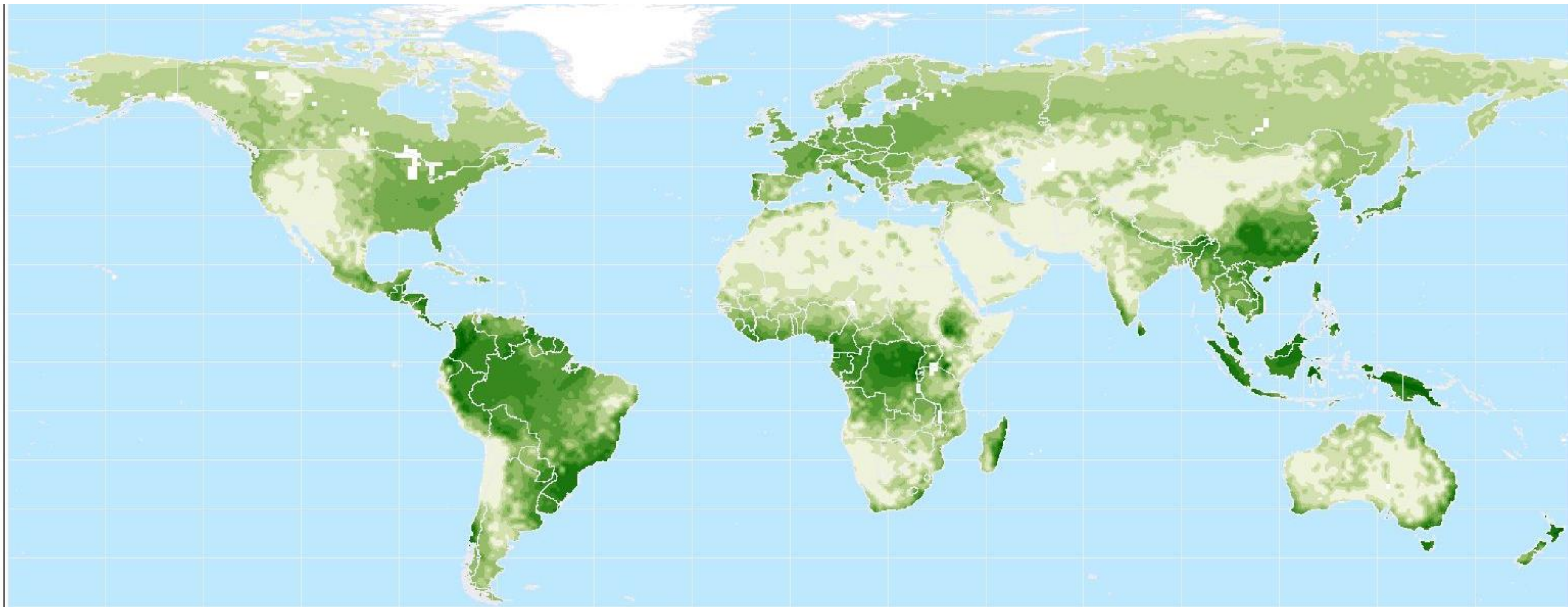
european social fund in the czech republic
EUROPEAN UNION
MINISTRY OF EDUCATION, YOUTH AND SPORTS
OP Education for Competitivness
INVESTMENTS IN EDUCATION DEVELOPMENT

Mapping and modeling species distributions

Department of Botany and Zoology, Masaryk University
Bi9661 Selected issues in Ecology, Autumn 2013
Borja Jiménez-Alfaro, PhD

