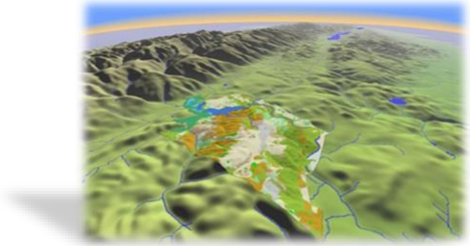
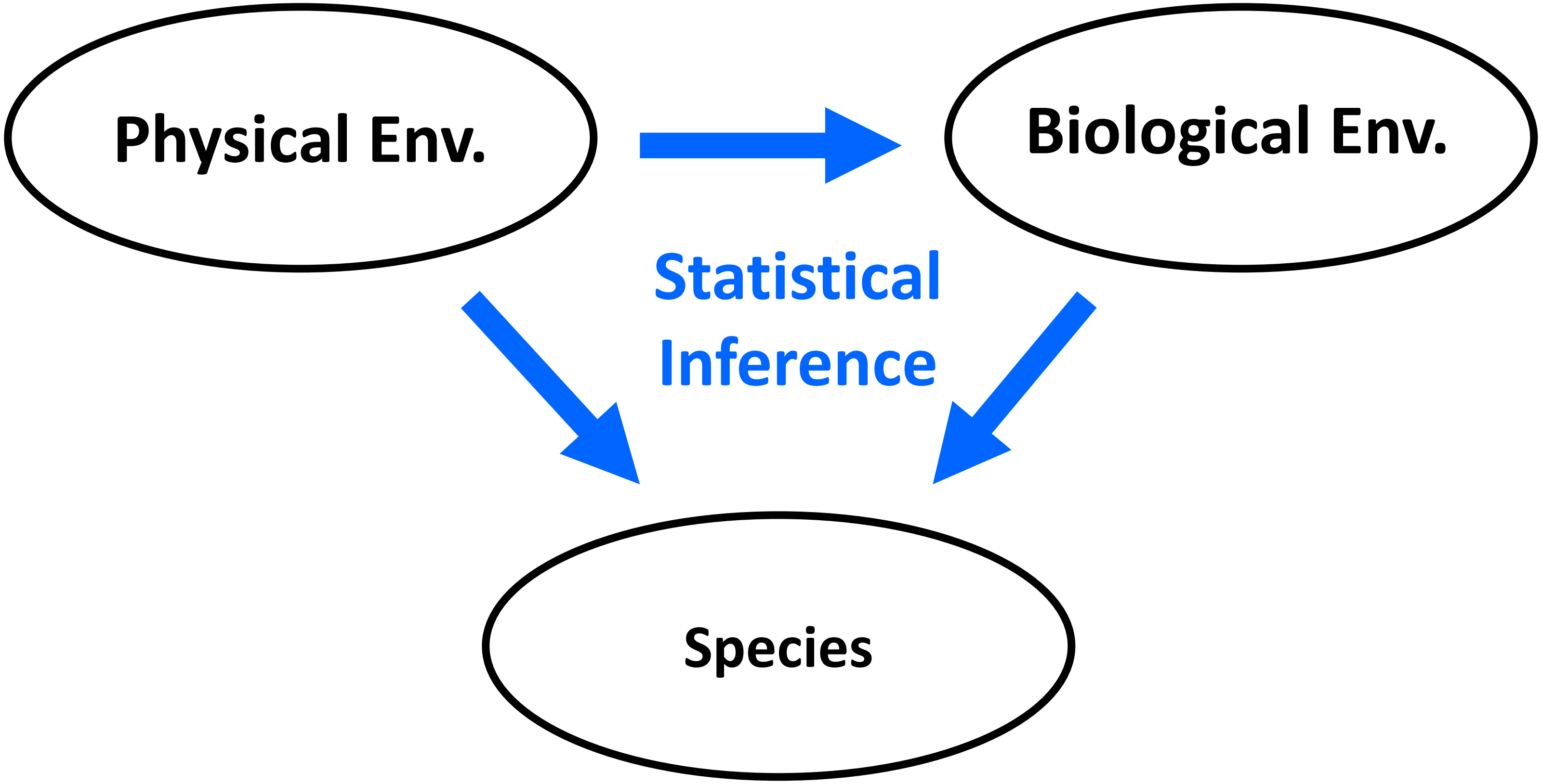
NICHOLAS SCHOOL OF THE
ENVIRONMENT AND EARTH SCIENCES
DUKE UNIVERSITY



ENVIRON 761:
Species Distribution Modeling - Model evaluation

Instructor: John Fay

Species distribution modeling



Physical Env.

Biological Env.

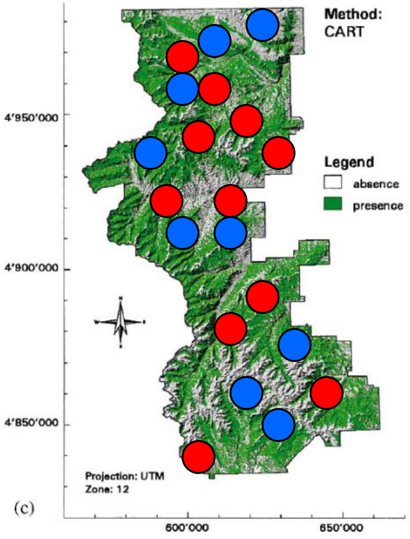
Statistical Inference

Species

Inductive modeling

GIS Applications

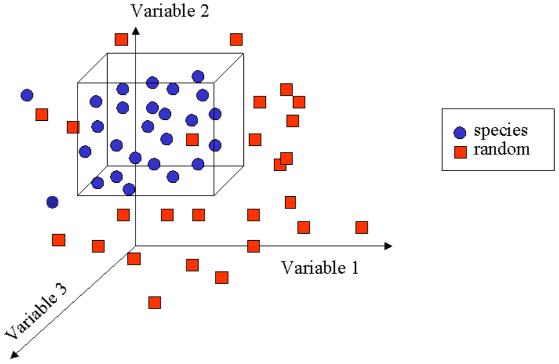
Geographic Space



Sample Data

Statistical Applications Data Space

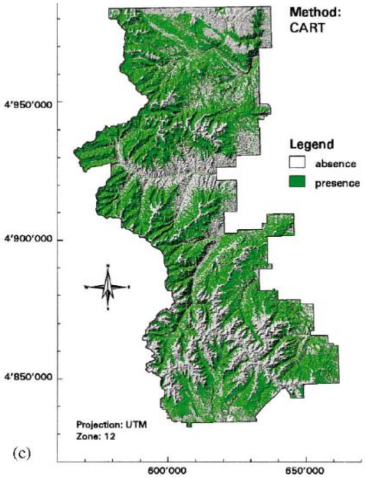
Species ~ Variable 1 + Variable 2 + Variable 3



