from a very limited number of localities. Few areas have been exhaustively sampled (Figure 1). Many of the collections were obtained from accessible

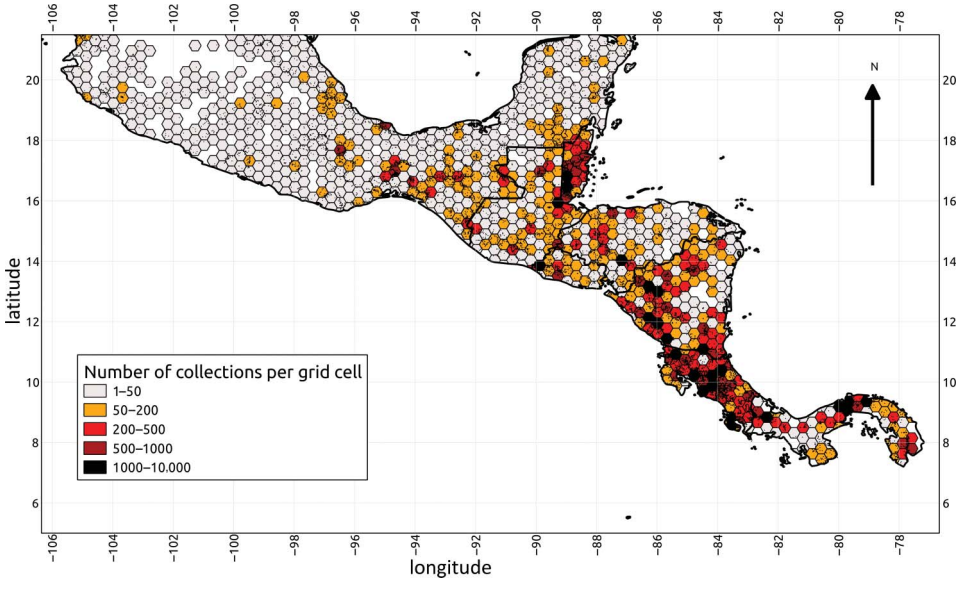

Figure 1. Spatial pattern of input data shown as number of botanical collections within 0.104 square degree hexagonal cells (1165–1272 km^2). Collection points are highly aggregated within these cells. At a higher resolution, they can be seen to be closely associated with roads, urban areas and a few highly studied field sites.

sites close to roads and urban areas. This provides the usual justification for modelling using presence-only data.

The input to models that aimed to predict the species potential distribution based on climatic variables were layers derived from the WorldClim data set. After analysis of co-linearity using PCA, we selected seven original and derived variables based on the criteria of orthogonal properties and interpretability, namely

(1) maximum temperature in June (tmax6),
(2) minimum temperature in January (tmin1),
(3) the difference between the maximum monthly maximum temperature and the minimum monthly minimum temperature (AMTD),
(4) the maximum difference between the maximum and minimum temperature for any single month (DMTD),
(5) mean monthly precipitation in January (prec1),
(6) mean monthly precipitation in June (prec6) and
(7) number of months with over 100 mm precipitation (GM).

The layers were resampled to a 3-arc second grid through nearest neighbour re-sampling. Figure 2 shows the spatial patterns of these seven variables on a comparable scale.

In order to produce layers with no ecological interpretation, we used the R package random fields (Schlather 2008). We generated seven stable Gaussian random fields with mean 20, variance 5, nugget 0, alpha 2 and scale parameters varying between 10 and 100. Figure 3 shows the spatial pattern of these 'pseudo-variables'. Any associations between the values and the pattern of variability shown by a genuine environmental variable are the result of chance and shared spatial autocorrelation. Random fields with low-scale parameters have lower spatial autocorrelation.

We investigated a range of modelling techniques including MAXENT, regression trees, generalized linear model (GLM) and generalized additive model (GAM). The results from MAXENT were found to be similar for a given combination of species and predictor variables to those obtained using GAM with pseudo-absences derived from a sample of 5000 background pixels. For brevity, we therefore present the results from fitting GAM models using the gam package (Hastie 2008) in R (R Development Core Team 2010). Initial tests showed that our general qualitative conclusions applied equally to all other modelling techniques, with the exception of rule-based models such as classification and regression trees (Brieman *et al*. 1984).

For each species, we generated 5000 random pseudo-absence points with no spatial constraints. This is equivalent to the default method of generating background points used by MAXENT. The presence and absence points were randomly split into two equal parts: one part was used for fitting the model and the second for validation. ROCs were calculated and visualised using the PresenceAbsence package (Freeman 2007). In order to test whether models fitted to single random fields could compete with models fitted to single climate variables, we fitted models to each single variable in turn. We then produced full models by including all seven climate variables and all seven random fields in separate models.

As a potential alternative to using AUC, we also tested the utility of AIC by applying automated backwards stepwise variable selection to the models fitted using all 14 variables at once.