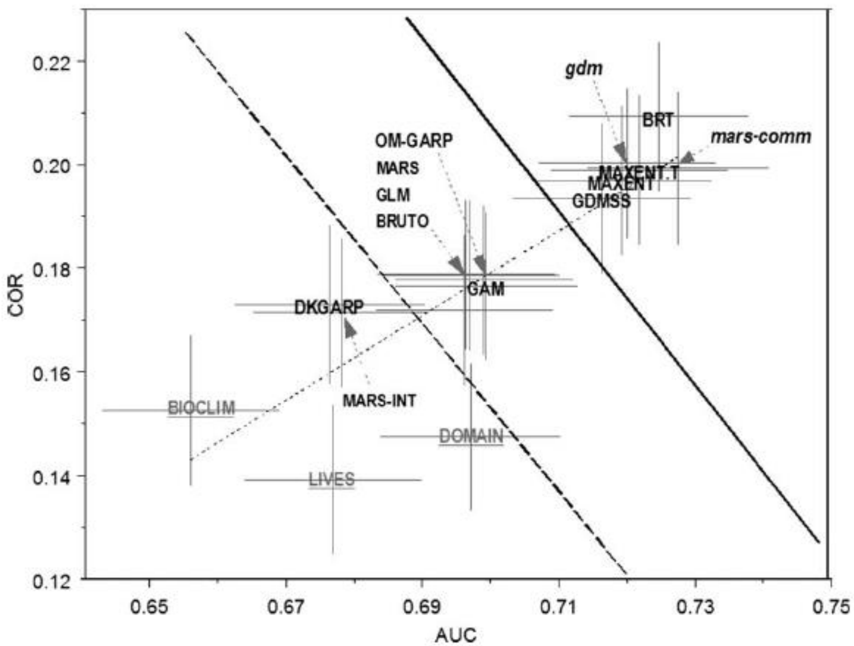


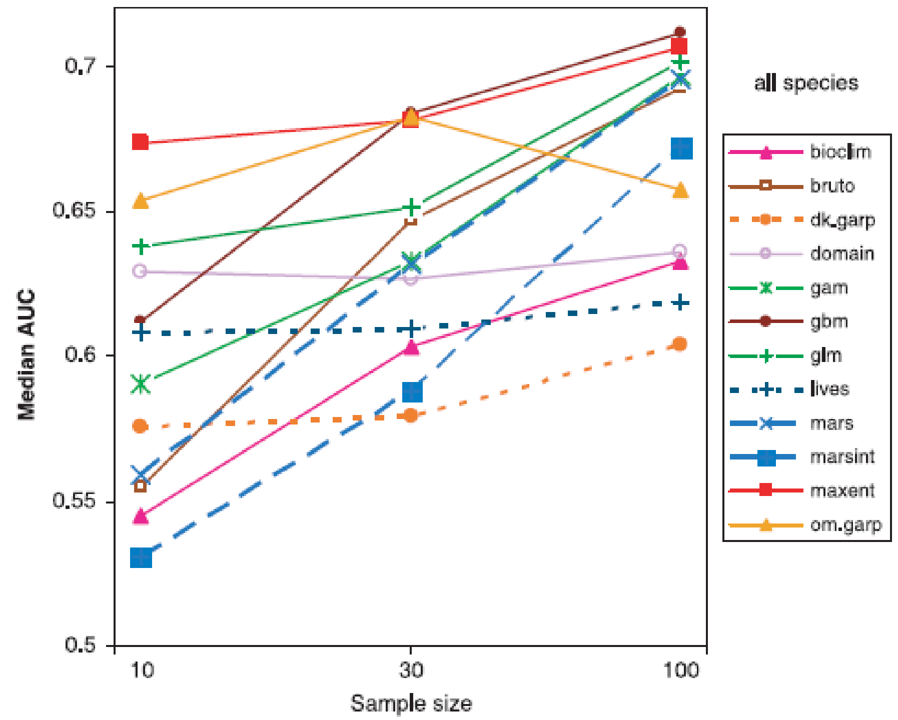
Part 3:

MAPPING + MODELING

Model evaluation and implementation

The question: how to estimate model accuracy?

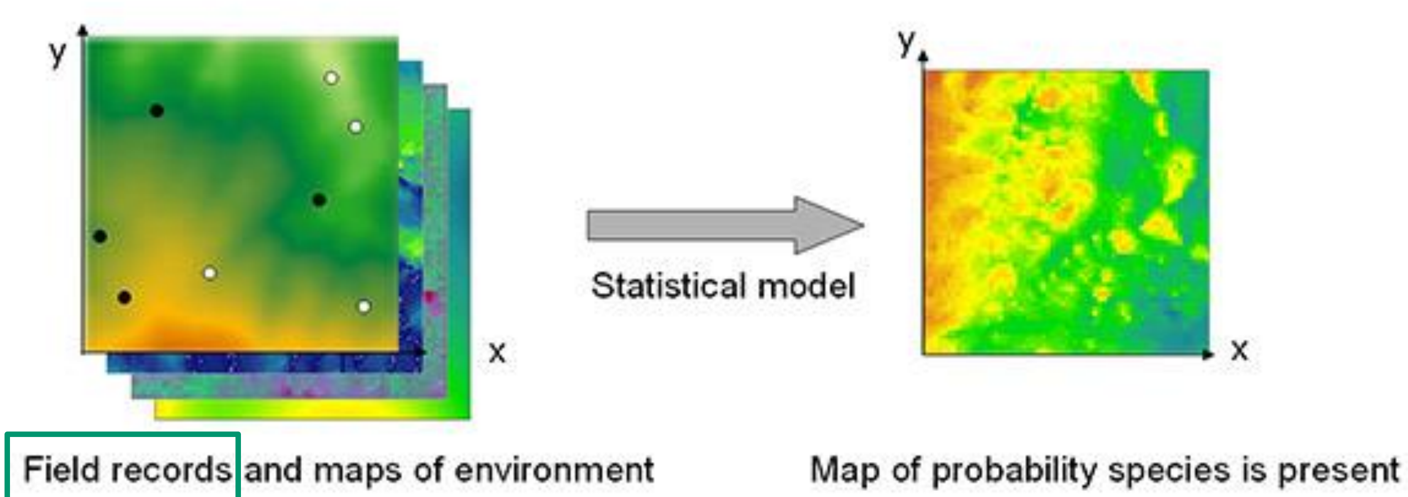




Data preparation

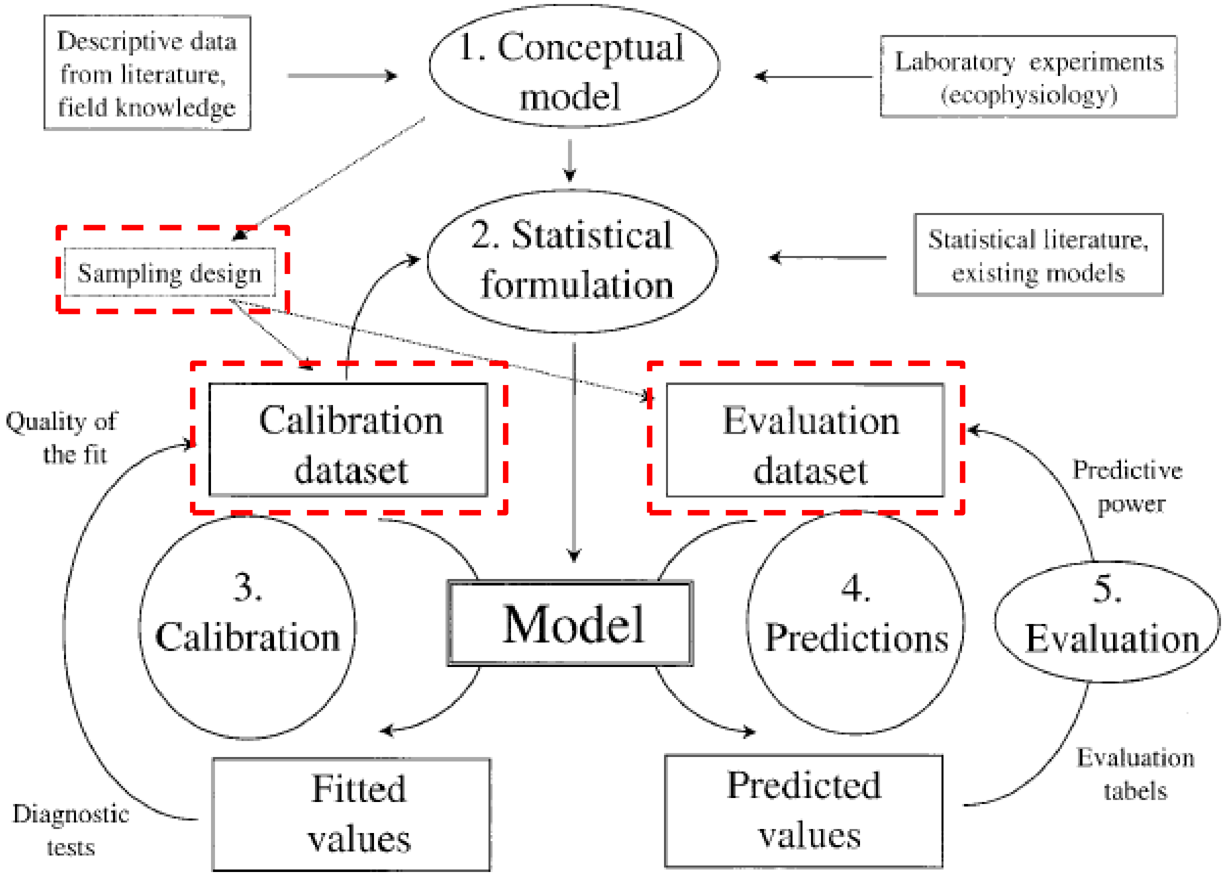
How did you organize your modeling project?