Refine Model

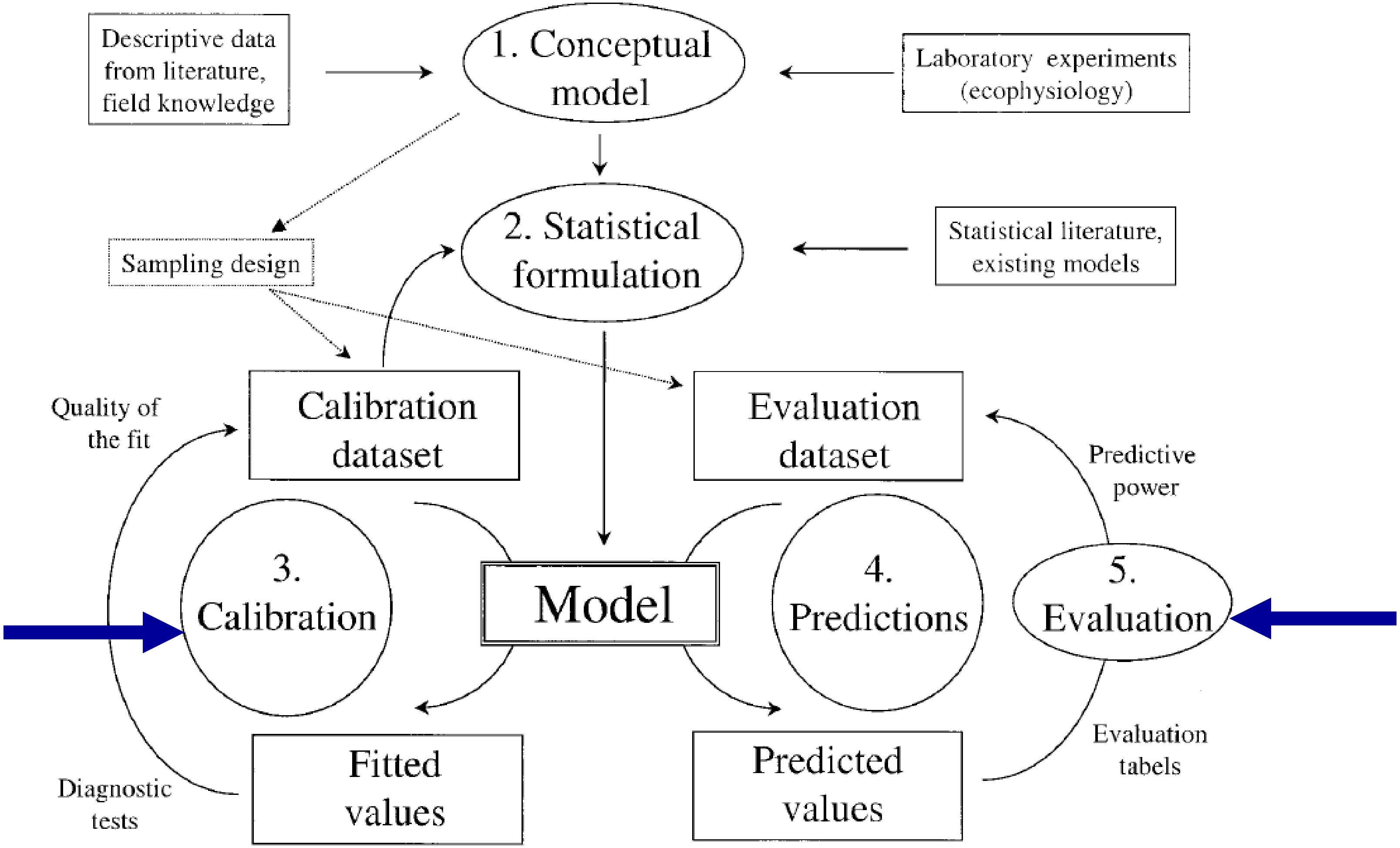
GIS Applications

Geographic Space



Model Habitat

Modeling process

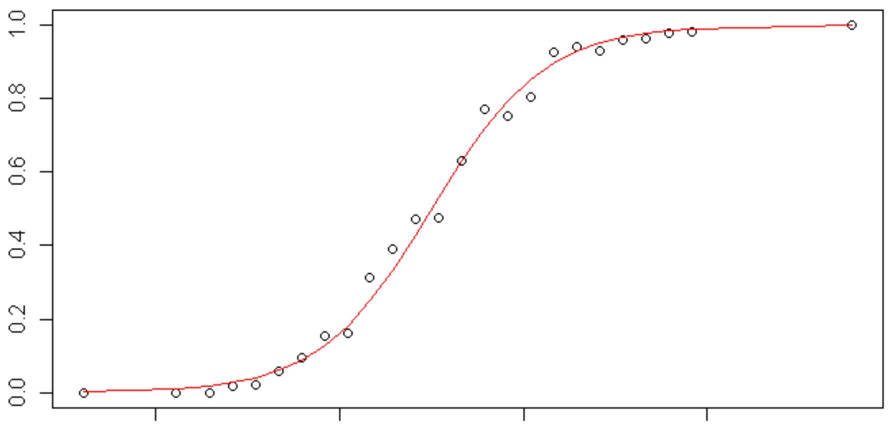


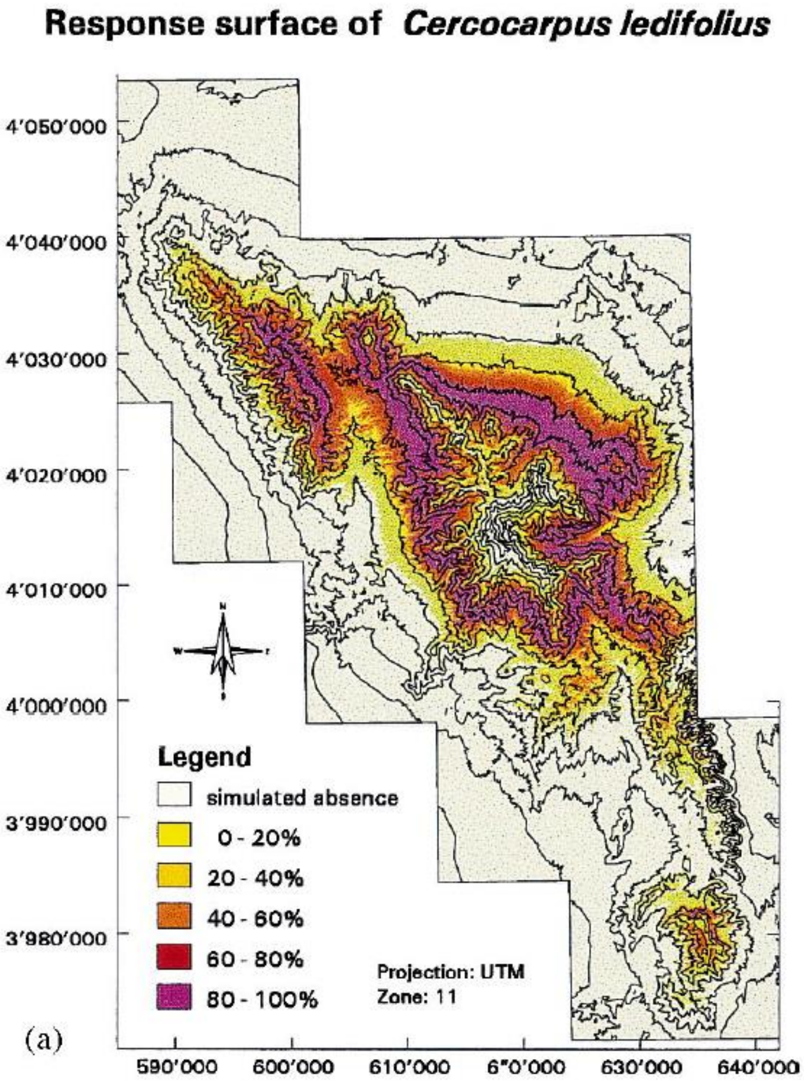
From: Guisan and Zimmermann 2000

Model examples

Probability of Occurrence

From logistic GLM's



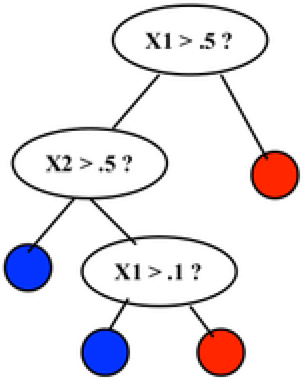

Response surface of *Cercocarpus ledifolius*

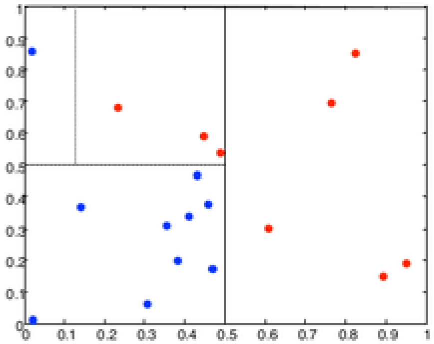
From: Guisan and Zimmermann 2000

Model examples

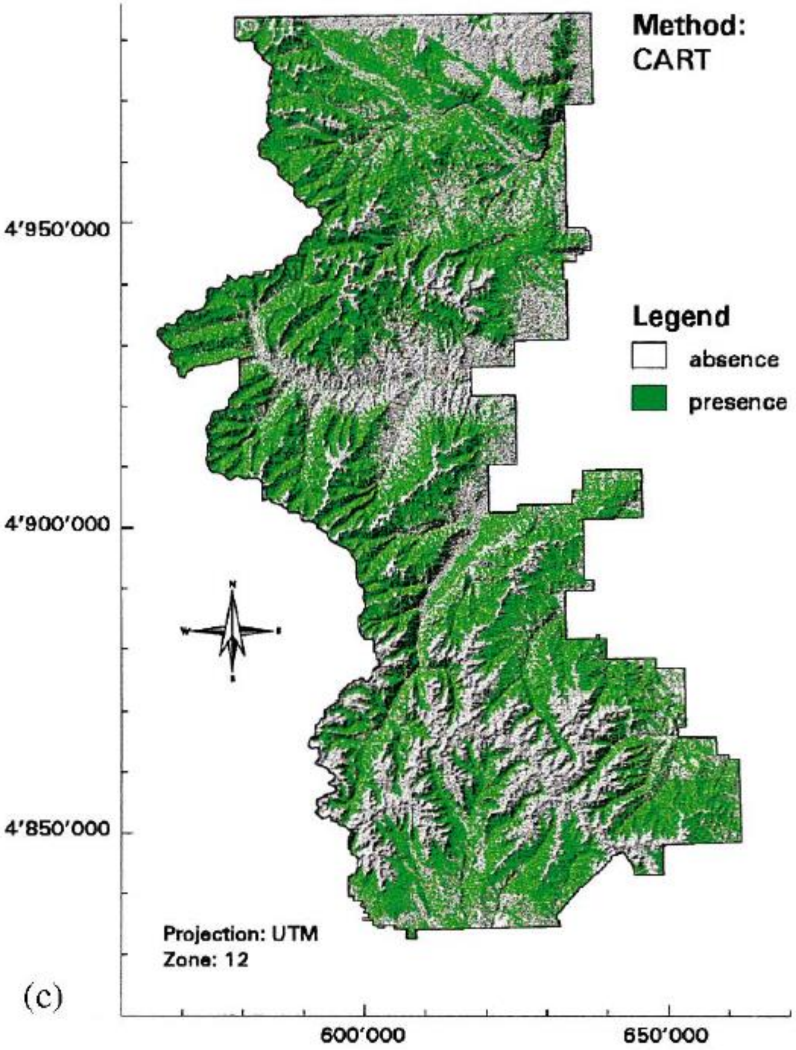
Predicted Occurrence

From non-probabilistic metric CART





Simulated presence: *Picea engelmannii*



From: Guisan and Zimmermann 2000

How good is a particular SDM?

Good models are both reliable and discriminatory

- **Reliable**:
  predicted probability is an accurate assessment of likelihood of finding a species at a given site.

- **Discriminatory**:
  a model's ability to separate habitat from non-habitat

A model can produce reliable predictions, but if it doesn't distinguish habitat from non-habitat, it's not very useful

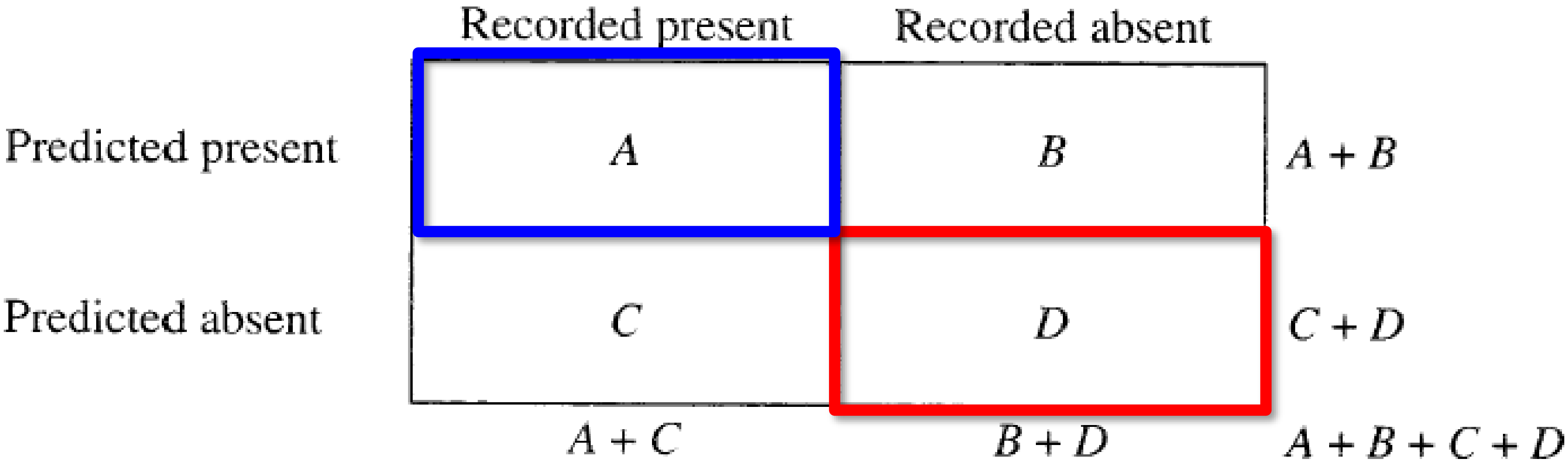
Components of model performance

- Accuracy of model predictions
    - Does the model make valid predictions?
- Rationality
    - May stumble upon a seemingly explanatory model, but one that makes little sense ("[Paul the Octopus…"](http://en.wikipedia.org/wiki/Paul_the_Octopus))

http://en.wikipedia.org/wiki/Paul_the_Octopus

- Interpretability of response variables
    - Are the predictions useful beyond habitat/non-habitat