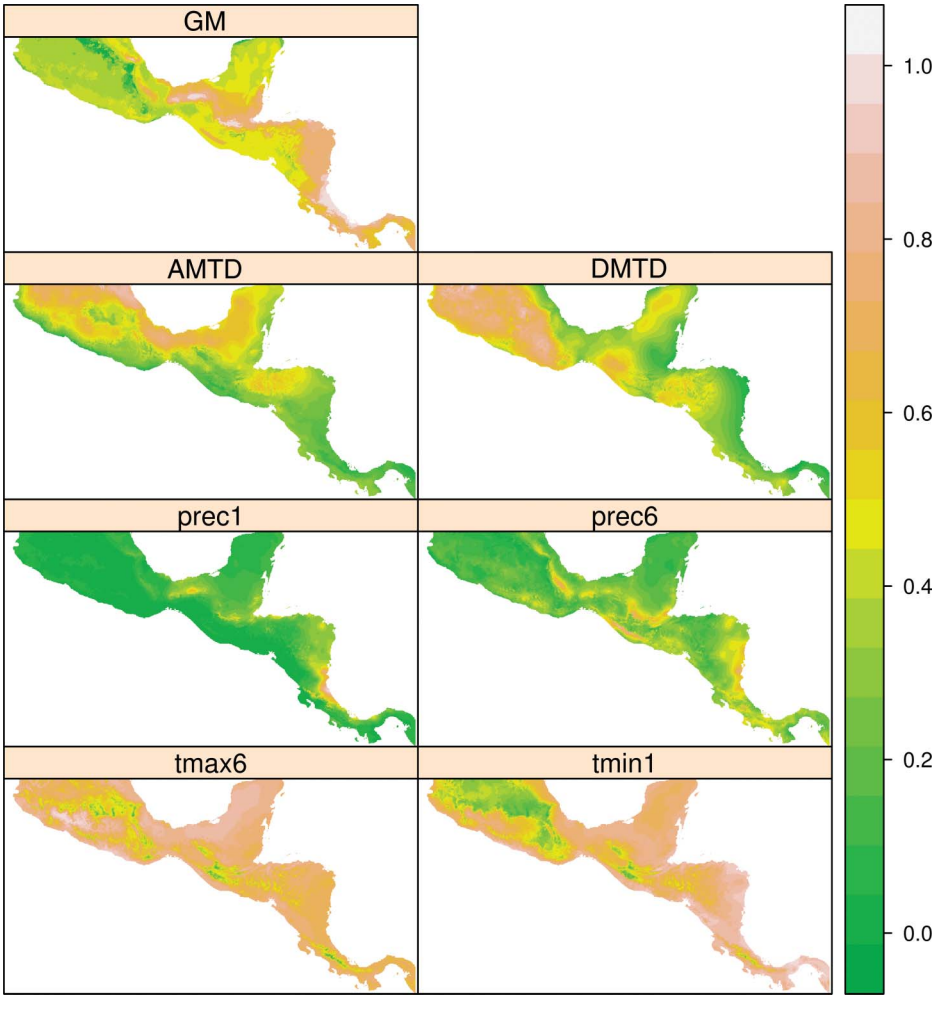

Figure 2. Spatial pattern of the climatic variables used in the models. The layers are shown after normalisation onto a scale between 0 and 1 in order to facilitate visual comparisons.

Results

The distribution of AUC scores for the full set of 2000 species is shown in Figure 4. Simple models based on one single variable had the lowest AUC values. All models based on a single variable had low discriminatory ability. They also produced spatial predictions that probably overestimated the geographic ranges for most of the species. In the case of these very simple models, the values of AUC provided a reasonable measure of the relative importance of each environmental variable. AUC scores for models fitted using informative variables were consistently higher than those for models with random fields as input. Paired comparison showed AUC to be 0.14 ($\pm$0.3) points higher overall for models fitted to genuine environmental variables than to random fields. Random fields with low spatial autocorrelation had the worst performance as measured by AUC. However, there was considerable overlap between the scores of environmental variables such as precipitation,