Did you think about model evaluation when sampling?



Main atributes: quantity and quality

Calibration versus Evaluation dataset



From Guisan and Zimmerman 2000

Option A – INDEPENDENT DATA

You should test your model using completely different data

- Using alternative data from different sources

- Or a new sampling design to collect NEW data

- Thus you will have **training** data for calibration
  testing data for evaluation

Option B – DATA PARTITION

When option A is not posible, a common procedure is to separate a subset of your own data for validation (although sampled in a similar way)

- You will have again **training** data and **testing** data

- Common procedure is to separate 80% of occurrences for training and 20% for testing

- For only two predictors, a ratio of 50/50 is recommended

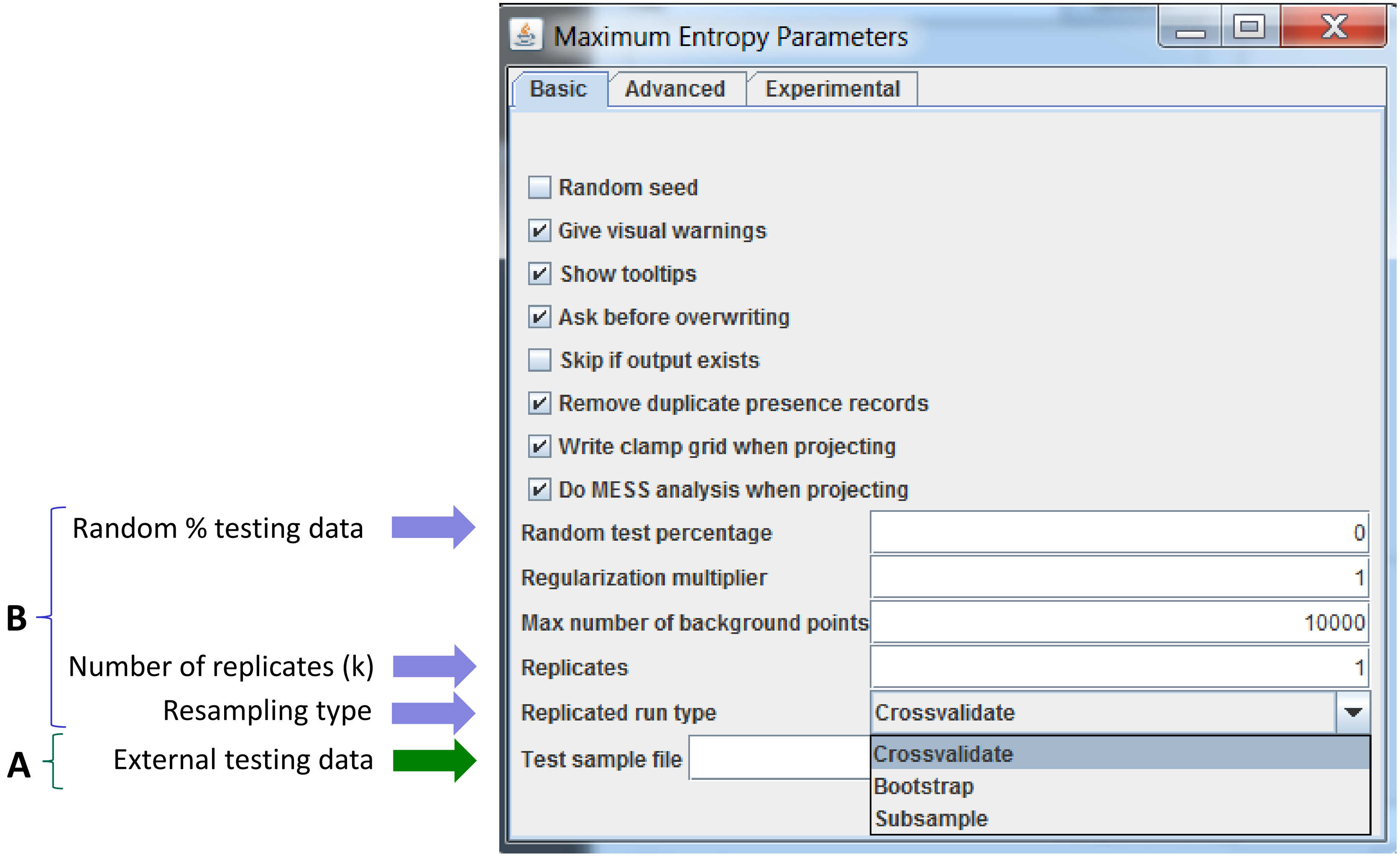
Option B – DATA PARTITION

With few samples, you can apply general techniques:

K-fold crossvalidation (leave-one out)
(if k = 10) you split the data into 10 subsets, and compute 10 models using 9 subsets for training and 1 for calibration. You can then average the models and the validation statistics

Bootstrap sampling
You can compute multiple models using a random selection of occurrences (sampling with replacement) to estimate prediction accuracy

For example, in MaxEnt



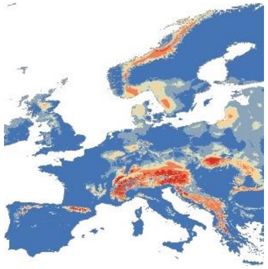
The properties of model evaluation

Training data

Model predictions

a) Categorical (1/0)

b) Probabilistic (0.01……1)