Presence-absence confusion matrix

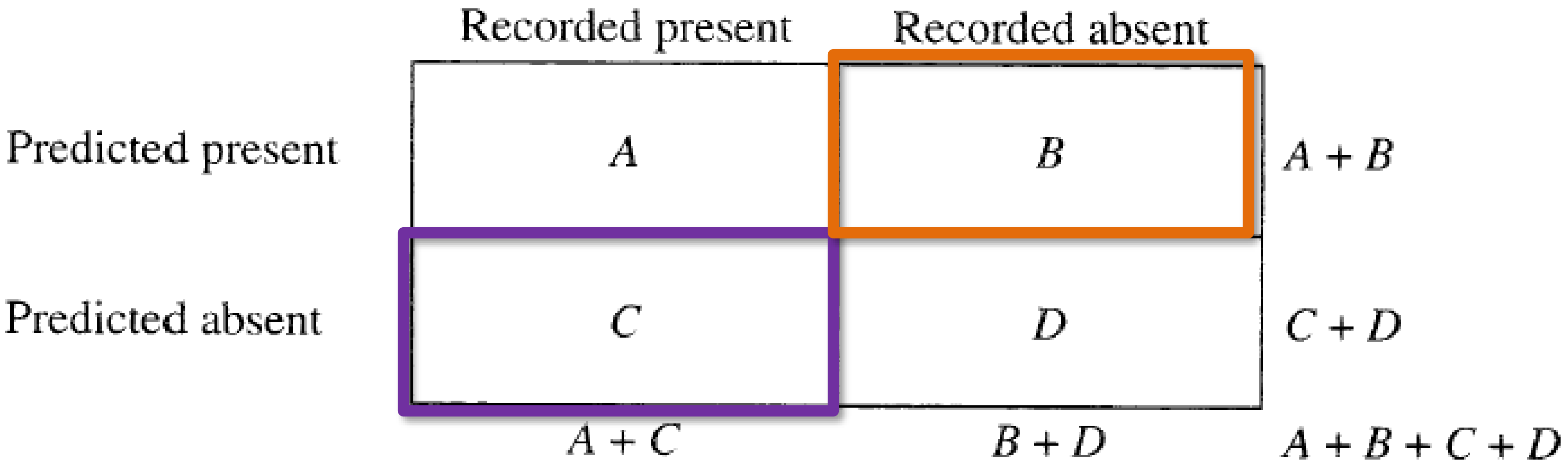


	Recorded present	Recorded absent	
Predicted present	A	B	$A + B$
Predicted absent	C	D	$C + D$
	$A + C$	$B + D$	$A + B + C + D$

(A)**True positive**: Species observed where predicted to be present

(D)**True negative**: Species absent where predicted to be absent

Presence-absence confusion matrix



	Recorded present	Recorded absent	
Predicted present	A	B	$A + B$
Predicted absent	C	D	$C + D$
	$A + C$	$B + D$	$A + B + C + D$

MISCLASSES

(B) **False positive**: Species absent where expected to be present

--- Errors of commission ---

(C) **False negative**: Species present where expected to be absent

--- Errors of omission ---

Presence-absence confusion matrix

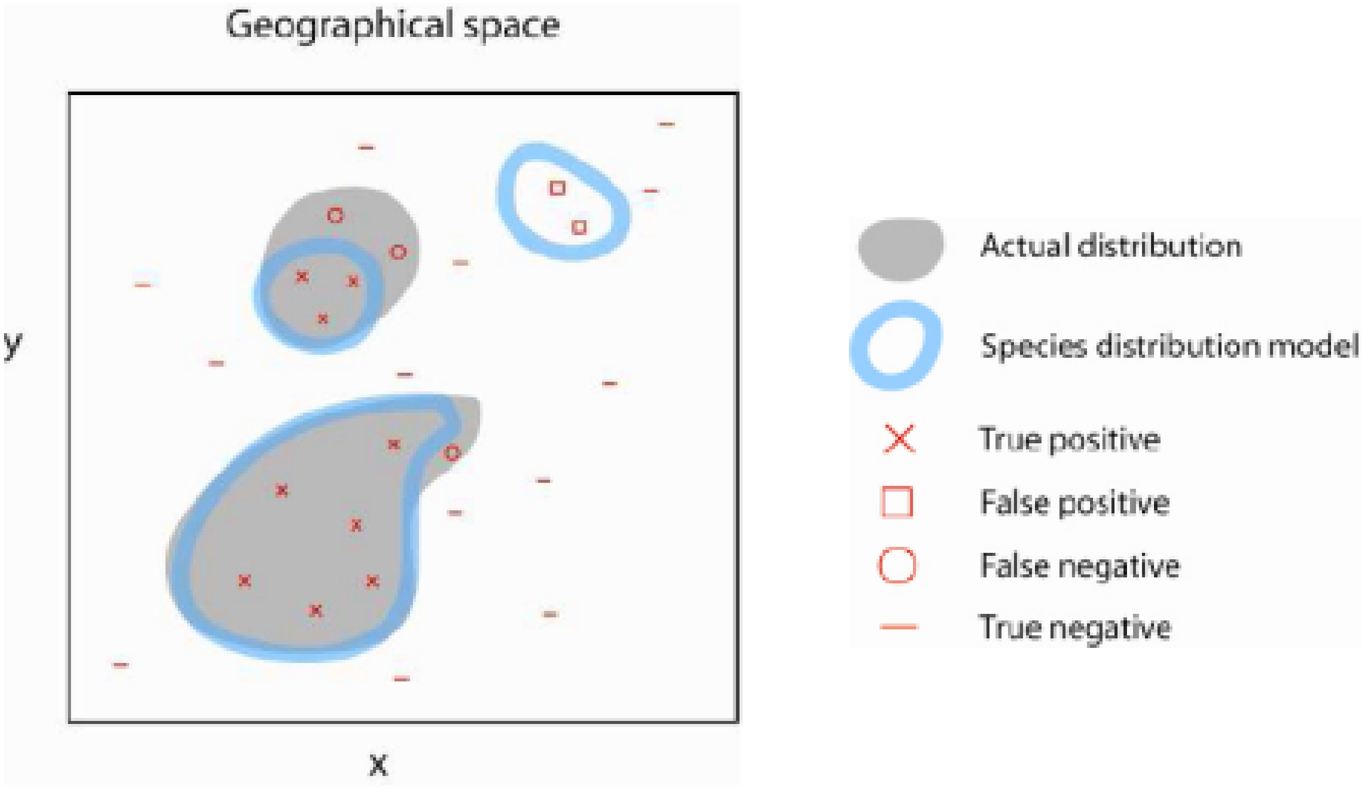
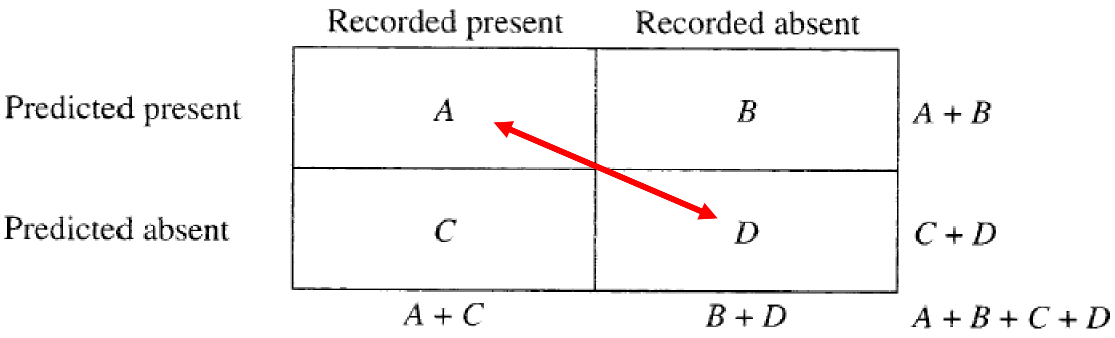


Figure 4. Diagram illustrating the four types of outcomes that are possible when assessing the predictive performance of a species distribution model: true positive, false positive, false negative and true negative. The diagram uses the same hypothetical actual and modeled distributions as in Figure 3. Each instance of a symbol (x, □, ○, -) on the map depicts a site that has been surveyed and presence or absence of the species recorded (it is assumed here that if a site falls within the actual distribution then the species will be detected). These survey records constitute the test data. Frequencies of each type of outcome are commonly entered into a confusion matrix (see main text).

Pearson 2008

Model accuracy



	Recorded present	Recorded absent	
Predicted present	A	B	$A + B$
Predicted absent	C	D	$C + D$
	$A + C$	$B + D$	$A + B + C + D$

$$\text{Accuracy} = \frac{A + D}{A + B + C + D}$$

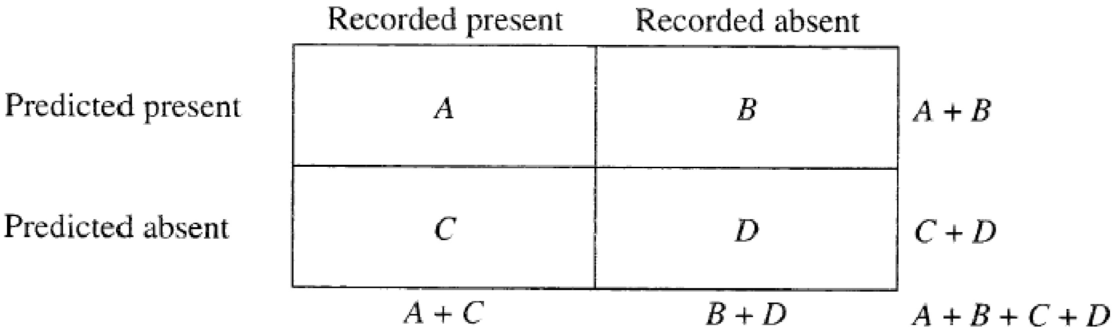
BUT…

- What if your species is *rare* (i.e., it doesn't usually occupy all available habitat)?

→ Your model's "accuracy" would falsely increase if you under-predicted habitat...

Kappa Statistic

$$\frac{\text{(observed accuracy – chance agreement)}}{\text{(1 – chance agreement)}}$$



	Recorded present	Recorded absent	
Predicted present	A	B	$A + B$
Predicted absent	C	D	$C + D$
	$A + C$	$B + D$	$A + B + C + D$

$$\frac{[(a+d)-(((a+c)(a+b)+(b+d)(c+d))/n)]}{[n-(((a+c)(a+b)+(b+d)(c+d))/n)]}$$

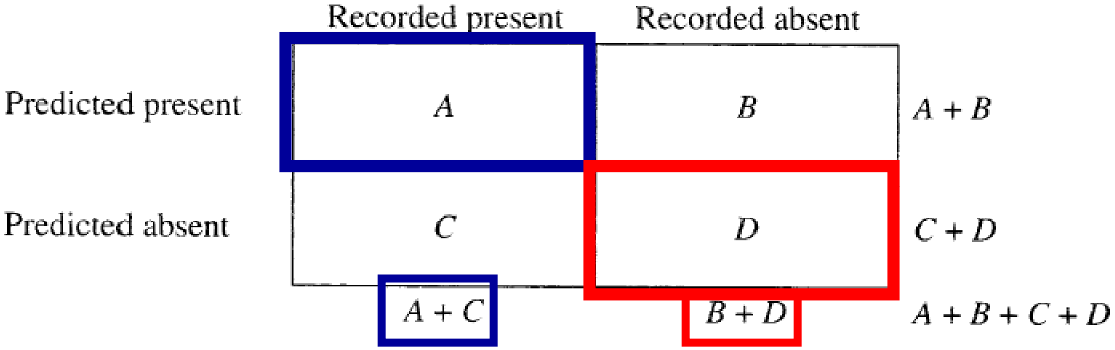
Adjustment to accuracy to account for chance agreement between predicted and observed values

Accuracy & Kappa Statistics

- *Accuracy* and *Kappa statistics* use all values in the confusion matrix and therefore require both presence and absence data.

- However, absence data are often unavailable (e.g. when using specimens from museum collections) and are inappropriate for use when the aim is to estimate the potential distribution (since the environment may be suitable even though the species is absent).