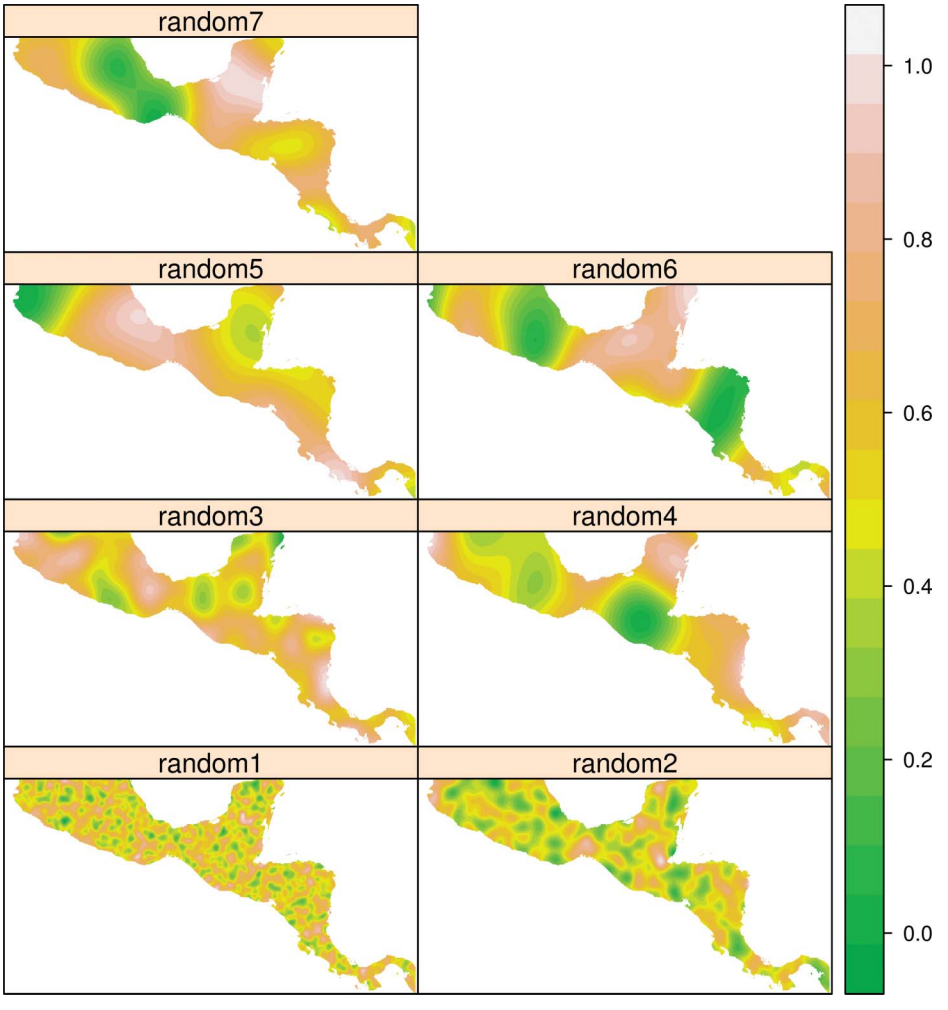

Figure 3. Spatial pattern of the random fields used in the models. The layers are shown after normalisation onto a scale between 0 and 1 in order to facilitate visual comparisons.

which is auto-correlated over quite a large range, and the scores for random field models with high range parameters.

When models were built using seven variables, the complex climate-based models were found to have statistically significantly higher AUC values than complex pseudo-models based on random fields ($P < 0.05$). However, the effect size for paired comparisons was small (0.027 ± 0.005). The overlap between the scores was very high (Figure 3). On a case-by-case basis, AUC values were higher for the pseudo-models for 354 species out of 2000. Therefore, if model selection were to be based on AUC, a misleading pseudo-model might be selected for 17% of the species.

Backwards stepwise model selection based on AIC resulted in the full model being retained for either sets of variables in almost all cases. This demonstrated that AIC is not useful for model selection within this particular context. This occurs as a result of

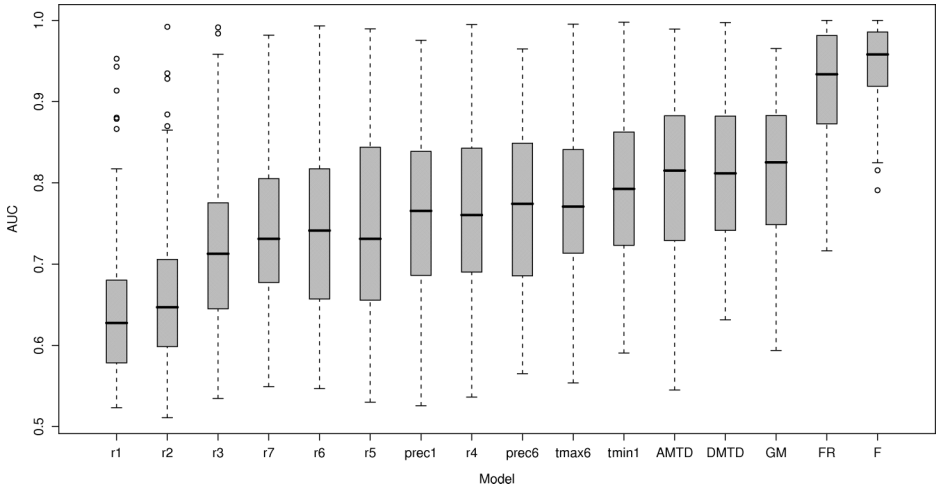

Figure 4. AUC scores for single variable models and full models. r1–r7 are the results from fitting models to single random fields arranged in order of the scale parameter. Models were also fit to the single variables: prec1, precipitation in January; prec6, precipitation in June; AMTD, annual mean temperature difference; DMTD, daily mean temperature difference; GM, growing months with rainfall exceeding 100 mm; tmin1, minimum temperature in January; tmax6, maximum temperature in June. FR is a 'full' random model using all seven random fields as input and F is the full climate model.

pseudo-absences, making an overly large contribution to the calculation of total deviance. The likelihood calculations are unreliable on *a priori* grounds.

Models fit to species with a small number of presence points had a significant tendency to produce higher AUC scores, although the relationship was not linear. The shape of the trend fitted using loess smoothing and the amount of scatter is shown in Figure 5. All the smoothed curves were statistically significant, given the large sample size ($n = 2000$). The relationship between small number of presence points and high AUC values was clearer when the full models using all seven variables were considered. Figure 6 shows that AUC values were highest when points were spatially aggregated, as measured by the mean distance between points.