Testing data

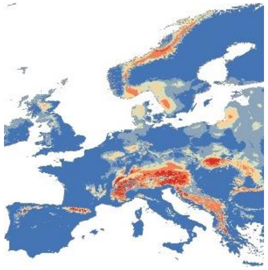
Categorical (1/0)
(i.e. presences/absences)

Measures of accuracy (= model performance)

For categorical models:

Threshold-dependent measures (e.g. KAPPA)

(you define a threshold between suitable/unsuitable)



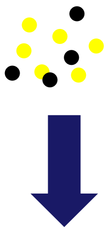
For probabilistic models:

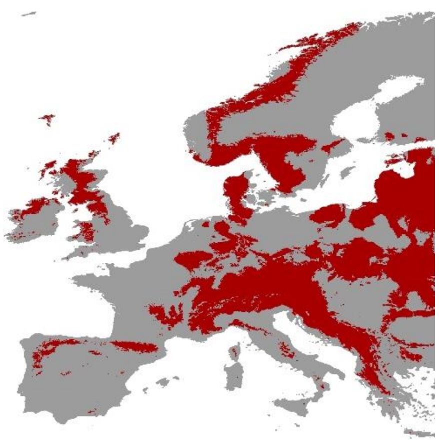
Threshold-independent measures (e.g. AUC)

(you assess the complete range of probabilities)

Threshold-dependent measures

The confusion (error) matrix



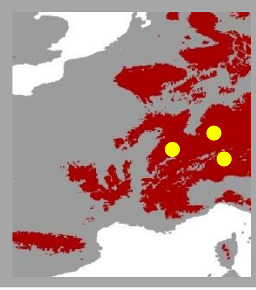


		TESTING DATA	
		(1) Presence	(0) Absence
THE MODEL	(1) Presence	*n*	*n*
	(0) Absence	*n*	*n*

Threshold-dependent measures

The confusion matrix (error) matrix

TESTING DATA

		(1) Presence	(0) Absence
THE MODEL	(1) Presence		
	(0) Absence		

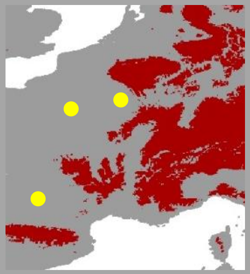
Threshold-dependent measures

The confusion (error) matrix

		TESTING DATA	
		(1) Presence	(0) Absence
THE MODEL	(1) Presence		
	(0) Absence		

Threshold-dependent measures

The confusion (error) matrix

		TESTING DATA	
		(1) Presence	(0) Absence
THE MODEL	(1) Presence		
	(0) Absence		

Threshold-dependent measures

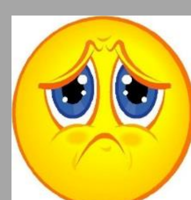
The confusion (error) matrix

		TESTING DATA	
		(1) Presence	(0) Absence
THE MODEL	(1) Presence		(commission error) 
	(0) Absence	(omission error) 	

Threshold-dependent measures

The confusion (error) matrix

TESTING DATA

THE MODEL		(1) Presence	(0) Absence	
	(1) Presence			→ **Sensitivity** (% true positives)
	(0) Absence			→ **Specificity** (% true negatives)

Evaluating models

Table 9.3. *Threshold-dependent accuracy measures for species presence–absence models based on the error matrix, where n is the total number of observations used for validation; $n = TP + TN + FP + FN$ (Table 9.2). These accuracy measures can be calculated for any probability threshold used to define categorical predictions, except for the true skill statistic which, by definition, is based on the probability threshold for which the sum of sensitivity and specificity is maximized*

Measure	Calculation
Sensitivity	TP/(TP + FN)
False negative rate	1 − Sensitivity
Specificity	TN/(TN + FP)
False positive rate	1 − Specificity
Percent correct classification	(TP + TN)/n
Positive predictive power	TP/(TP + FP)
Odds ratio	(TP × TN)/(FP × FN)
Kappa	$\dfrac{[(TP+TN)-(((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/n)]}{[n-(((TP+FN)(TP+FP)+(FP+TN)(FN+TN))/n)]}$
True skill statistic	1 − maximum (Sensitivity + Specificity)

From Franklin 2009

Evaluating models

Most common measures of accuracy for categorical models:

KAPPA (from 0 to 1)
Pros Widely recognized measure of agreement for categorical data
Cons In some cases is sensitive to prevalence of the data
(better to be used when prevalence is c. 50%)

TRUE STILL STATISTIC (TSS) (from -1 to +1)
Pros An alternative to Kappa, less senstitive to prevalence
Cons Sometimes it can be negatively related to prevalence

MODEL EVALUATION

An example of using Kappa for model evaluation

Diversity and Distributions, (Diversity Distrib.) (2007) 13, 397–405

BIODIVERSITY RESEARCH

A comparative evaluation of presence-only methods for modelling species distribution