Measuring discrimination performance



	Recorded present	Recorded absent	
Predicted present	A	B	$A + B$
Predicted absent	C	D	$C + D$
	$A + C$	$B + D$	$A + B + C + D$

Sensitivity **True Positive Fraction**

$$= \frac{\text{Number of positive sites correctly predicted}}{\text{Total number of positive sites in sample}}$$

$$= \frac{A}{(A + C)} \qquad (4)$$

Specificity **True Negative Fraction**

$$= \frac{\text{Number of negative sites correctly predicted}}{\text{Total number of negative sites in sample}}$$

$$= \frac{D}{(B + D)} \qquad (5)$$

Measuring discrimination performance

	Recorded present	Recorded absent	
Predicted present	A	B	$A + B$
Predicted absent	C	D	$C + D$
	$A + C$	$B + D$	$A + B + C + D$

False positive fraction **"Omission rate"**

$$= \frac{\text{Number of false positive predictions}}{\text{Total number of negative sites in sample}}$$

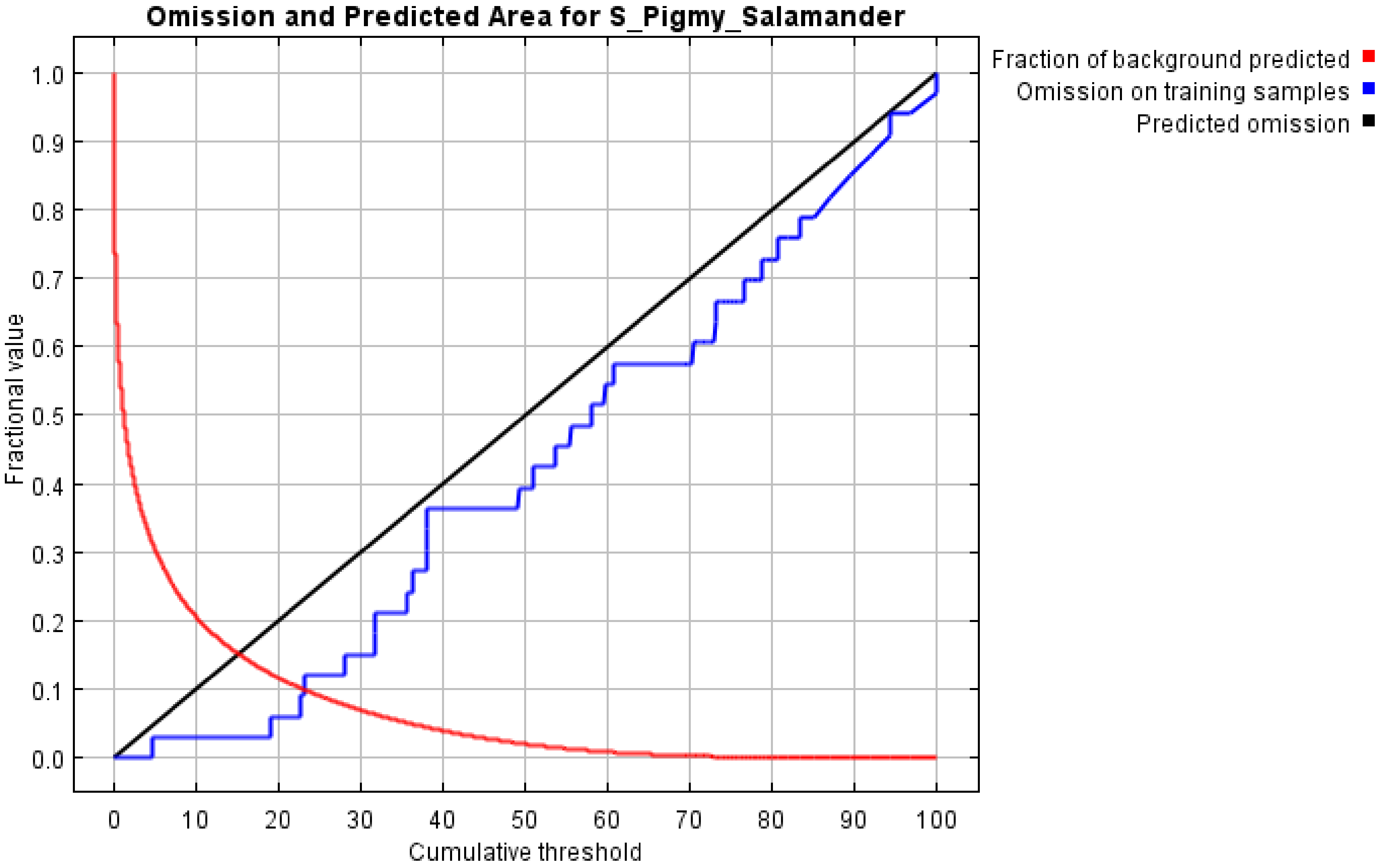
$$= \frac{B}{(B + D)} \qquad (6)$$

False negative fraction

$$= \frac{\text{Number of false negative predictions}}{\text{Total number of positive sites in sample}}$$

$$= \frac{C}{(A + C)} \qquad (7)$$

Measuring discrimination performance



Model *sensitivity*

$$\text{Sensitivity} = \frac{\text{Number of positive sites correctly predicted}}{\text{Total number of positive sites in sample}}$$

High sensitivity → low omission rate

"How likely is a model to correctly predict presence"

Can always achieve high sensitivity by classifying <u>all</u> area as "habitat"

Model *specificity*

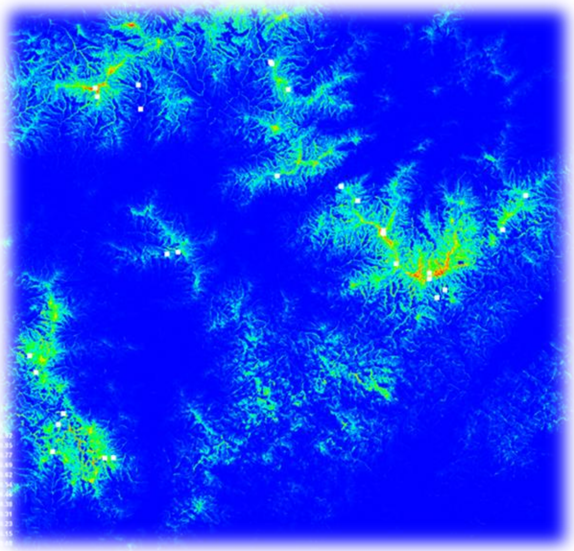
$$\text{Specificity} = \frac{\text{Number of negative sites correctly predicted}}{\text{Total number of negative sites in sample}}$$

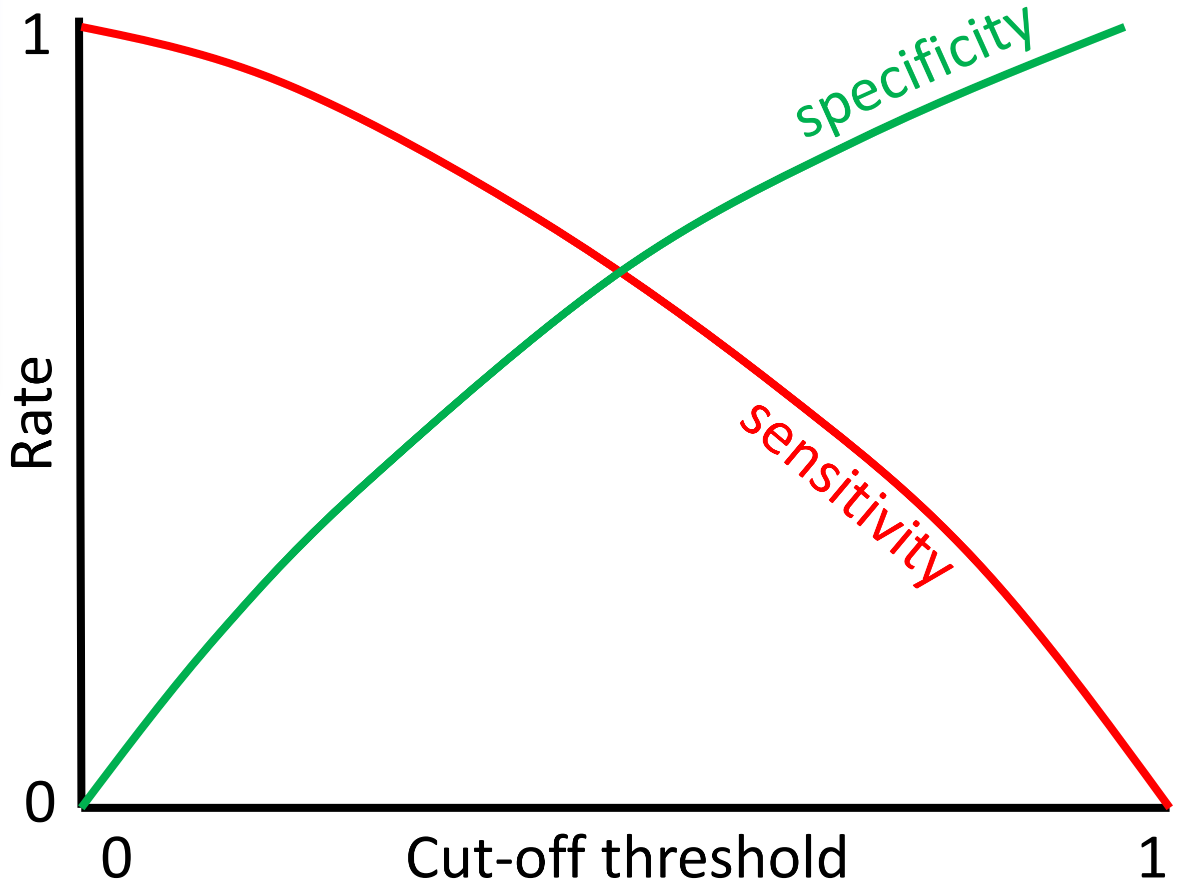
High specificity → low commission rate

"How likely is a model to correctly predict absence"

