Borneo : a quantitative analysis of botanical richness, endemicity and floristic regions based on herbarium records

Raes, N.

CHAPTER 3

A null-model for significance testing of presence-only species distribution models

Niels Raes and Hans ter Steege

Forum

Ecography 30 (2007) 727-736

Species' distribution models (SDMs) attempt to predict the potential distribution of species by interpolating identified relationships between species' presence/absence, or presence-only data on one hand, and environmental predictors on the other hand, to a geographical area of interest. Currently, they are widely applied in biogeography, conservation biology, ecology, palaeo-ecology, invasive species studies, and wildlife management (Guisan & Zimmermann, 2000; Araújo & Pearson, 2005; Thuiller *et al.*, 2005; Araújo & Guisan, 2006; Guisan *et al.*, 2006; Peterson, 2006). More recently, vast numbers of herbarium and natural history museum collections have become available (Graham *et al.*, 2004) and techniques to apply this special type of presence-only data have been developed (Hirzel *et al.*, 2002; Anderson *et al.*, 2003; Elith *et al.*, 2006; Pearce & Boyce, 2006; Phillips *et al.*, 2006). Despite the widespread use of SDMs, several high-priority research interests remain to be investigated (Guisan & Thuiller, 2005; Araújo & Guisan, 2006). One of these is the improvement of SDM validation, or the quantification of a model's predictive performance (Araújo & Guisan, 2006).

The fact that the standard validation procedures for an SDM are not sufficient to assess the applicability of an SDM in a predictive context, was first shown by Olden *et al.* (2002). They showed that after SDM validation it is critical to assess whether the SDM prediction differs from what would be expected on the basis of chance alone. SDMs producing random predictions are neither helpful nor useful (Olden *et al.*, 2002). Thus, in this paper we introduce a null-model methodology that allows testing whether SDMs developed with presence-only data differ significantly from what would be expected by chance. We also demonstrate that it is critical and possible to correct for collector-bias in specimen data in this test.

SDM validation and measures of accuracy

Validation of SDMs can be carried out with several different measures of model accuracy. The most widely applied measures of model accuracy include sensitivity, specificity, Cohen's kappa, and the area under the curve (AUC) of the receiver operating characteristic (ROC) plot (Fielding & Bell, 1997; Manel *et al.*, 2001; McPherson *et al.*, 2004). Most measures of SDM accuracy, including the four mentioned above, are directly or indirectly derived from a confusion matrix (see Fielding and Bell 1997). Sensitivity quantifies the proportion of observed presences correctly predicted as presence, the true positive fraction. Specificity quantifies the true negative fraction. Cohen's kappa quantifies overall agreement between predictions and observations, corrected for agreement expected to occur by chance.

These three measures of accuracy require that probabilities of occurrence obtained with SDMs are transformed into discrete presences or absences, for which purpose a threshold of 0.5 is commonly used (McPherson *et al.*, 2004; Liu *et al.*, 2005; Jiménez-Valverde & Lobo, 2007).

The AUC value of the ROC plot is a method that does not require discrete presence/absence predictions, and is therefore a measure of accuracy that is threshold independent (Pearce & Ferrier, 2000; McPherson *et al.*, 2004).

The ROC plot is obtained by plotting sensitivity as a function of the falsely-predicted positive fraction, or commission error (1-specificity), for all possible thresholds of a probabilistic prediction of occurrence. The resulting area under the ROC curve provides a single measure of overall model accuracy, which is independent of a particular threshold. AUC values range from 0 to 1, with a value of 0.5 indicating model accuracy not better than random, and a value of 1.0 indicating perfect

model fit (Fielding & Bell, 1997). An AUC value can be interpreted as indicating the probability that, when a presence site (site where a species is recorded as present) and an absence site (site where a species is recorded as absent) are drawn at random from the population, the presence site has a higher predicted value than the absence site (Elith *et al.*, 2006; Phillips *et al.*, 2006).
All four measures of model accuracy were tested extensively for statistical artefacts, and the AUC value was the only measure of SDM accuracy that was invariable to the proportion of the data representing species' presence, known as prevalence (Pearce & Ferrier, 2000; Manel *et al.*, 2001; McPherson *et al.*, 2004). Insensitivity to prevalence is of special relevance when the AUC values are used to assess model accuracy for SDMs that have been developed with presence-only data. When the required absences are lacking, they are replaced by pseudo-absences. Pseudo-absences are sites, randomly selected across the geographical area of interest, at localities where no species presence was recorded and for which species occurrence is set as absent (Ferrier *et al.*, 2002; Anderson *et al.*, 2003; Elith *et al.*, 2006; Phillips *et al.*, 2006). A sufficiently large sample of pseudo-absences is needed to provide a reasonable representation of the environmental variation exhibited by the geographical area of interest, typically 1,000-10,000 points (Stockwell & Peters, 1999; Ferrier *et al.*, 2002; Phillips *et al.*, 2006). These large numbers of pseudo-absences automatically result in low prevalence values. The number of records by which a species is represented in herbaria and natural history museums range from one to 150-200 records (Stockwell & Peterson, 2002). Even when a species is represented by 200 unique presence-only records and 1,000 pseudo-absences are used, prevalence is only 16.7% (200/1200).
A major drawback of using pseudo-absences, however, is that the maximum achievable AUC value indicating perfect model fit, is no longer 1, but 1-*a*/2 (where *a* is the fraction of the geographical area of interest covered by a species' true distribution, which typically is not known (Phillips *et al.*, 2004; Phillips *et al.*, 2006). Nevertheless, random prediction still corresponds to an AUC value of 0.5. Therefore, standard thresholds of AUC values indicating SDM accuracy (e.g., the threshold of AUC>0.7 that is often used; Pearce and Ferrier 2000, Swets *et al.,* 2000, Manel et al. 2001), do not apply.