Asaf Tsoar*, Omri Allouche, Ofer Steinitz, Dotan Rotem and Ronen Kadmon

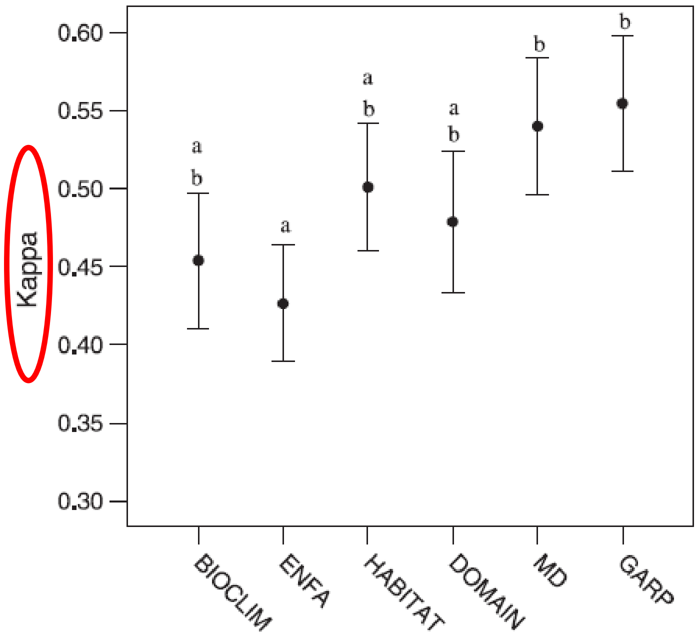


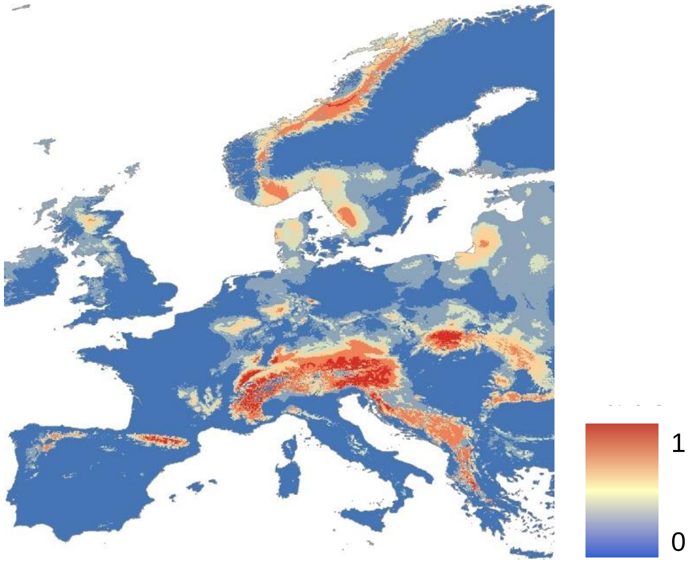
Figure 2 Differences in Kappa among modelling methods (BIOCLIM, HABITAT, GARP, ENFA, DOMAIN, and MD) when data for all taxa (snails, birds, and bats) are pooled. Error bars represent mean ± 1 standard error. Models sharing the same letters do not differ from each other significantly ($P > 0.05$ following Bonferroni corrections for multiple comparisons).

Threshold-independent measures

Are based on continuous probabilistic outputs

Are independent of the prevalence

Useful for comparing the accuracy of different models
(e.g. with different frequencies and prevalences)



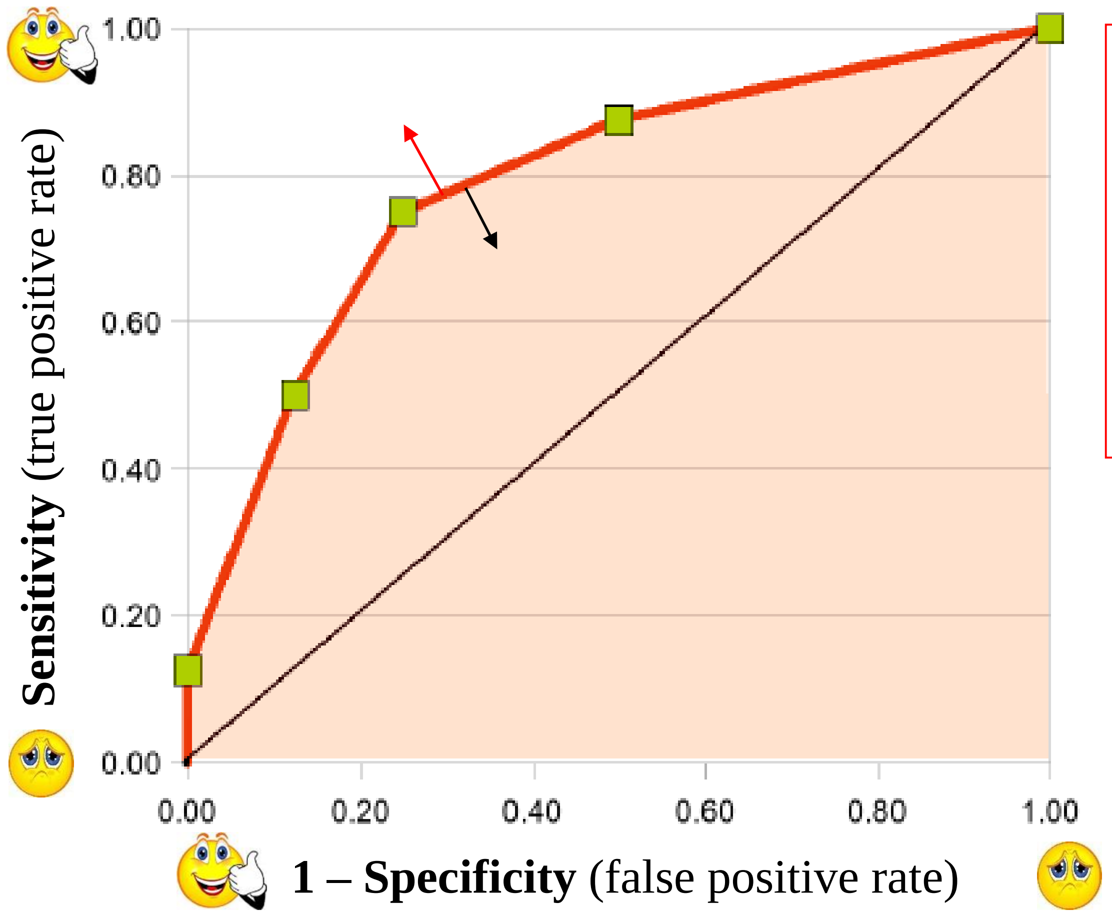
The ROC plot

(ROC = Receiving Operating Characteristic)

Values referred to different prob. thresholds (0, 0.1… 1)

AUC (Area under the Curve) of the ROC plot

Prob. that a random selection classify > suitability for presence than for absence



Model performance
(AUC values)

0.9 - 1.0: very good
0.8 - 0.9: good
0.7 - 0.8: moderate
0.6 - 0.7: low
0.5 - 0.6: very low

The ROC space

A (good)