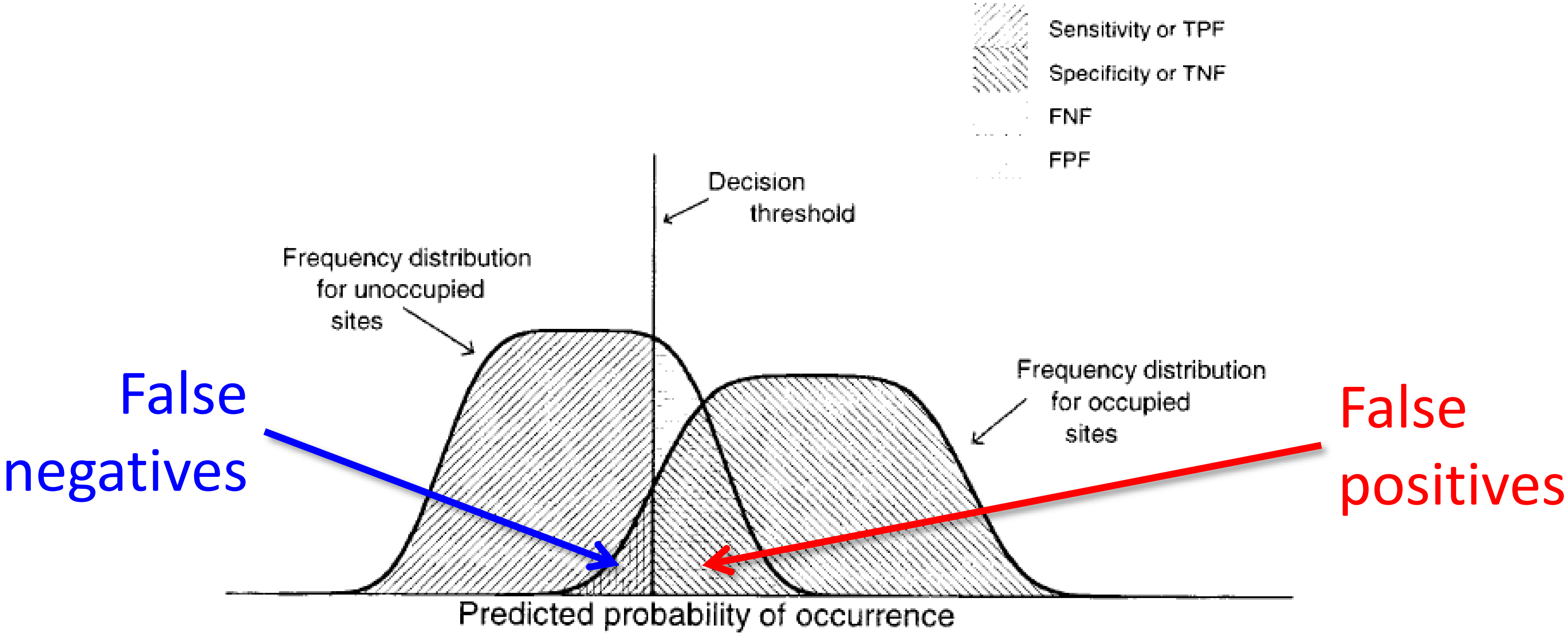
Can always achieve high specificity by classifying <u>no</u> area as "habitat"

Sensitivity/Specificity





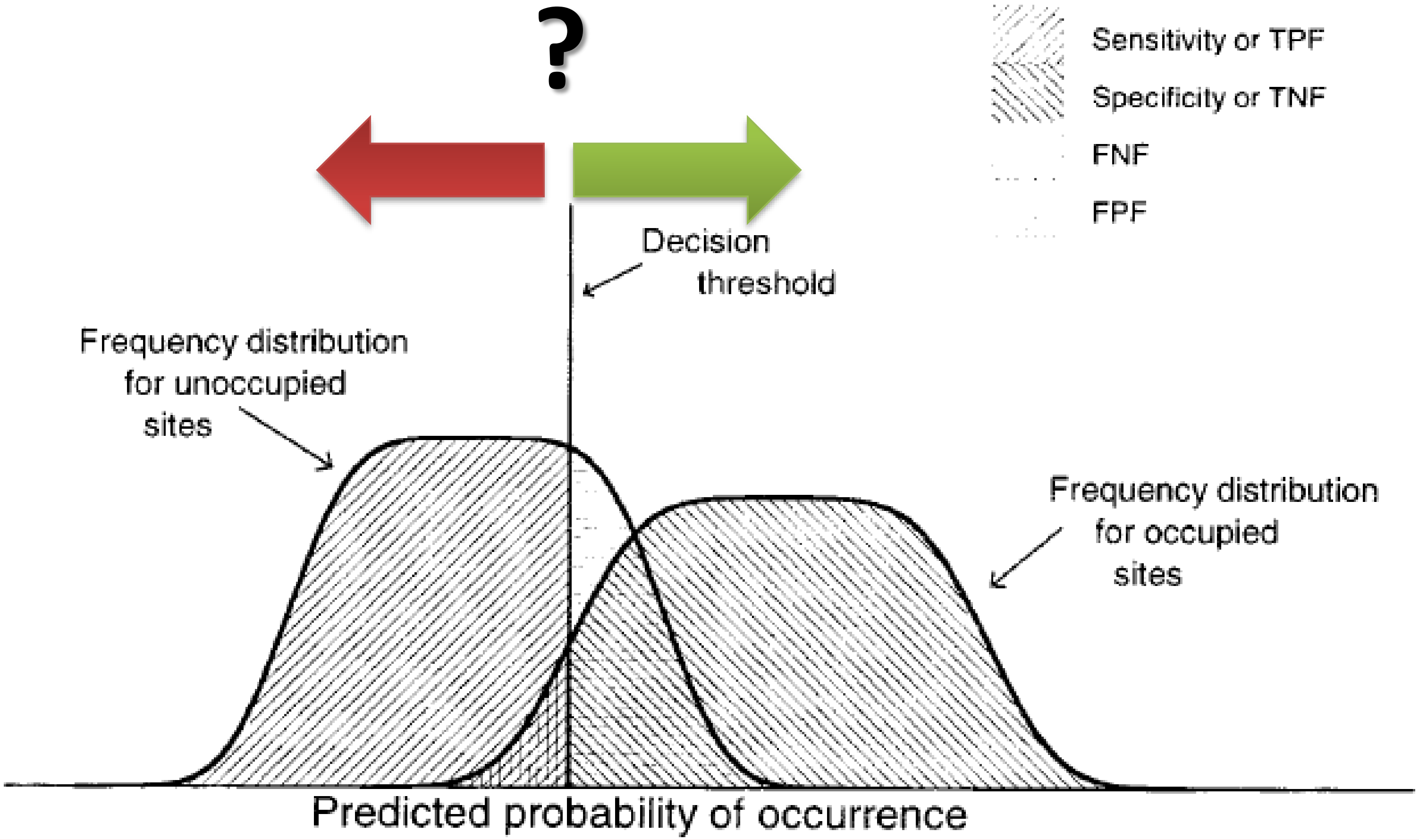
Sensitivity/Specificity



The curves represent the frequency distribution of probabilities predicted by a model for occupied and unoccupied sites within a data set for which the real distribution of the species is known.

A threshold probability, represented by the vertical line, separates sites predicted to be occupied from sites predicted to be unoccupied.

Model tuning

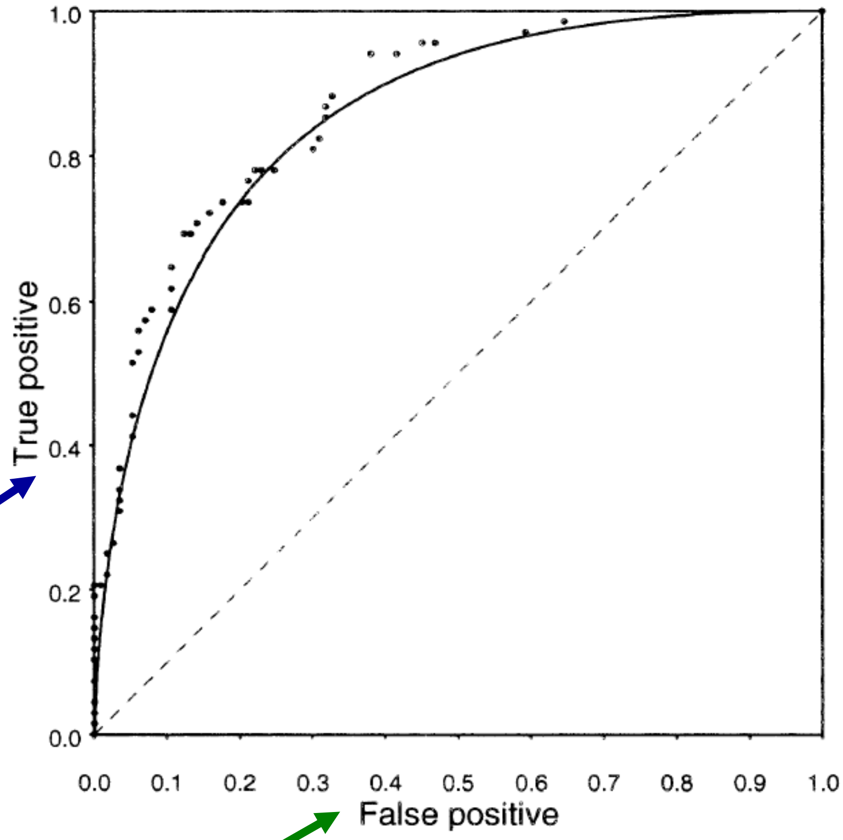


Where is the optimal place to put the decision threshold to minimize false positives & negatives?

Receiver operating characteristic (ROC)

The ROC graph in which the sensitivity (true positive proportion) is plotted against the false positive proportion for a range of threshold probabilities.

A smooth curve is drawn through the points to derive the ROC curve. The 45° line represents the sensitivity and false positive values expected to be achieved by chance alone for each decision threshold.



True Positive Fraction

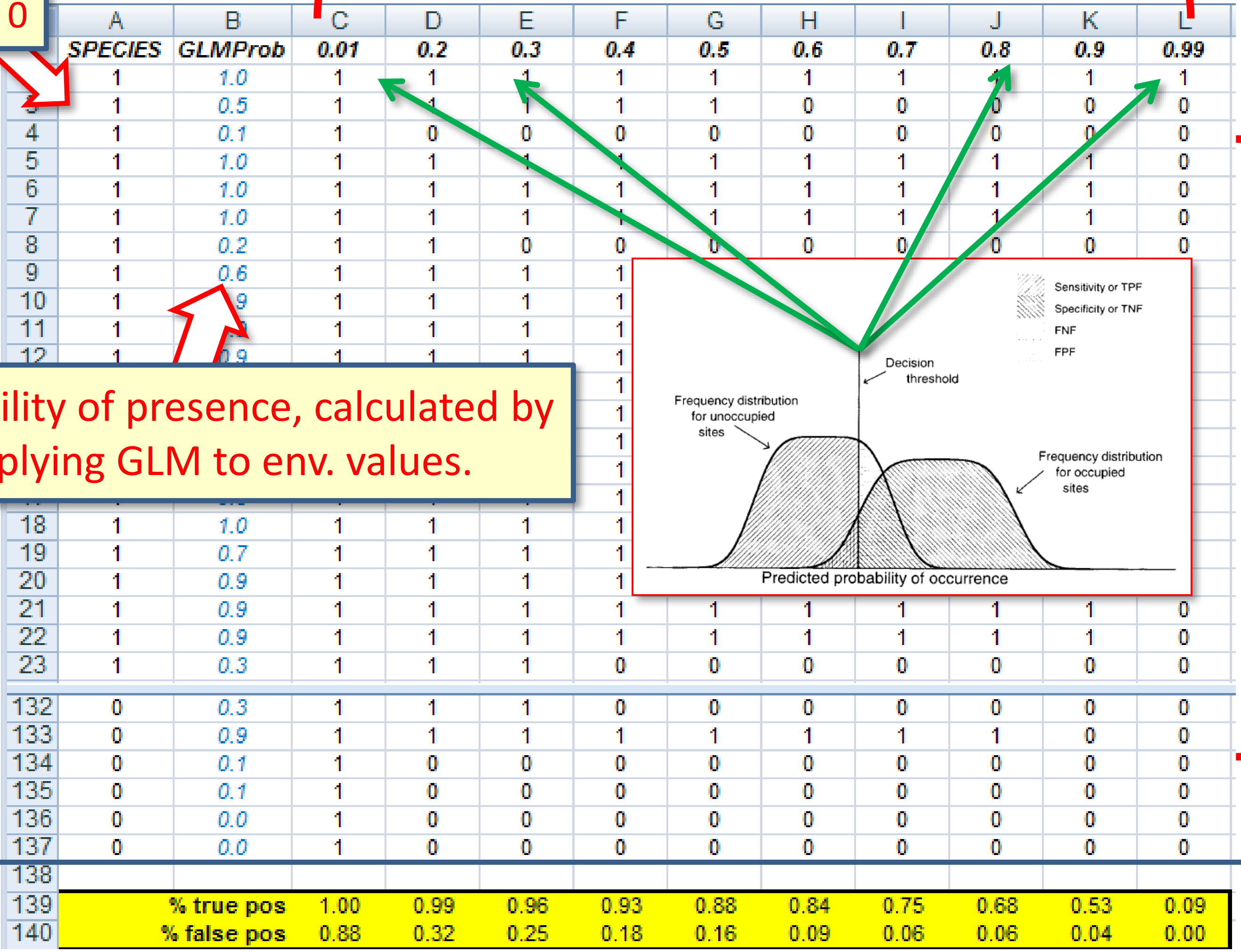
False Positive Fraction

	Recorded present	Recorded absent	
Predicted present	A	B	$A + B$
Predicted absent	C	D	$C + D$
	$A + C$	$B + D$	$A + B + C + D$