A null-model approach for significance testing of presence-only SDMs

To test the significance of an SDM we propose to test the AUC value (of the SDM) against a null distribution of expected AUC values based on random collection data (sensu Olden *et al.* 2002). A null-distribution, or null-model, is a model that is based on randomizations of ecological data or random sampling from a known or imagined distribution (Swets *et al.*, 2000; Jetz *et al.*, 2004; Gotelli & McGill, 2006). A null-model is straightforward in theory and closely resembles hypothesis testing in conventional statistical analysis. To build a null-model, first the AUC value of the real SDM is determined. Next, a null-model is generated by randomly drawing collection localities without replacement, from the geographical area for which the species distribution is modelled. The number of randomly drawn collection localities is equal to the actual number of collections for that species. This is repeated 999 times to generate a frequency histogram of AUC values, expected if the null hypothesis is true. The position of the observed AUC value in the null distribution of the 'randomly' generated AUC values is then used to assign a probability value, just as in a conventional statistical analysis (Dolédec *et al.*, 2000; Olden *et al.*, 2002; Gotelli & McGill, 2006). We use a one-sided 95% confidence interval (C.I.) since we are only interested in whether an SDM performs significantly better than expected by chance, rather than assessing whether it performs significantly worse. We interpret a significant model to indicate that the relations between species' presence localities and the predictor variable values at those locations are stronger than can be expected by chance.
An additional advantage of significance testing of an SDM with a null-model is that we can use all presence records to develop and test the SDM. Common practice in measuring an SDM's accuracy is the split-sample approach. This approach splits the available species records into a training and test sample (Fielding & Bell, 1997). It is assumed that a randomly selected test sample from original data constitutes independent observations, which can be used for statistical testing (Araújo *et al.*, 2005). However, such a test sample is not fully independent due to spatial autocorrelation (Araújo *et al.*, 2005; McPherson & Jetz, 2007). Moreover, dependent on the random split, different values of SDM accuracy may be obtained (Phillips *et al.*, 2006). Phillips *et al.* (2006) showed that SDMs for a species represented by 128 records and 10 different random splits, yielded AUC values ranging from 0.819 until 0.903. More extremely, our unpublished results yielded AUC values for a species represented by 8 records ranging between 0.079 and 0.912 based on 100 random splits.
Testing an SDM against a null-model, however, could suffer from one more problem. When drawing random points from a geographical area one assumes that collectors visited all localities equally well. If this condition is not met, which is likely to be the case (Reddy & Davalos, 2003; Romo *et al.*, 2006; Hortal *et al.*, 2007), the randomly drawn points, that are used to develop the null-model, might include ecological conditions that are not represented by the localities from where actual collections were gathered. This bias could results in a significant deviation from the null-model for species that are randomly distributed over the actual collection localities.

The impact of collection bias on significance testing

SDMs predict the presence and absence of a species for a given geographical area, based on the localities where the records were collected and the values of environmental predictors at those sites. SDMs are especially useful when only part of the entire geographical area has been sampled, as is generally the case. This works fine as long as the collection localities are randomly spread over the complete geographical area. Unfortunately, collectors tend to visit areas which are easily accessible, such as areas close to cities, roads, rivers, and nature reserves resulting in serious collection biases (Parnell *et al.*, 2003; Reddy & Davalos, 2003; Kadmon *et al.*, 2004; Hortal *et al.*, 2007). The influence of collection biases on the accuracy of SDMs largely depends on the range of values of each of the environmental variables covered by the collection localities, known as climatic, or environmental bias (Kadmon *et al.*, 2003, 2004). Kadmon *et al.* (2003) showed that environmental biases, expressed as the degree of sampling bias

with respect to the environmental conditions under which a species is known to occur, had a significant negative effect on the predictive accuracy of the SDM. Although this is a serious issue of concern (Araújo & Guisan, 2006), it is not specific to any methodology used to develop SDMs. However, it is relevant when the accuracy of an SDM is tested against a null-model.