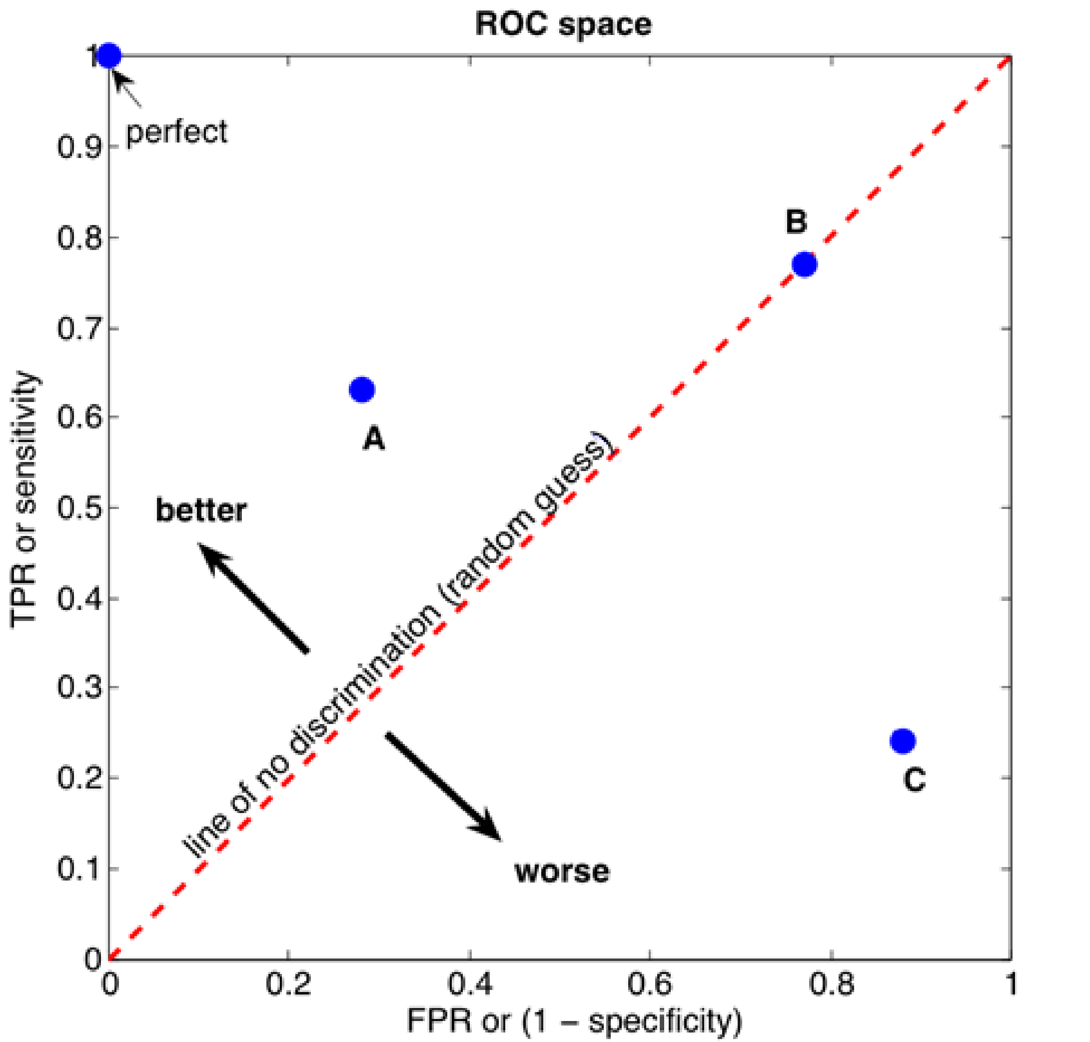
	P=100	N=100
P=91	TP=63	FP=28
N=109	FN=37	TN=72

B (random)

	P=100	N=100
P=154	TP=77	FP=77
N=46	FN=23	TN=23

C (bad)

	P=100	N=100
P=112	TP=24	FP=88
N=88	FN=76	TN=12



What happens with presence-only methods?

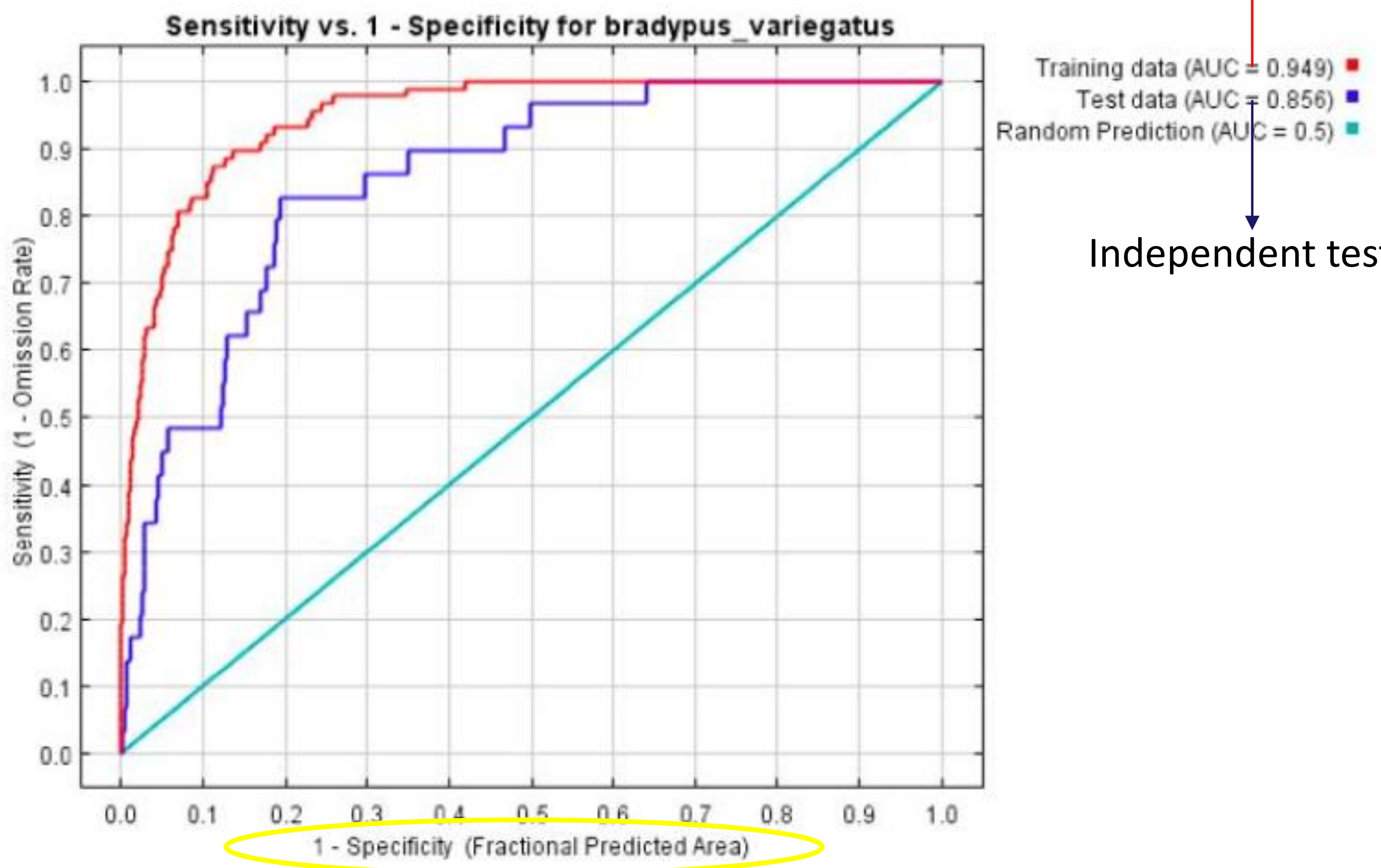
Only presences means only **sensitivity**
It is necessary to use pseudo-absences or background data