Figures 7 and 8 show two examples of the spatial predictions from the models. The threshold used to produce binary maps was placed at a sensitivity of 0.8 in order to include at least 80% of the presence points. The single variable models, such as the temperature only model, clearly have low discriminatory power as measured by the AUC. This extends the modelled potential area of occurrence when compared with the full model based on either the random fields or the set of climatic variables. However, the more complex models have similar discriminatory ability in both cases. The match between the spatial predictions of the full models and the pseudo-models can be seen to vary, although models have very similar AUC scores.

Discussion

Effect of presence-only data on AIC and AUC

The results cast doubt on the level of confidence that should be placed on AUC as a guide for model selection using presence-only data. Stepwise selection using AIC is not an

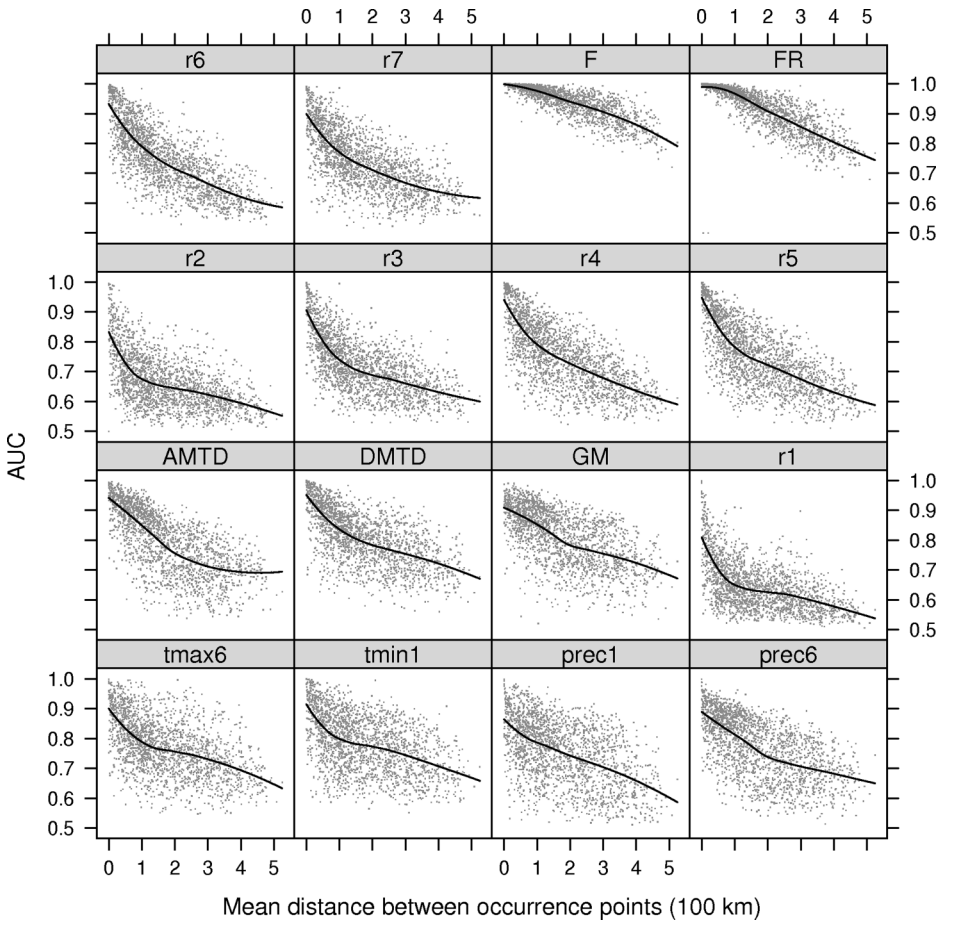

Figure 5. Relationship between AUC and number of occurrence points used in modelling. Lines show the results of fitting splines to show the shape of the trend. All relationships were significant at $P < 0.01$.

alternative, as the calculation of the likelihood is based on an ill-defined class leading to predictable over fitting, as demonstrated by the failure of backwards stepwise model selection to remove variables. The problems with both AIC and AUC arise through a common cause. They are the result of the use of a single class of observations within an analysis that is designed to use two classes. The absence class is too ill defined to be used as a valid measure of model performance. The 'errors of commission' that are used in a conventional ROC analysis cannot be equated to the commission of background points. In the case of presence-only modelling, the errors of commission occur when a species is 'erroneously' predicted to occur at a point on the map where the species has not been recorded. However, if a model predicts the species to be present only where it has already been detected, then it would have a perfect AUC score, but at the same time, it would serve no useful purpose. Such a model would simply replicate the pattern of known data points. In order to be predictive, a species distribution must, by definition, commission some points in order to be considered successful. Errors of commission must therefore be down-weighted in some way in comparison to errors of omission (Anderson 2003). However, ROC analysis