When collecting is environmentally biased, an SDM is more likely to deviate significantly from a random null-model that does not include such bias. When, for example, collection localities are biased for mean annual temperature, a significant part of the species' actual temperature range could remain unsampled. When these data are used in an SDM that is tested against a null-model, based on records that were randomly drawn from the entire study area, this species will possibly show a preferred mean annual temperature range compared to the randomly drawn points. It will accordingly more likely deviate significantly from the null-model than its actual range would justify. Such collection bias might thus result in certain areas being systematically under predicted by the SDM. It should be noted, however, that this is true for all distribution modelling methods and can only be solved by additional data collection.

Fortunately, the problem of having a higher chance of significantly deviating from a randomly drawn null-model if collections are biased, can be solved by restricting the randomly drawn points to all known collection localities. Thus, drawing the null-model from a biased distribution. To test for environmental bias in known collection localities a distribution model using all known collection localities is tested against a null-model developed by 100 -1000 times drawing an equal number of random points from the entire study area. If the distribution model's accuracy of known collection localities deviates significantly from this 'second' null-model, then we conclude that the collection localities are environmentally biased. If this is the case then the SDMs have to be tested against a null-model that is based on actual collection localities.

A case study based on Bornean plant collections

To illustrate the applicability of a null-model approach to select SDMs that deviate significantly from random expectation, we selected all occurrences of the genus *Shorea* (Dipterocarpaceae) on the Malesian island Borneo (approx. 8°N - 5°S, 108° - 120°E; Fig. 3.3) from the BRAHMS database of plant collections present at the National Herbarium of the Netherlands, Leiden University, the Netherlands. *Shorea* was selected because this genus has been thoroughly taxonomically revised and species identifications are reliable (Ashton, 1983). The database contained 4466 records of 147 Shorea species for Borneo. Out of these 147 species, 116 were represented by 5, or more, unique collection localities. For those species, we developed SDMs.

To model the species distributions we used environmental predictor variables with a 5 arc-minutes resolution (~10km at the equator). We selected the digital elevation model (DEM) and the 19 bioclimatic variables of the current conditions (~1950-2000) from the WORLDCLIM dataset (hhtp://www.worldclim.org) for Borneo (Hijmans ***et al.***, 2005). Additionally, we selected 15 FAO soil variables (FAO, 2002). We also included a measure of the effect of the El Niño Southern Oscillation Event (ENSO). This variable was expressed as the relative average annual difference in Normalized Difference Vegetation Index (NDVI) between the months of an ENSO, and a non-ENSO year. To this dataset

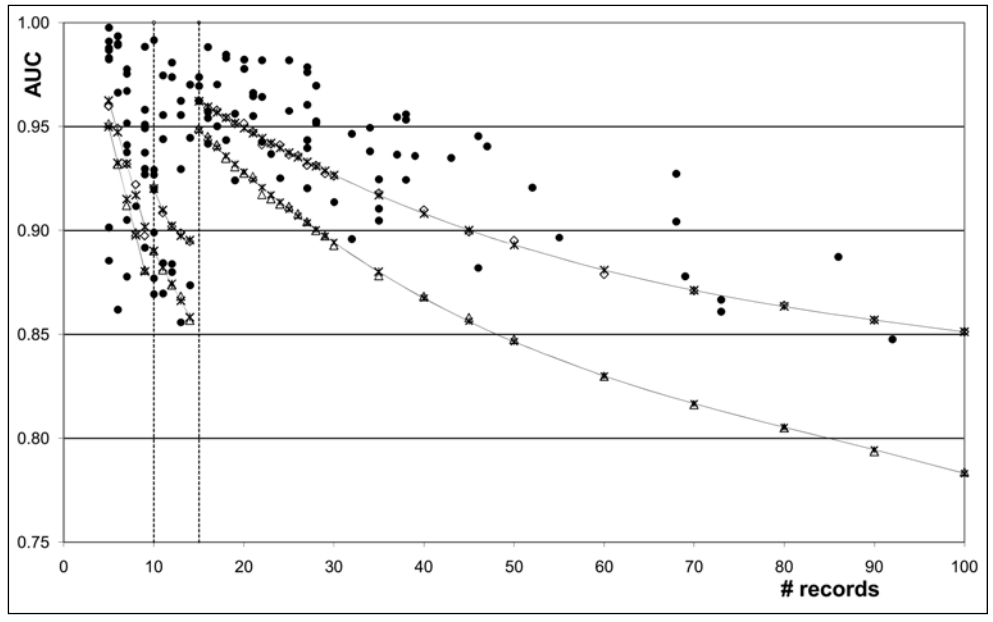

Figure 3.1. Species' distribution model (SDM) AUC values (•), the 95% confidence interval (C.I.) AUC values of the randomly drawn null-models (Δ), and the 95% C.I. AUC values of the environmentally bias corrected null-models (◊). Asterisks give the fitted 95% C.I. AUC values for both series of null-models connected by a line. Vertical dotted lines indicated the consecutive addition to the initial linear modelling features, of quadratic, and hinge features by Maxent. SDM AUC values that are higher than their corresponding 95% C.I. AUC value of the fitted null-model, significantly deviate from what would be expected by random chance ($p<0.05$).