Building a ROC

OBSERVED
Presence = 1
Absence = 0

Decision threshold values

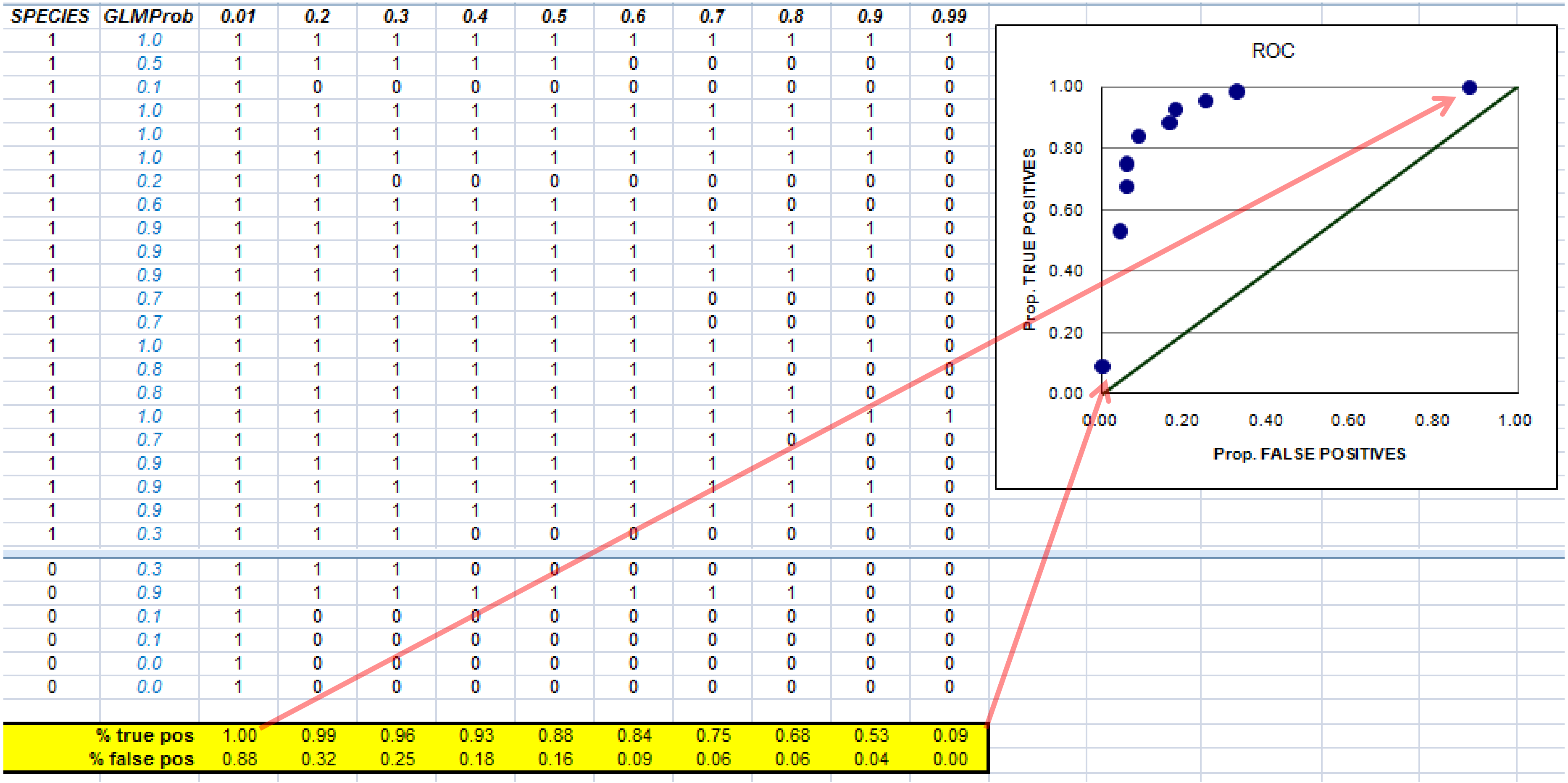


	A	B	C	D	E	F	G	H	I	J	K	L
	SPECIES	*GLMProb*	*0.01*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*	*0.99*
[illegible]	1	1.0	1	1	1	1	1	1	1	1	1	1
[illegible]	1	0.5	1	1	1	1	1	0	0	0	0	0
4	1	0.1	1	0	0	0	0	0	0	0	0	0
5	1	1.0	1	1	1	1	1	1	1	1	1	0
6	1	1.0	1	1	1	1	1	1	1	1	1	0
7	1	1.0	1	1	1	1	1	1	1	1	1	0
8	1	0.2	1	1	0	0	0	0	0	0	0	0
9	1	0.6	1	1	1	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
10	1	[illegible]	1	1	1	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
11	1	[illegible]	1	1	1	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
12	1	0.9	1	1	1	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
18	1	1.0	1	1	1	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
19	1	0.7	1	1	1	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
20	1	0.9	1	1	1	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
21	1	0.9	1	1	1	1	1	1	1	1	1	0
22	1	0.9	1	1	1	1	1	1	1	1	1	0
23	1	0.3	1	1	1	0	0	0	0	0	0	0
132	0	0.3	1	1	1	0	0	0	0	0	0	0
133	0	0.9	1	1	1	1	1	1	1	1	0	0
134	0	0.1	1	0	0	0	0	0	0	0	0	0
135	0	0.1	1	0	0	0	0	0	0	0	0	0
136	0	0.0	1	0	0	0	0	0	0	0	0	0
137	0	0.0	1	0	0	0	0	0	0	0	0	0
138												
139		**% true pos**	1.00	0.99	0.96	0.93	0.88	0.84	0.75	0.68	0.53	0.09
140		**% false pos**	0.88	0.32	0.25	0.18	0.16	0.09	0.06	0.06	0.04	0.00

Probability of presence, calculated by applying GLM to env. values.

PREDICTED
set to **1** if
GLM prob.
exceeds
threshold;
0 if not

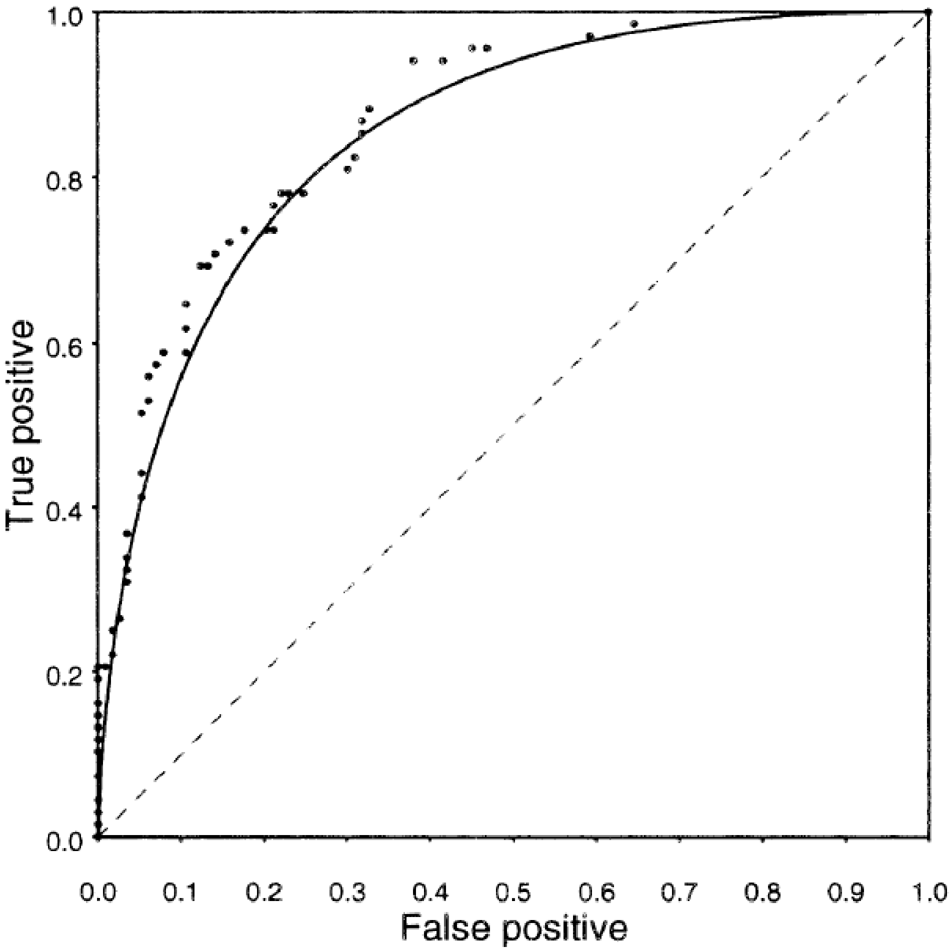
Building a ROC



SPECIES	*GLMProb*	*0.01*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*	*0.9*	*0.99*
1	*1.0*	1	1	1	1	1	1	1	1	1	1
1	*0.5*	1	1	1	1	1	0	0	0	0	0
1	*0.1*	1	0	0	0	0	0	0	0	0	0
1	*1.0*	1	1	1	1	1	1	1	1	1	0
1	*1.0*	1	1	1	1	1	1	1	1	1	0
1	*1.0*	1	1	1	1	1	1	1	1	1	0
1	*0.2*	1	1	0	0	0	0	0	0	0	0
1	*0.6*	1	1	1	1	1	1	0	0	0	0
1	*0.9*	1	1	1	1	1	1	1	1	1	0
1	*0.9*	1	1	1	1	1	1	1	1	1	0
1	*0.9*	1	1	1	1	1	1	1	1	0	0
1	*0.7*	1	1	1	1	1	1	0	0	0	0
1	*0.7*	1	1	1	1	1	1	0	0	0	0
1	*1.0*	1	1	1	1	1	1	1	1	1	0
1	*0.8*	1	1	1	1	1	1	1	0	0	0
1	*0.8*	1	1	1	1	1	1	1	1	0	0
1	*1.0*	1	1	1	1	1	1	1	1	1	1
1	*0.7*	1	1	1	1	1	1	1	0	0	0
1	*0.9*	1	1	1	1	1	1	1	1	0	0
1	*0.9*	1	1	1	1	1	1	1	1	1	0
1	*0.9*	1	1	1	1	1	1	1	1	1	0
1	*0.3*	1	1	1	0	0	0	0	0	0	0
0	*0.3*	1	1	1	0	0	0	0	0	0	0
0	*0.9*	1	1	1	1	1	1	1	1	0	0
0	*0.1*	1	0	0	0	0	0	0	0	0	0
0	*0.1*	1	0	0	0	0	0	0	0	0	0
0	*0.0*	1	0	0	0	0	0	0	0	0	0
0	*0.0*	1	0	0	0	0	0	0	0	0	0
	% true pos	1.00	0.99	0.96	0.93	0.88	0.84	0.75	0.68	0.53	0.09
	% false pos	0.88	0.32	0.25	0.18	0.16	0.09	0.06	0.06	0.04	0.00

