

# Null-model for Species' Distribution Modelling with Presence-only Data

Niels Raes and Hans ter Steege



National Herbarium of the Netherlands – Leiden University branch,  
the Netherlands, raes@nhn.leidenuniv.nl

## Introduction

Species' distribution models attempt to relate species' presence data to environmental predictors. Greatest challenge in species' distribution modelling is the assessment of model predictive accuracy. Model accuracy is a measure of how well test data are predicted by the model build with training data. To obtain independent test data, all available unique presence records of a species are partitioned in train/test data. Statistical tests assessing model accuracy require absence data. When absence data are lacking, these are replaced by pseudo-absences. Pseudo-absences are needed in a sufficient large sample to represent the environmental variation of the area studied. This results in a low sampling prevalence, or proportion representing presence, since the number of presence records of a species is often less than 50 and pseudo-absences are represented by 1,000-10,000 localities. Of all measures of accuracy used in species' distribution modelling, the Area Under the Receiver Operating Characteristic Curve (AUC) is the only measure invariant to low sampling prevalence (McPherson, Jetz and Rogers 2004) and threshold independent. AUC values range from 0 to 1, with a value of 0.5 indicating model accuracy no better than random, and 1 indicating perfect model fit. However, when applied with presence-only data and pseudo-absences the maximum AUC value is no longer 1 but  $1-a/2$ , where  $a$  represents a species' true distribution (Phillips, Anderson and Shapire 2006). Consequently standard measures of accuracy using AUC values no longer apply, since a species' true distribution is not known. Further, it was shown that differently partitioned data have different model accuracies (Phillips, Anderson and Shapire 2006).

## Research Aims

1. Introduce a null-model to assess whether a species' distribution model is better than can be expected by chance alone.
2. Introduce a new measure of model accuracy as the percentage of differently partitioned models performing better than the null-model.

## Discussion / Conclusions

- *In the absence of a maximum AUC value, null-models are a reliable measure of model accuracy.*
- *Results show the influence of data partitioning on model performance, single model runs can result in very high and very low model accuracies.*
- 58% (55/95) of the *Shorea* species can be modelled with reliable or high accuracy. This is a slightly better result than the less than 50% accurately modelled fish distributions based on presence-absence data and tested with a null-model (Olden, Jackson and Peres-Neto 2002).
- The larger range of AUC values at low number of presences is conform larger standard errors of the mean AUC at low and high prevalence using presence/absence data (McPherson, Jetz and Rogers 2004).

## Materials and Methods

**Species data:** All unique *Shorea* collections of the NHN-Leiden database of Borneo with more than 8 collections, 95 species.  
Pseudo-absences: 1,000.

**Environmental predictors:** 20 WorldClim variables, 15 FAO soil properties data, and 3 additional ones (38 in total) at 5 arc minute resolution (in total 8614 cells).

**Software:** Maxent version 2.1; maximum entropy method for modelling species' distributions with presence-only data (Phillips, Anderson and Shapire 2006).

### Null-model

- Draw 999 times 8, 16, 32, 64, 128, and 256 random points.
- Partition data 50/50 in train/test data.
- Model random points (999 x 6 models) and assess the AUC value of the test data.
- Establish the 95% confidence interval for each number of random points.
- Apply a curve fit through the upper 95% C.I. limit.

### Model accuracy

- Randomly partition the presence records of each species 100 times (95 x 100).
- Model the differently partitioned data.
- Assess which percentage of the models have a higher test AUC value than the null-model for the corresponding number of records and assess model accuracy according table.
- Finally predicted distributions should be modelled using all presence localities for those species which can be modelled with reliable or high accuracy.

Table. Model accuracy expressed as percentage of differently partitioned models with an AUC value above the 95% C.I. upper limit of the null-model.

% > null-model	Accuracy
< 70	Low
70-90	Reliable
≥ 90	High

## Results

The upper limit of the 95% C.I. values of the null-model showed perfect fit with the negative power function  $y = 0.4051 x^{-0.4915}$  ( $p < 0.001$ ;  $R^2 = 0.9994$ ) (red striped lines figure).