we added the Walsh's index (Walsh, 1996; Leigh Jr., 2004). This index integrates the effects of annual rainfall and its seasonality. Finally, the elevation range derived from the SRTM 90m Digital Elevation Data (http://srtm.csi.cgiar.org/) was added. All data layers were scaled to 5 arc-minute resolution, and resampled to the geographical extent of the most restricted FAO soil variable data layers. This resulted in 8577 data cells for Borneo. All data layer manipulations were performed with Manifold GIS (Manifold Net Ltd).

To model *Shorea* species distributions of Borneo we used Maxent (version 2.3.0; http://www.cs.princeton.edu/~shapire/maxent/) (Phillips ***et al.***, 2006). Maxent, or the maximum entropy method for species' distribution modelling, estimates the most uniform distribution ("maximum entropy") across the study area, given the constraint that the expected value of each environmental predictor variable under this estimated distribution matches its empirical average (average values for the set of species' presence records) (Hernandez *et al.*, 2006; Phillips *et al.*, 2006). Maxent was specifically developed to model species distributions with presence-only data and has outperformed most other modelling applications (Elith *et al.*, 2006; Hernandez *et al.*, 2006; Pearson *et al.*, 2007). An added advantage of Maxent is that it also performs the ROC statistical analysis. Since we tested whether an SDM's AUC value deviates significantly from a null-model, the 'random test percentage' was set to zero resulting in training data only. To avoid the inclusion of multiple presence records in one grid cell per species we set Maxent to 'remove duplicate presence records'. This reduced the total available presence records for the 116 *Shorea* species represented

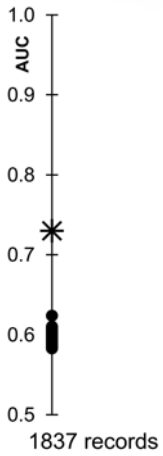

Figure 3.2. The AUC value of the model based on the 1837 collection cells (∗) and the 100 AUC values (•) of models based on 1837 randomly drawn cells from the total 8577 cells of Borneo, indication the 1837 collection cells are significantly environmentally biased ($p<0.01$).

by at least five records to 2552. The modelling rules were set to 'Auto Features' using only linear features when less than 10 records were available, adding quadratic features for SDMs developed with 10 or more and less than 15 records, and including hinge features for species with 15 or more records. Maxent adds product and threshold features for those species represented by 80, or more, records. However, we set Maxent to use linear, quadratic and hinge features for all species represented by at least 15 records, due to odd behaviour of Maxent when product and threshold features were added (explained in the discussion). For each of the 116 *Shorea* species we developed an SDM with Maxent using all presence records under the modelling rules as described above. The number of unique records per species ranged from 5 until 92 (Table S3.1, '# records'). The AUC values of all *Shorea* SDMs are presented as dots in Figure 3.1, and under 'AUC' in Table S3.1.

Testing SDMs against a null-model

To test whether *Shorea* SDMs significantly differed from what would be expected by chance, we calculated the 95% C.I. AUC value for each number of records by which the *Shorea* species were represented. We developed frequency histograms of expected AUC values by randomly drawing points without replacement from all 8577 available cells of Borneo (999 times), and model these with Maxent under the same conditions as the *Shorea* species. We developed frequency histograms of expected AUC values for 5 – 30 records (26 distributions), for 35 – 50 records with intervals of 5 records (4 distributions), and for 60 – 100 records with intervals of 10 records (5 distributions). For each frequency histogram, we assessed the 95% C.I. upper limit AUC value, by ranking the 999 AUC values and selecting the 949th value (0.95 x 999 = 949; Fig. 3.1, triangles). For each of the three resulting sets of 95% C.I. AUC values we applied a curve-fit (Fig. 3.1, asterisks). The fitted 95% C.I. AUC values of the null-models for the number of records by which each *Shorea* species is represented, are given in Table S3.1, '95% C.I. All'.

With the fitted 95% C.I. AUC values, it is now easy to assess which of the *Shorea* species has an accuracy of its SDM that is significantly higher than expected by chance alone ($p<0.05$). This was the case for 105 of the 116 *Shorea* species (91%) which were modelled (Table S3.1, '95% C.I. All').