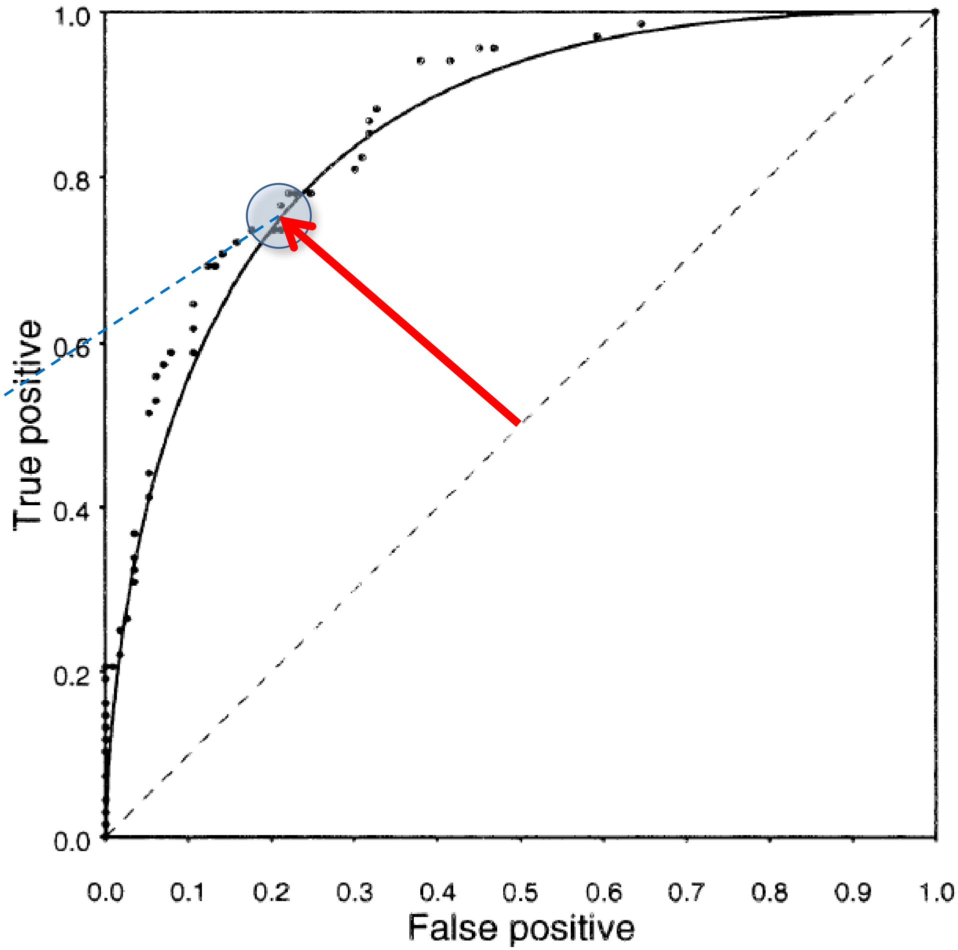
ROC

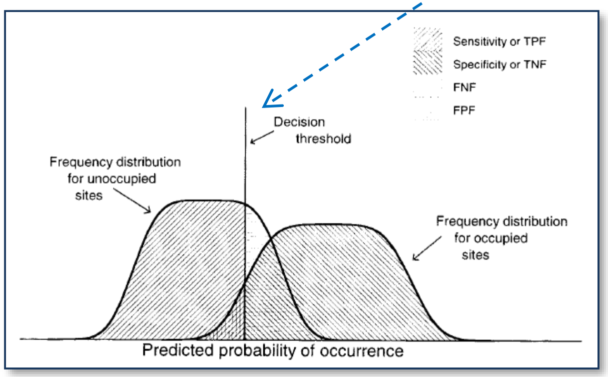
- Each dot represents a plot of true positives against false positives for a given decision threshold
- The diagonal line represents what's expected by chance alone (GLM probs are random)
- The further from the diagonal line, the more discriminating your model is



ROC

- The point furthest from the diagonal represents an optimal decision threshold, (i.e. the best balance between false negatives and false positives)

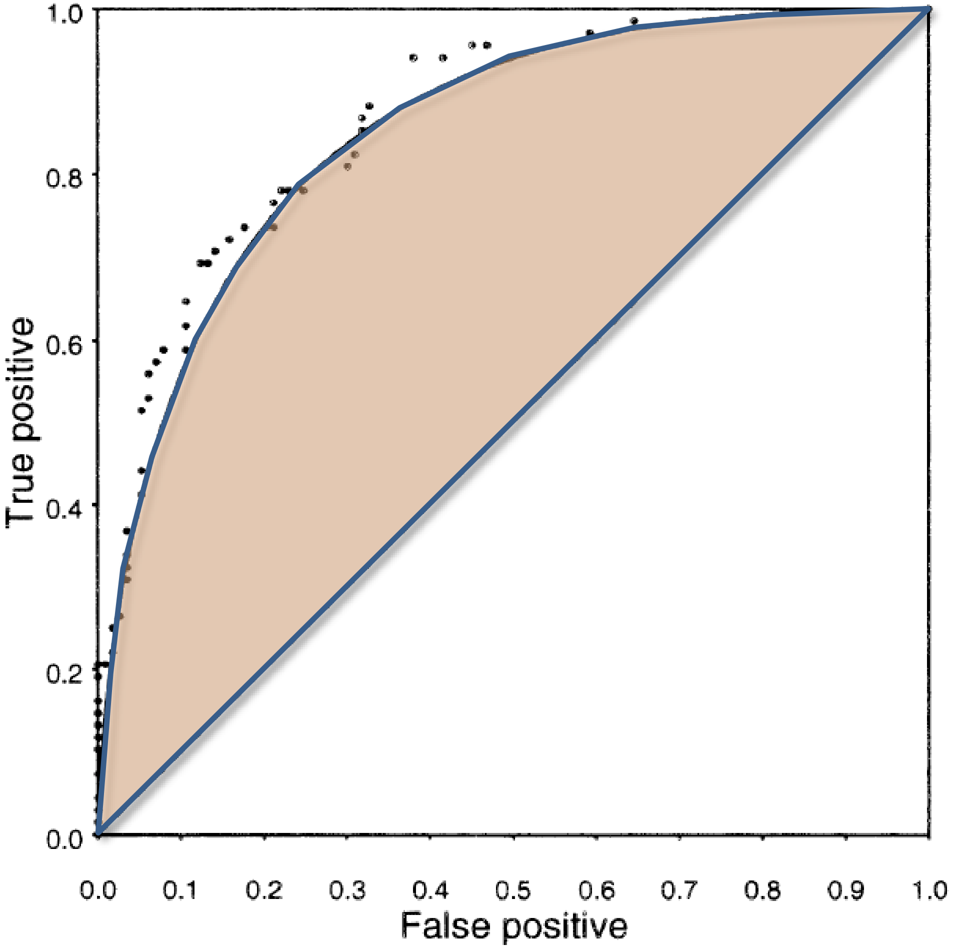




Area under the curve (AUC)

- Measuring the area under the ROC curve gives a quantifiable estimate of the overall goodness of your model.

- The area under an ROC curve (AUC) has a natural statistical interpretation. Pick a random positive example and a random negative example. The area under the curve is the probability that the classifier correctly orders the two points (with random ordering in the case of ties). A perfect classifier therefore has an AUC of 1. (Phillips et al 2004)



Tuning your model

Choosing a <u>higher</u> probability threshold:

- Increases false negatives (actual habitat that may not get mapped as habitat)
- Areas that do get labeled habitat in final map are more certain to be habitat

Choosing a <u>lower</u> threshold:

- Increases false positives (maps habitat areas where habitat may not really exist)
- Less certain that habitat areas are truly habitat

Tuning your model: Spotted Owl

Mexican spotted owl:

- Threatened species: the goal (law) is to manage potential habitat to minimize impacts on the species

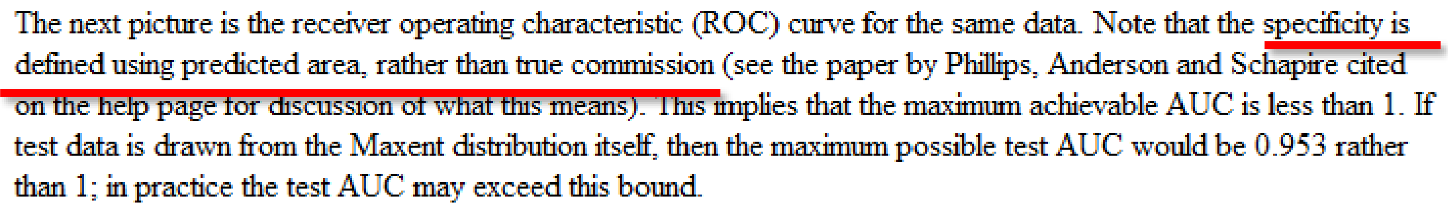
Northern spotted owl:

- More pressure: the NWFP identifies special management areas that will be highly protected (off-limits)