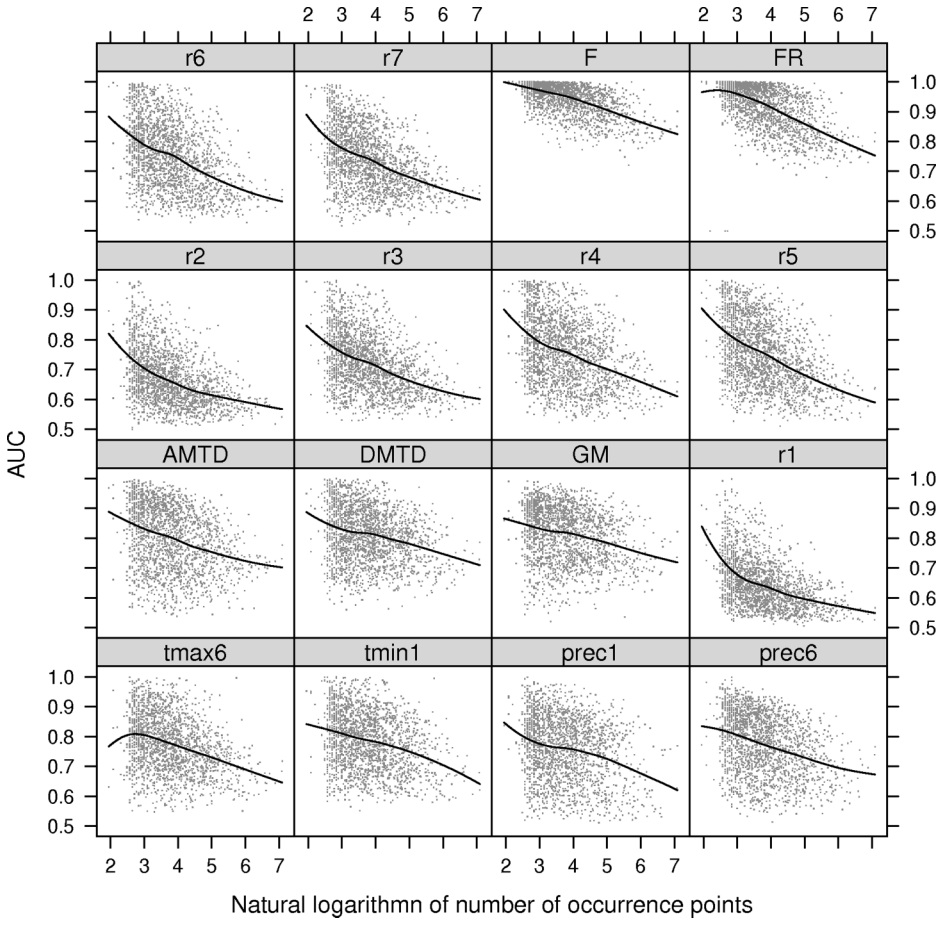

Figure 6. Relationship between AUC and mean distance between pairs of points. Lines show the results of fitting splines to show the shape of the trend. All relationships were significant at $P < 0.01$.

provides no guidance regarding how much down-weighting is optimal, nor how to assign weights to points with differing probabilities of true absence. It is therefore impossible to evaluate whether the model is over-fitting or under-fitting using AUC values alone.

Effect of small numbers of data points and spatial aggregation

A further practical problem arises. All else being equal, a model built from a small number of data points would be expected to be less reliable than a model built using a large number of data points. However, the results of AUC analysis suggest the opposite. This has been pointed out previously by Lobo *et al*. (2008). When random background points are used to calculate the AUC, the abscissa of the ROC may be interpreted as the proportion of the total area predicted as present, rather than a measure based on correctly or incorrectly predicting true absence, which is unknown (Peterson *et al*. 2008). This issue is further compounded by the strong positive relationship between AUC values and spatial aggregation. Spatially aggregated observations of predictor variables are naturally auto-correlated. This