Testing SDMs against a bias corrected null-model

In order to assess whether the known collection localities are environmentally biased, we selected all databased and georeferenced plant specimen records from Borneo that were present in the BRAHMS database of the National Herbarium of the Netherlands. In total the database contained 142,097 properly georeferenced records. These records could be assigned to 1837 of the total of 8577 grid cells of Borneo This means that only 21.4% of the grid cells of Borneo have been

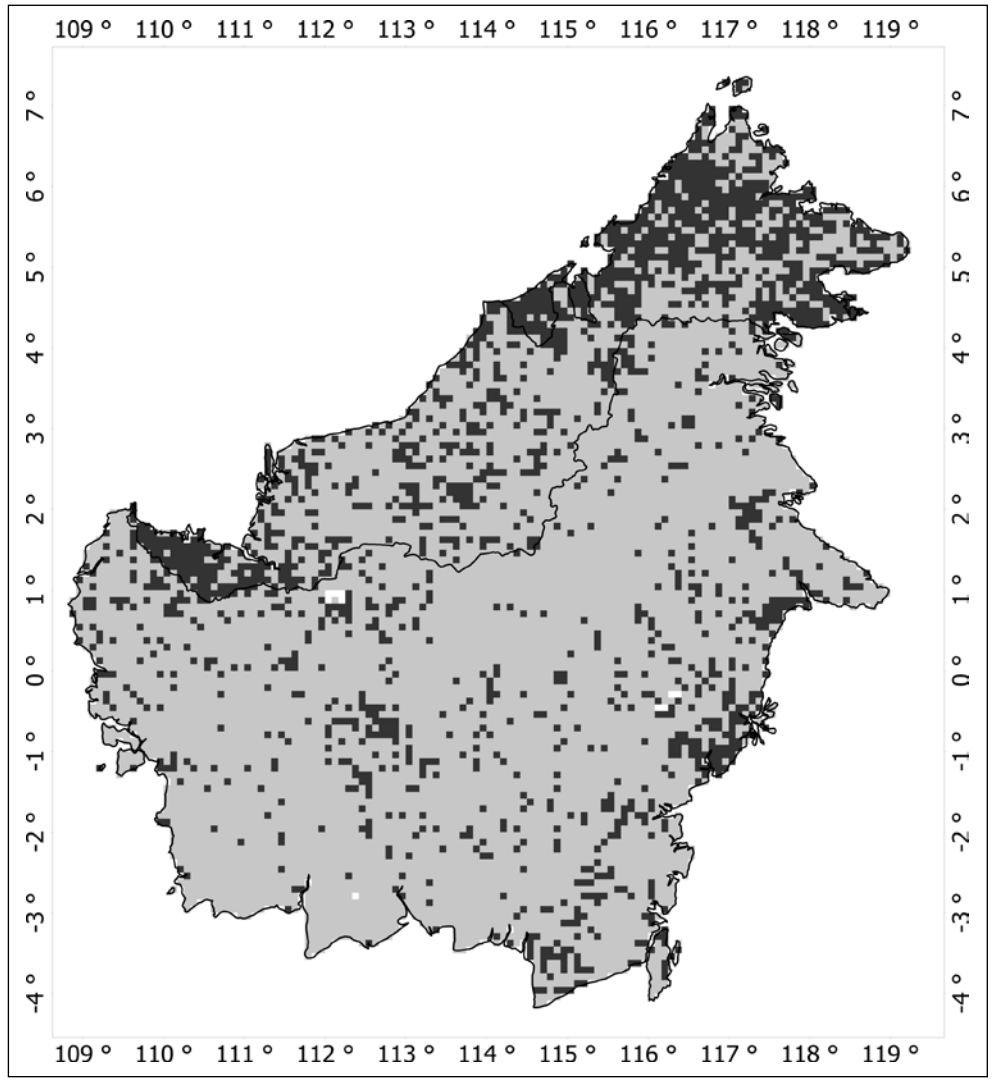

Figure 3.3. Spatial distribution of the 1837 cells, from the 8577 cells for Borneo, where at least one of the 142 097 collections was made (indicated by dark grey squares). Light grey squares indicate the remaining 6740 unsampled cells. White cells indicate large lake areas for which no environmental data were available.

visited by collectors who actually made any collections (Fig. 3.3). The collections are clearly geographically biased, as evident from the geographical distribution of the dark grey squares in Figure 3.3. However, predicting species presences or absences in non-visited areas is one of the major applications of the use of SDMs, so this should not be a major problem. More importantly, it is to assess whether these localities are environmentally biased, or whether certain conditions are over- or under-represented with respect to the environmental conditions for the entire geographical area of Borneo. For this purpose,

we first developed a distribution model of the 1837 collection localities and assessed the model's AUC value. Then, we developed a frequency histogram of expected AUC values on basis of 1837 randomly drawn localities from the 8577 cells of Borneo (100 reps). Unfortunately the AUC value of the distribution model based on the collections localities, is significantly different from random expectation (p<0.01; Fig. 3.2), hence, the collection localities are also environmentally biased.
The implication that collecting effort is environmentally biased for Borneo is that SDMs cannot be tested with null-models drawn randomly from all 8577 grid cells of Borneo.
To overcome this problem we developed a second series of null-models, in the same way as described above, but now randomly drawing from the 1837 known collection locality cells. The resulting 95% C.I. AUC values of these null-models are presented as diamonds in Figure 3.1. Again, we applied a fit through these values to establish the 95% C.I. AUC values against which the SDM AUC values were tested. These values are given in Table S3.1 under '95% C.I. Bias'. Now only 80 of the 116 *Shorea* species (69%) have a SDM AUC value significantly different from a (bias corrected) null-model (Table S3.1, '95% C.I. Bias'; Fig. 3.4a,c). This means that an additional 25 SDMs were rejected, compared to testing against environmentally unbiased null-models.