### Model Accuracy:

- 32 high (red dots)
- 23 reliable (black dots)
- 45 low (grey crosses)

### References:

- McPherson, J. M., Jetz, W. and Rogers, D. J. 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? – *J. Appl. Ecology* 41: 811-823.
- Olden, J. D., Jackson, D. A. and Peres-Neto, P. R. 2002. Predictive Models of Fish Species Distributions: A Note on Proper Validation and Chance Predictions. – *Transactions of the American Fisheries Society* 131: 329–336.
- Phillips, S. J., Anderson, R. P. and Schapire, R. E. 2006. Maximum entropy modeling of species geographic distributions. – *Ecological Modelling* 190: 231-259.

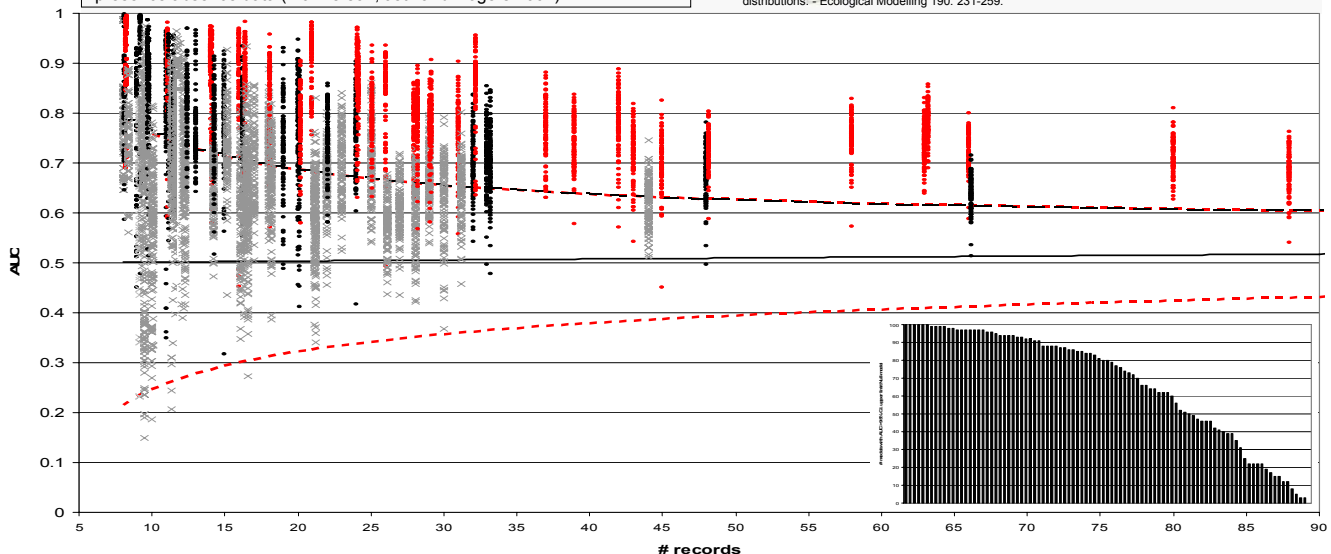


Figure. Results of the null-model and randomized partitioned data of all *Shorea* species of Borneo with more than 8 unique records. Red striped lines represent the 95% confidence interval (C.I.) limits of the null-model. The black solid line is the average AUC of the null-model. Red dots show the AUC values of 100 times random partitioned data for species that have more than 90% of their models above the upper limit of the 95% C.I. of the null-model. Black dots show the AUC values of 100 random partitioned data for species that have 70-90% of their models above the upper limit of the 95% C.I. of the null-model. Grey crosses show species with less than 70% of their models above the upper limit of the 95% C.I. of the null-model. Species with the same number of records are shown next to each other. The black striped line shows the value of the 95% C.I. upper limit for the corresponding number of records. Inset: % of models above the null model of all species in decreasing order.