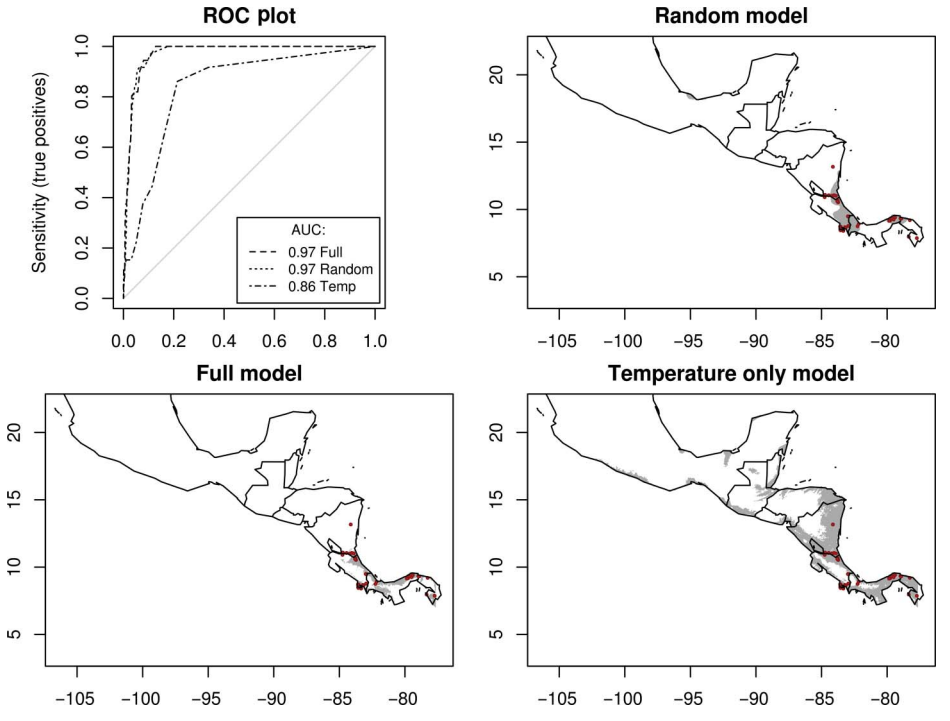

Figure 7. Example of model output showing the AUC and the spatial predictions for three models using the presence points for *Garcinia madruno*. Grey shaded areas represent the predicted spatial distribution using a threshold placed at a sensitivity of 0.8 in order to include at least 80% of the presence points shown in red.

lack of independence leads to fewer degrees of freedom for the predictor variables than would otherwise be the case. Lennon (2000) points out that correlation between a spatially clustered response variable and a set of explanatory variables is strongly biased in favour of those explanatory variables that are themselves highly auto-correlated. Similarly, multiple regression analysis finds highly auto-correlated explanatory variables to be significant much more frequently than it should. Thus, the chances of mistakenly identifying a significant slope across an auto-correlated pattern are very high when classical regression is used. These results show a similar pattern. Auto-correlated spatial data have complex statistical properties that make any form of inference challenging.

Diagnostic issues and interpretation of AUC

If models are simply used as pragmatic tools for interpolating between known occurrences and guiding the process of estimating potential distributions, then the issues may not be critical. AUC does provide some measure of the ability of a model to discriminate between background points and presence points. However, users of SDMs often wish to interpret models in terms of a species niche. Specialised species with very narrow niches often occur in a limited number of geographical locations. In such cases, a fitted model may be interpreted as describing the precise combination of conditions that the species requires. However, the results show that models fitted using random fields can just perform as well as models fitted to genuine environmental variables. In fact, any delimited region of space

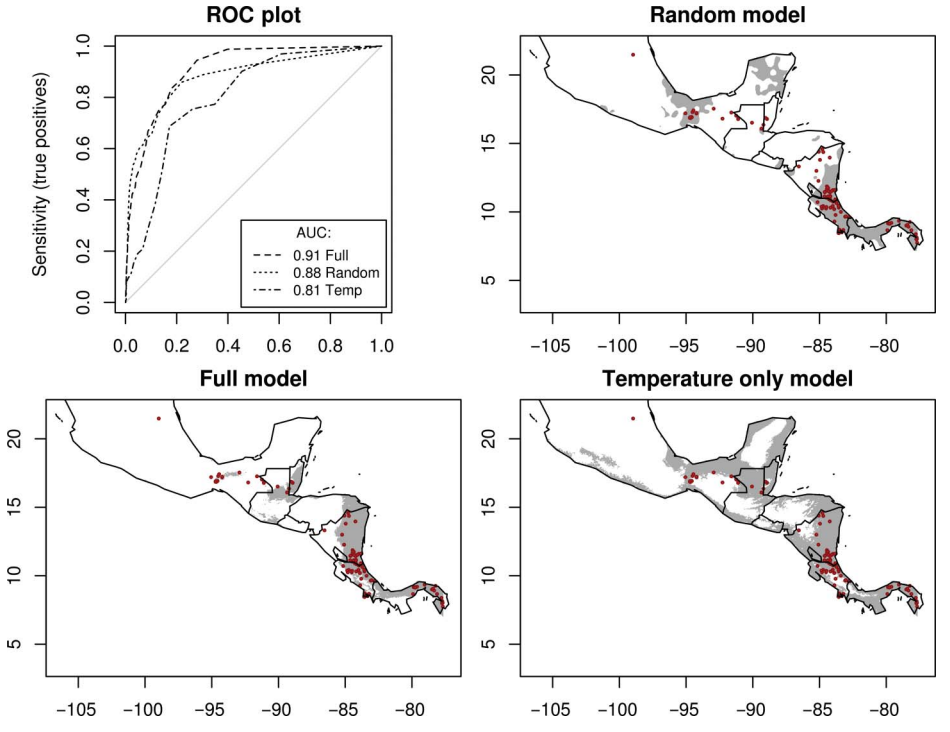

Figure 8. Example of model output showing the area under the curve and the spatial predictions for three models using the presence points for *Hamelia axillaris*. Grey shaded areas represent the predicted spatial distribution using a threshold placed at a sensitivity of 0.8 in order to include at least 80% of the presence points shown in red.

that approximates to the distribution of collection points, for example, a convex hull drawn around the points, will also have a high AUC. If the data points are highly aggregated, a simple convex hull could outperform niche models. However, such a purely spatial construction would provide no insight into species relationship with the environment. The use of multiple environmental variables and a machine learning algorithm leads to a similar effect that is not being made explicit. Complex combinations of variables provide flexibility, allowing any model to fit itself neatly around the data points. Thus, there is real danger of inferring 'pseudo-niches'.