Discussion

By proposing the use of null-models in the field of presence-only species' distribution modelling, we introduce a novel methodology that allows for significance testing of SDMs. The new methodology makes use of all presence records to develop an SDM and to test its accuracy with the AUC procedure, a threshold- and prevalence-independent single measure of SDM accuracy. A significant SDM indicates that correlations between species' presence localities and the environmental predictor variables, as identified and interpolated by Maxent, deviate from random chance.
Secondly, we show the importance of correcting for environmental biases in data collection. Null models which incorporate the environmental bias within the collection data reject a significant fraction of SDMs which are significant based upon a randomly drawn null-model. If the collection localities are environmentally biased and a species is found throughout the subset of values represented by the collection localities, this species is likely to differ significantly from a null-model which is drawn from the total range of values. This results in an SDM that is an underestimation of the true geographical range of the species. This, because under these conditions the full range of values under which the species truly occurs is not incorporated in the SDM.
Although we introduce a null-model approach to the field of presence-only species' distribution modelling, the use of null-models for significance testing was successfully applied by Olden *et al.* (2002) for presence-absence SDM testing, and by Dolédec *et al.* (2000) in the field of community analysis. Our methodology differs from Olden *et al.* (2002) in that we adapted the null-model approach to make use of presence-only data, and test an SDM accuracy with the threshold- and prevalence independent AUC procedure (Swets, 1988; Manel *et al.*, 2001; McPherson *et al.*, 2004; Guisan *et al.*, 2006). This is important as in our case study the number of species presence records ranged from 5 to 92. Combined with 1,000 pseudo-absences this resulted in prevalence values as low as 0.5 to 8.4%.
We interpret that species, for which the SDM AUC value significantly deviates from a null-model, have specific niche requirements that were met at the localities where they were collected. This agrees with the reasoning of Dolédec *et al.* (2000). They analysed community data with a new multivariate method they called OMI (for Outlying Mean Index), to measure the distance between mean habitat conditions used by a species, and the mean habitat conditions of the sampling area (Dolédec *et al.*, 2000). The OMI value (analogous to the SDM AUC value) of a species is tested against the null-distribution of '1000 random permutation values obtained