In Maxent:
(1 – specificity) or commission error….
…is substituted by the fraction of the study area predicted as presence

MODEL EVALUATION

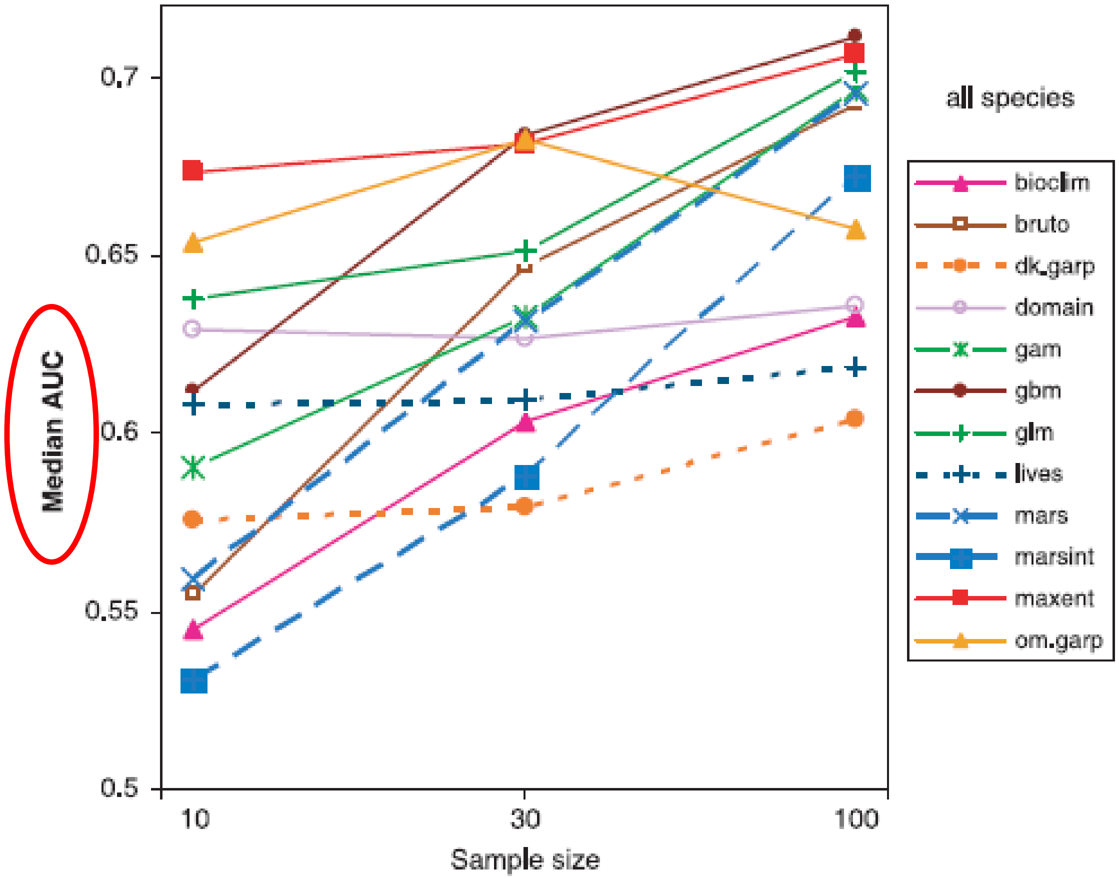
All presences predicted as presence

Probably over fitted



Independent testing data

AUC is widely used for assesing model performance



Probability thresholds

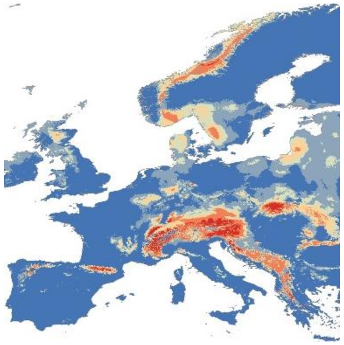
Thresholds are necessary for:

- Obtaining categorical models (presence/absence)

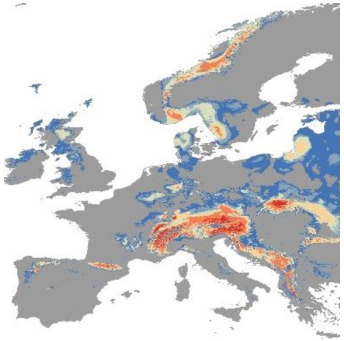
- Comparing model performance (Kappa, TSS, etc)

- Documenting model outputs (suitable areas for a species)

Probability thresholds



Without threshold (from 0 to 1)