Dangers of overinterpretation

There are already published reports of AUC that could be misleading. Broenniman *et al*. (2006) report average AUC values of 0.95 when modelling endemic life forms of South Africa using presence-only data, stating that this showed that models had excellent predictive ability. Although their study split the available data into a training and validation component, the approach does not provide for truly independent evaluation (Chatfield 1995). They reported that only 3 species out of 975 showed poor predictive ability (AUC < 0.7). In contrast, Wisz *et al*. (2008) used fully independent presence–absence data in order to calculate AUC. In the second study, species with relatively large sample sizes (>100) obtained AUC values below 0.7. They stated that models built using smaller sample sizes had AUC

values that were only slightly better than chance predictions. This supports our suggestion that reliable models built with sound data can be expected to have lower AUC values than empirically weaker models based on spatially aggregated presence-only data.

The issue of museum collection bias

The problems are compounded by the *ad hoc* nature of the sampling that led to the presence data. If collection effort has been spatially concentrated, the apparent relationship with environmental variables may simply reflect the limited extent of collecting activity. We suspect this is an explanation for the pattern of occurrence of many of the species of trees in the data set we used. When this occurs, a calculation of AUC using truly independent data will provide much lower values. When no formal system has been used for data collection, systematic error is expected to greatly outweigh random error. This should be explicitly taken into account (Elith *et al*. 2002). Although data from historical museum collections may be all that is available for many parts of the world, their extremely poor statistical properties should be fully recognised. It has been suggested that only points previously visited by collectors should be used to generate pseudo-absences. However, if a large proportion of the geographical area remains unvisited by collectors, this is not practical and imposes its own biases. The best long-term solution is to begin a process of systematic data collection, aimed at providing the insight into ecological processes and the predictive power that will always be impossible to achieve using data from historical collections.

Solutions

These results should serve as a warning when interpreting AUC. In an applied setting, there is a real danger that pseudo-niches could be over optimistically inferred from models fitted using presence-only data. Wisz *et al*. (2008) state that 'given the highly dimensional, complex nature of ecological niches of species large numbers of samples *may* be needed to allow for accurate description of the range of conditions over which a species occurs'. The results here suggest that a large number of samples are not an option. Rather, it is essential to avoid the suggestion of misleading niche spaces. A much stronger conceptual framework is needed for selecting variables (Austin 2007, Williams *et al* 2007).

The issue may be avoided by stating clearly that AUC values are simply relative measures of model's discriminatory ability. Predictions of probabilities of occurrence cannot be derived from presence-only modelling. Output from a presence-only based model is in fact unlikely to have a simple linear correspondence with the scores produced by a model based on known presences and absences. Thus, without independent validation, the maps produced by SDMs are simply pragmatic tools that may aid in determining the limits of species distributions, but do not by themselves predict them. If true presence–absence data are available, the AUC could be a valid measure of predictive ability. In this case, models may also be attributed with some explanatory ability. Even so, there are many other valid caveats that must be carefully considered (Lobo *et al*. 2008). ROC should never be applied uncritically.

Conclusion

The results show that the measure of discrimination provided by AUC can select spurious 'pseudo-models' with neither predictive nor explanatory value. AUC is a valid measure of the ability of a model to discriminate between two classes of input data. However, when

one class is ill-defined, the interpretation of the AUC score is challenging. The use of pseudo–absence or background values does not allow AUC to be meaningfully interpreted as a measure of strength of evidence in support of a model. The results demonstrate that comparisons of model performance using AUC scores naturally lead to the selection of complex models which include many variables. Because a chosen set of predictor variables could include auto-correlated features with no biological meaning, inference on multidimensional niche spaces should not usually be supported by AUC values alone. We suggest that AUC-based model validation should be restricted to data from inventories in which true absence can be calibrated against detection probability. Claims of 'excellent' predictive ability based on AUC values should be approached with caution. Models built from intrinsically limited data can never be expected to address the 'Wallacean shortfalls' that arise as a result of incomplete spatial sampling (Bini *et al*. 2006).

References

Akaike, H., 1974. New look at statistical-model identification. *IEEE Transactions on Automatic Control*, AC19 (6), 716–723.

Anderson, R.P., 2003. Real vs. artefactual absences in species distributions: tests for oryzomys albigularis (rodentia: Muridae) in venezuela. *Journal of Biogeography*, 30 (4), 591–605.

Araújo, M.B., *et al*., 2004. Would climate change drive species out of reserves? An assessment of existing reserve-selection methods. *Global Change Biology*, 10 (9), 1618–1626.

Araújo, M.B. and Rahbek, C., 2007. Conserving biodiversity in a world of conflicts. *Journal of Biogeography*, 34 (2), 199–200.

Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecological Modelling*, 200 (1–2), 1–19.