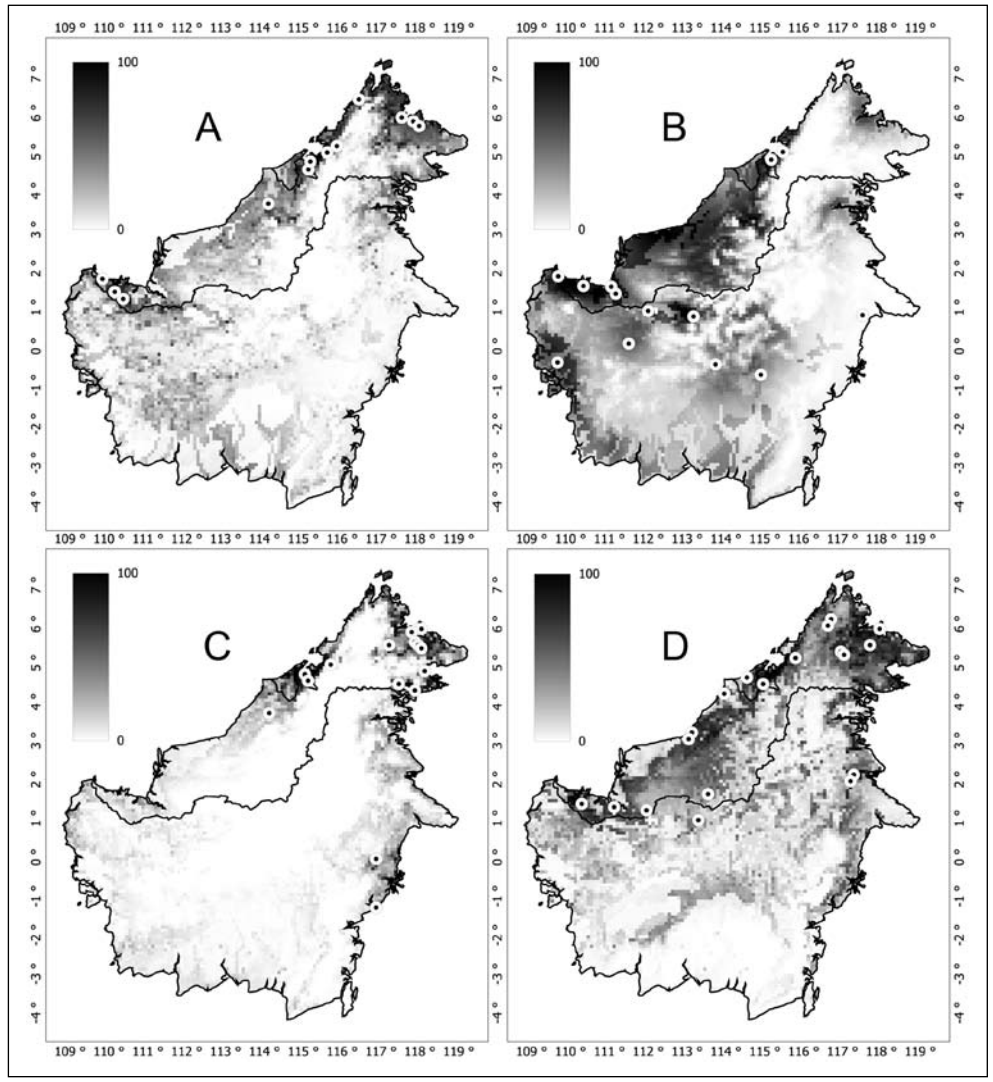

Figure 3.4a-d. Maxent predictions for two significant SDMs (A, C), and two non-significant SDMs (B, D). Collection localities are indicated by dots. A) *Shorea isoptera P.S. Ashton*, (Appendix, Table S3.1, #45), B) *S. platycarpa Heim* (Appendix, Table S3.1, # 49), C) *S. confusa P.S. Ashton* (Appendix, Table S3.1, #57), and D) *S. macroptera Dyer* (Appendix, Table S3.1, #66).

under the null hypothesis that the species is indifferent to its environment'. For species that significantly deviated from this, 'theoretical ubiquitous species that tolerates the most general habitat conditions', it was concluded that the observed species position in habitat differed significantly from what would be expected by chance. This OMI-methodology was later implemented in a species distribution modelling technique called Ecological-Niche Factor Analysis (ENFA) (Hirzel *et al.*, 2002), but testing against a null-distribution was never formalized.

The first to notice that accuracy assessment of presence-only SDMs alone was not sufficient, and SDMs should be tested against a random null hypothesis were Anderson *et al.* (2002). They used the split sample approach, dividing the available presence records of a species in a 75% training and 25% test dataset. After SDM development using the training data, they tested whether test points fell into areas predicted presence more often than expected at random, given the overall proportion of pixels predicted presence vs. predicted absence for that species (Anderson *et al.*, 2002). The latest advances in this methodology were recently made, by introducing a jack-knife (or 'leave-one-out') procedure for SDM accuracy assessment and a combined *p*-value significance test for significance testing of the presence-only SDMs (for details see Pearson *et al.* 2007). However, this methodology does not take into account possible environmental bias in collection localities. If the full niche of a species is not represented by the collection localities, the species' predicted distribution will be smaller than its true distribution. Modelling applications, such as Maxent, are very well capable of predicting the species' distribution based on the available presence records without model under-fitting. A smaller predicted species' distribution automatically results in higher chance of significantly deviating from the random null hypothesis, the same way as in our case study more species significantly deviate from a randomly drawn null-model than from a null-model that is corrected for environmental bias (Fig. 3.1; Table S3.1). Additionally, the jack-knife validation approach may lead to overoptimistic estimates of the predictive power with larger sample sizes (Pearson *et al.*, 2007).

Our results showed the importance of correcting for environmental bias in known collection localities when null-models are used for significance testing of presence-only SDMs. However, at the same time this requirement hampers the general applicability of the methodology. In our case study, we could make use of the full herbarium record database of the National Herbarium of the Netherlands, containing 142,097 georeferenced plant specimen records found in 1837 of the 8577 grid cells of Borneo. We recognize that this amount of data will not always be available. However, since the majority of collections has been made in close proximity to roads, rivers, cities, and nature reserves (Reddy & Davalos, 2003; Kadmon *et al.*, 2004; Hortal *et al.*, 2007), an alternative could be to use a distance buffered road-river map, including cities and nature reserves, to select the grid cells and test these cells for environmental biases. If these cells are environmentally biased, the SDMs can then be tested against a null-model drawn from this pool of cells. However, this approach is less accurate and requires further testing.