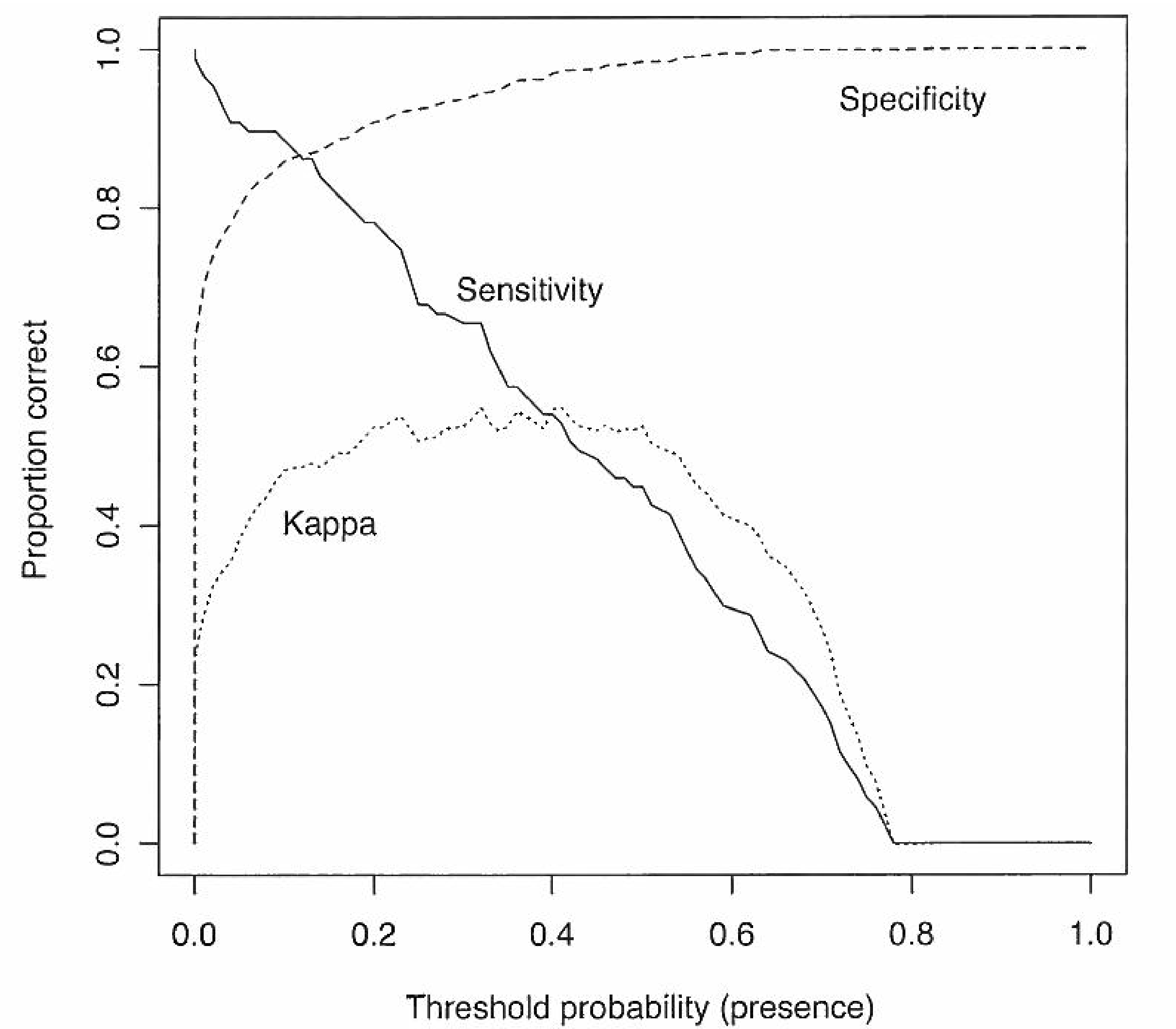
Minimum threshold (from 0.17 to 1)

Threshold 0.17 for binary output (0 or 1)

TABLE 7.1. Some published methods for setting thresholds of occurrence, to convert continuous or ordinal model output to binary predictions of "present" and "absent."

Method	*Definition*	*Occurrence data type*[b]	*Example reference(s)*
Fixed value	An arbitrary fixed value (e.g., probability = 0.5).	None needed	Manel et al. 1999; Robertson et al. 2004
Least training presence	The lowest predicted value corresponding to an occurrence record.	Presence-only	Pearson et al. 2007; Phillips et al. 2006
Fixed sensitivity[a]	The threshold at which an arbitrary fixed sensitivity is reached (e.g., 0.95, meaning that 95% of calibration occurrence localities will be included in the prediction).	Presence-only	Pearson et al. 2004
Sensitivity-specificity[a] equality	The threshold at which sensitivity and specificity are equal.	Presence/absence	Pearson et al. 2004
Sensitivity-specificity sum maximization	The sum of sensitivity and specificity is maximized.	Presence/absence	Manel et al. 2001
Maximize Kappa[a]	The threshold at which Cohen's Kappa statistic is maximized.	Presence/absence	Huntley et al. 1995
Average probability/ suitability	The mean value across model output.	None needed	Cramer 2003
Equal prevalence	Species' prevalence (the proportion of presences relative to the number of sites) is maintained the same in the prediction as in the calibration data.	Presence only	Cramer 2003

From Peterson et al. 2011



From Franklin 2009

For example, in MaxEnt

Some common thresholds and corresponding omission rates are as follows. If test data are available, binomial probabilities are calculated exactly if the number of test samples is at most 25, otherwise using a normal approximation to the binomial. These are 1-sided p-values for the null hypothesis that test points are predicted no better than by a random prediction with the same fractional predicted area. The "Balance" threshold minimizes 6 * training omission rate + .04 * cumulative threshold + 1.6 * fractional predicted area.

Cumulative threshold	Logistic threshold	Description	Fractional predicted area	Training omission rate
1.000	0.025	Fixed cumulative value 1	0.557	0.006
5.000	0.105	Fixed cumulative value 5	0.369	0.041
10.000	0.170	Fixed cumulative value 10	0.283	0.061
0.280	0.011	Minimum training presence	0.665	0.000
19.872	0.257	10 percentile training presence	0.182	0.100
26.446	0.327	Equal training sensitivity and specificity	0.137	0.137
23.444	0.292	Maximum training sensitivity plus specificity	0.156	0.112
1.935	0.043	Balance training omission, predicted area and threshold value	0.480	0.011
10.554	0.176	Equate entropy of thresholded and original distributions	0.276	0.061