Berry, P.M., *et al*., 2002. Modelling potential impacts of climate change on the bioclimatic envelope of species in Britain and Ireland. *Global Ecology and Biogeography*, 11 (6), 453–462.

Bini, L.M., *et al*., 2006. Challenging Wallacean and Linnean shortfalls: knowledge gradients and conservation planning in a biodiversity hotspot. *Diversity and Distributions*, 12 (5), 475–482.

Breiman, L., *et al*., 1984. *Classification and regression trees*. Belmont, CA: Wadsworth.

Broenniman, O., *et al*., 2006. Do geographic distribution, niche property and life form explain plants' vulnerability to global change? *Global Change Biology*, 12 (6), 1079–1093.

Burnham, K.P., *et al*., 2001. Suggestions for presenting the results of data analyses. *Journal of Wildlife Management*, 65 (3), 373–378.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A*, 158, 419–466.

Egan, J.P., 1975. *Signal detection theory and ROC analysis*. New York: Academic Press.

Elith, J., Burgman, M.A., and Regan, H.M., 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling*, 157 (2–3), 313–329.

Elith, J., *et al*., 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29 (2), 129–151.

Elith, J. and Leathwick, J., 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions*, 13 (3), 265–275.

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 (2), 263–274.

Freeman, E. (2007). *PresenceAbsence: An R Package for Presence-Absence Model Evaluation*. Ogden, UT: USDA Forest Service, Rocky Mountain Research Station.

García, A., 2006. Using ecological niche modelling to identify diversity hotspots for the herpetofauna of pacific lowlands and adjacent interior valleys of Mexico. *Biological Conservation*, 130 (1), 25–46.

Golicher, D.J., *et al*., 2008. Applying climatically associated species pools to the modelling of compositional change in tropical Montane forests. *Global Ecology and Biogeography*, 17 (2), 262–273.

Graham, C.H., *et al*., 2004. New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology & Evolution*, 19 (9), 497–503.

Hanley, J.A. and McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29–36.

Hastie, T., 2008. *gam: Generalized Additive Models* [online]. R package version 1.0. Available from: http://cran.r-project.org/ [Accessed May 2011].

Hernandez, P.A., *et al*., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29 (5), 773–785.

Kuper, W., *et al*., 2006. Deficiency in African plant distribution data – missing pieces of the puzzle. *Botanical Journal of the Linnean Society*, 150 (3), 355–368.

Lennon, J.J., 2000. Red-shifts and red herrings in geographical ecology. *Ecography*, 23 (1), 101–113.

Lobo, J.M., Jiménez-Valverde, A., and Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17 (2), 145–151.

Moilanen, A., *et al*., 2006. Uncertainty analysis for regional-scale reserve selection. *Conservation Biology*, 20, 1688–1697.

Newton, A.C. and Oldfield, S., 2008. Red listing the world's tree species: a review of recent progress. *Endangered Species Research*, 6, 137–147.

Papes, M. and Gaubert, P., 2007. Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (mammalia, carnivora) across two continents. *Diversity and Distributions*, 13, 890–902.

Pearson, R.G., *et al*., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. *Journal of Biogeography*, 34 (1), 102–117.

Peterson, A.T. and Kluza, D.A., 2003. New distributional modelling approaches for gap analysis. *Animal Conservation*, 6, 47–54.

Peterson, A.T., Papes, M., and Soberon, J., 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. *Ecological Modelling*, 213 (1), 63–72.

Peterson, A.T. and Vieglais, D.A., 2001. Predicting species invasions using ecological niche modeling: new approaches from bioinformatics attack a pressing problem. *Bioscience*, 51 (5), 363–371.

Phillips, S.J., Anderson, R.P., and Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190 (3–4), 231–259.

R Development Core Team (2010). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Schlather, M., 2008. *RandomFields: simulation and analysis of random fields*. R package version 1.3.35. Available from: http://cran.r-project.org/ [Accessed May 2011].

Stockwell, D.R.B., 2006. Improving ecological niche models by data mining large environmental datasets for surrogate models. *Ecological Modelling*, 192 (1–2), 188–196.

Stockwell, D.R.B. and Noble, I.R., 1992. Induction of sets of rules from animal distribution data – a robust and informative method of data-analysis. *Mathematics and Computers in Simulation*, 33 (5–6), 385–390.

Swets, J., 1988. Measuring the accuracy of diagnostic systems. *Science*, 240, 1285–1293.

Swets, J.A., Dawes, R.M., and Monahan, J., 2000. Better decisions through science. *Scientific American*, 283, 82–87.

Thuiller, W., Lavorel, S., and Araújo, M.B., 2005. Niche properties and geographical extent as predictors of species sensitivity to climate change. *Global Ecology and Biogeography*, 14 (4), 347–357.

Whittaker, R.J., *et al*., 2005. Conservation biogeography: assessment and prospect. *Diversity and Distributions*, 11 (1), 3–23.

Williams, K.J., *et al*., 2012. Which environmental variables should I use in my biodiversity model? *International Journal of Geographical Information Science*, 1–39. doi:10.1080/13658816.2012.698015.

Wisz, M.S., *et al*., 2008. Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, 14, 763–773.