Our results showed that for low prevalence values very high AUC values can be expected from randomly drawn points (Fig. 3.1; Table S3.1). Olden *et al.* (2002) too reported such high accuracy values for low (and high) prevalence. The 95% C.I. AUC value of the bias corrected null-model for 15 records (prevalence = 1.48%) was as high as 0.9622 (Fig. 3.1; Table S3.1). Nevertheless, 80 of the 116 *Shorea* species (69%) had an SDM AUC value higher than the 95% C.I. AUC value of the bias corrected null-model. Dolédec *et al.* (2000) reported that, for their application of a null-model for two case studies, 59% and 85% of their species respectively, had significant results. Pearson ***et al.*** (2006) report values from 62-100% depending on the modelling application and thresholds that were used. Our testing against a randomly drawn null model resulted in a comparably high percentage (91%) of significant SDM AUC values ($p<0.05$). All these results are higher than the 50% reported by Olden *et al.* (2002).

Both the AUC values of the two null-models, and the SDMs, show a decreasing trend with increasing number of records (Fig. 3.1; Table S3.1). This is most likely the result of applying ROC plots to SDMs, developed with presence-only data, reducing the maximum AUC value dependent on the species' true distribution (Phillips *et al.*, 2006). Assuming that the predicted species' distributions are a good proxy for the species' true distributions, we assessed the area for which species were predicted to be present by converting the continuous probabilistic Maxent predictions of occurrence to discrete presence-absence values. We used the maximized sensitivity-specificity sum threshold for this purpose (Liu *et al.*, 2005). Regressing significant SDMs AUC values against the area for which they were predicted to be present (Table S3.1; 'Area (in %)') revealed a significant negative linear correlation (AUC=0.9913-0.0029*Area; $p<0.001$; $R^2 = 0.576$). We consider this as a strong indication that it is not the accuracy of the models that is reduced but merely that the maximum achievable AUC value is reduced due to an increased true distribution of the species. We therefore do not support the statement that the predictive accuracy of the model decreases when the extent of a species distribution increases, as suggested by Hernandez *et al.* (2006). When an increased predicted distribution and related lower SDM AUC value is caused by a broad niche amplitude, however, as is the case for habitat generalists, an SDM accuracy is more likely not to deviate from a null-model and the SDM can therefore not be used. This is possibly the case for the SDMs presented in Figure 3.4b,d (Table S3.1; #49 and #66).

A consequence of implementing the proposed use of null-models for SDM evaluation, is that SDM accuracy is tested with the same data used to develop models, i.e., a form of model verification (Araújo & Guisan, 2006). A problem with this approach is that SDMs may over-fit the calibration, or training data (Araújo *et al.*, 2005). Over-fitting, however, is not considered a problem if the goal is to describe a pattern and simultaneously reduce false negatives: i.e., true observations that are not predicted by the model (Araújo & Guisan, 2006). An advantage is that all observations are used to develop the SDMs, making optimal use of all available information. If the modelled species' distributions are intended to be used for conservation planning, verification is an approved method to test whether an SDM performs as intended. However, if the models are used to predict range shifts under different climate change scenario's, or to assess the possible invasiveness of a species, an SDM's ability to correctly predict independent test data is preferred (Araújo & Guisan, 2006). It should be kept in mind, however, that SDMs, as they are applied in this study, predict the potential distribution of a species and do not take into account competition, and historical or present geographical barriers (Soberón & Peterson, 2005; Peterson, 2006). Most studies addressing these issues use data partitioning methods to allocate records to training and test datasets. The most familiar technique is one-time data-splitting (Araújo *et al.*, 2005). Our unpublished

results indicated, however, that dependent on the spatial distribution and the random split of the records, SDM accuracies could be very different.

An advantage of Maxent is its ability to counteract the tendency of SDMs to over-fit when few presence records are available, due to its regularization procedure (Hernandez *et al.*, 2006; Phillips *et al.*, 2006). Therefore, we used the standard settings of Maxent. However, the null-models developed for 80, 90, and 100 records developed with the modelling rules set to 'auto features' and the regularization multiplier set to 1, resulted in increasing 95% C.I. AUC values indicating over-fitting of the models (data not shown). For this reason, we set the modelling rules to use linear, quadratic and hinge functions to develop the null-distributions for those numbers of records and the SDMs developed with more than 79 records.

We are aware that spatial autocorrelation in the distribution of the species records and environmental variables may also influence SDM accuracy. Our intention was not to investigate the influence of spatial autocorrelation on SDM accuracy, however, but to provide a methodology for significance testing of presence-only SDMs. Simultaneously we showed that the evaluation of presence-only SDM quality based on subjective ROC plot thresholds (e.g. AUC≥0.7 = useful), cannot be applied. With this contribution, we hope to provide SDM users with a valuable tool to identify those species that can be accurately modelled, while providing an additional reason for being cautious about interpretations of SDMs that are not tested for significance.

Acknowledgements – We would like to thank Steven Phillips for adapting the Maxent application several times to allow us to perform the ROC analyses of the null-models. We also thank Pieter Baas, Chuck Cannon, Peter Hovenkamp, Marco Roos, Ferry Slik and four anonymous referees, all of whose comments allowed us to make useful improvements to the manuscript.