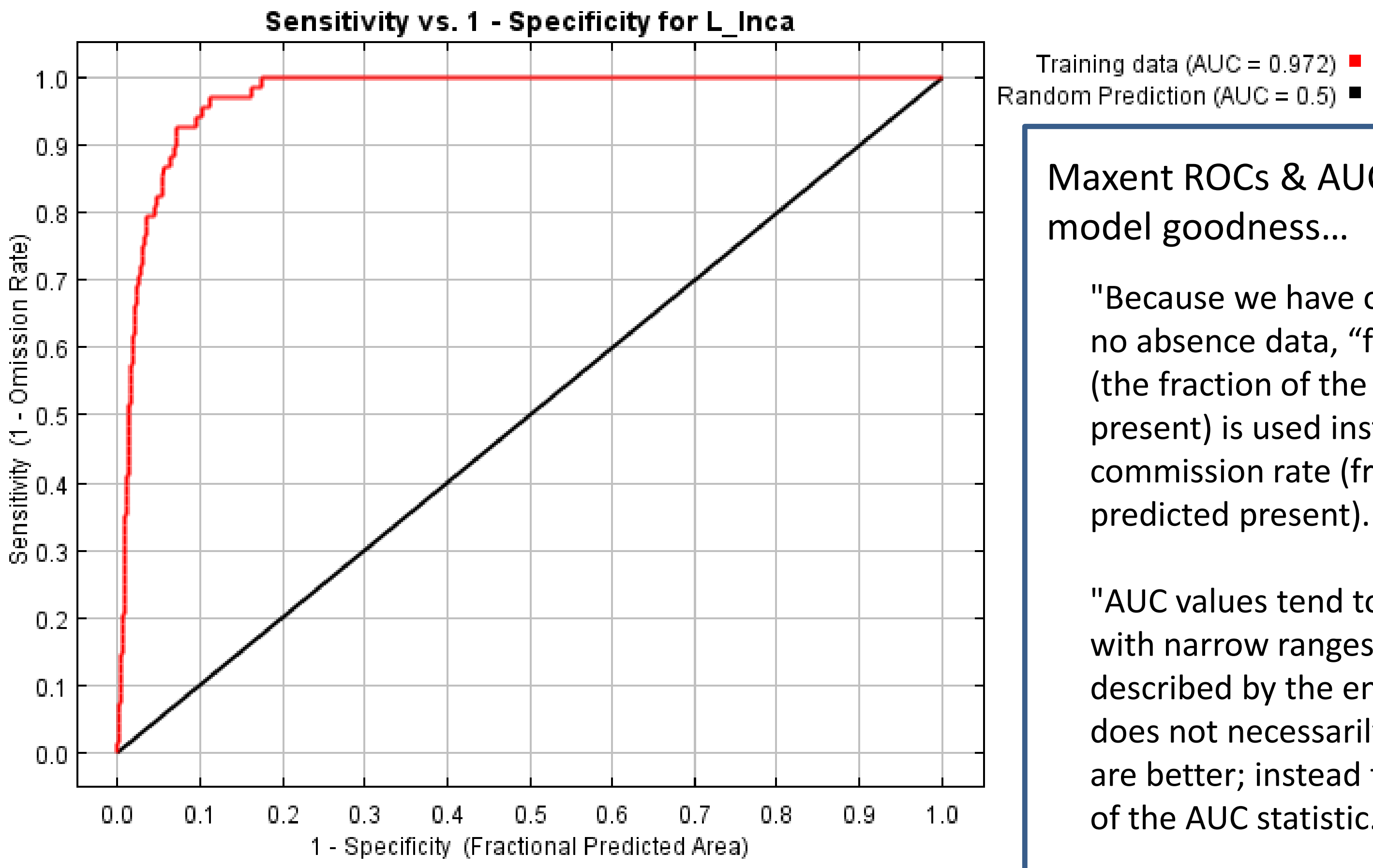
Tuning maxent models

- MaxEnt is a “presence-only” model so it doesn’t have “true negatives”
- the MaxEnt software uses pseudo-ROC to maximize “true positives” while minimizing total area predicted to be “habitat”
- there is no correct way to tune a MaxEnt model this way (as with others)

ROC/AUC - Maxent



The next picture is the receiver operating characteristic (ROC) curve for the same data. Note that the specificity is defined using predicted area, rather than true commission (see the paper by Phillips, Anderson and Schapire cited on the help page for discussion of what this means). This implies that the maximum achievable AUC is less than 1. If test data is drawn from the Maxent distribution itself, then the maximum possible test AUC would be 0.953 rather than 1; in practice the test AUC may exceed this bound.

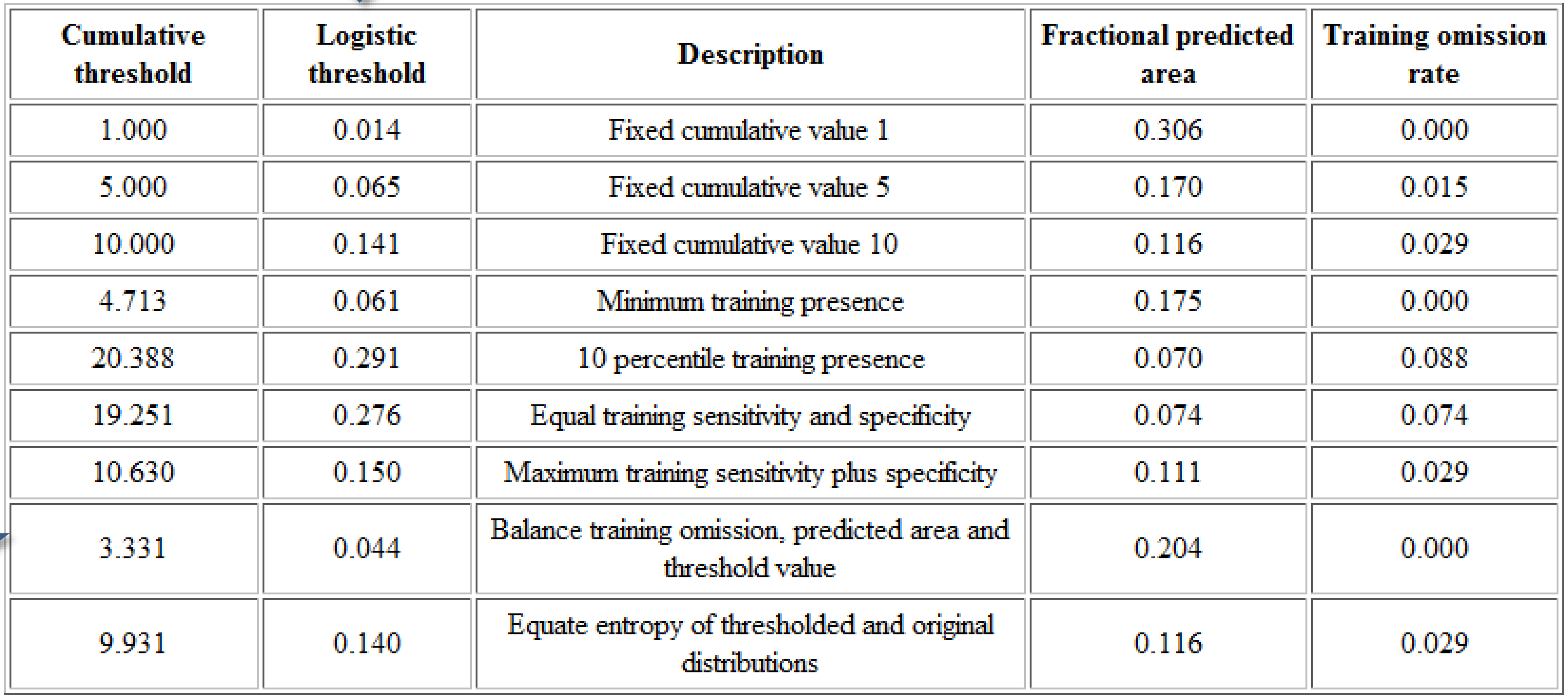


Maxent ROCs & AUCs tend to overestimate model goodness…

> "Because we have only occurrence data and no absence data, "fractional predicted area" (the fraction of the total study area predicted present) is used instead of the more standard commission rate (fraction of absences predicted present). "
>
> "AUC values tend to be higher for species with narrow ranges, relative to the study area described by the environmental data. This does not necessarily mean that the models are better; instead this behavior is an artifact of the AUC statistic. "

Phillips. et al 2004

ROC/AUC - Maxent



Cumulative threshold	Logistic threshold	Description	Fractional predicted area	Training omission rate
1.000	0.014	Fixed cumulative value 1	0.306	0.000
5.000	0.065	Fixed cumulative value 5	0.170	0.015
10.000	0.141	Fixed cumulative value 10	0.116	0.029
4.713	0.061	Minimum training presence	0.175	0.000
20.388	0.291	10 percentile training presence	0.070	0.088
19.251	0.276	Equal training sensitivity and specificity	0.074	0.074
10.630	0.150	Maximum training sensitivity plus specificity	0.111	0.029
3.331	0.044	Balance training omission, predicted area and threshold value	0.204	0.000
9.931	0.140	Equate entropy of thresholded and original distributions	0.116	0.029

ROC

Conclusions:

- ROC relative operating characteristic (ROC) curve provide a framework for evaluating habitat models.
- ROC methods are analogous to “Accuracy Assessment” methods.
- ROC methods provide diagnostic information on both model Calibration and Discrimination.
- ROC allows for informed "tuning" of model output

Habitat models: misclasses

Pulliam (2000):

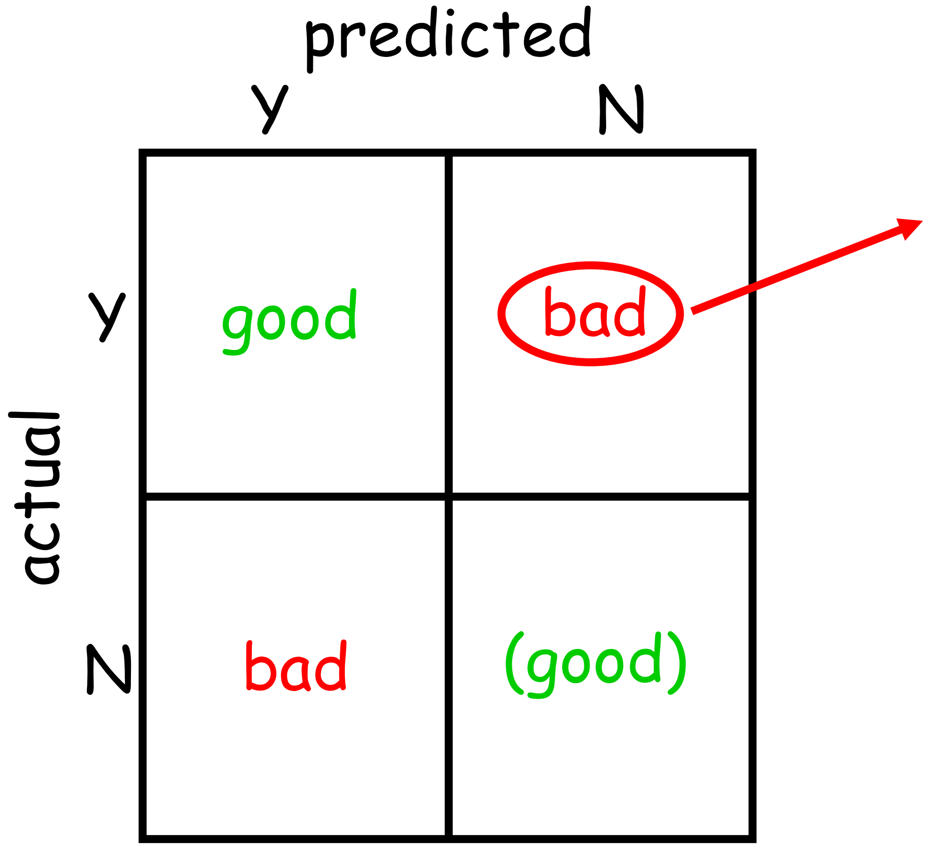
- Original context (1970’s): competition and the assembly of communities
- More recent: metapopulations and area vs isolation effects (especially at landscape scales where geospatial implementations dominate)

Habitat models: misclasses

Moving these analyses into GIS:

- Switches to geospatial predictors that are coarser-grained but of larger extent
- Switches the focus of the ecology from microhabitat (communities) to landscape ecology and meta-communities, or to biogeography

Habitat models: misclasses



predicted

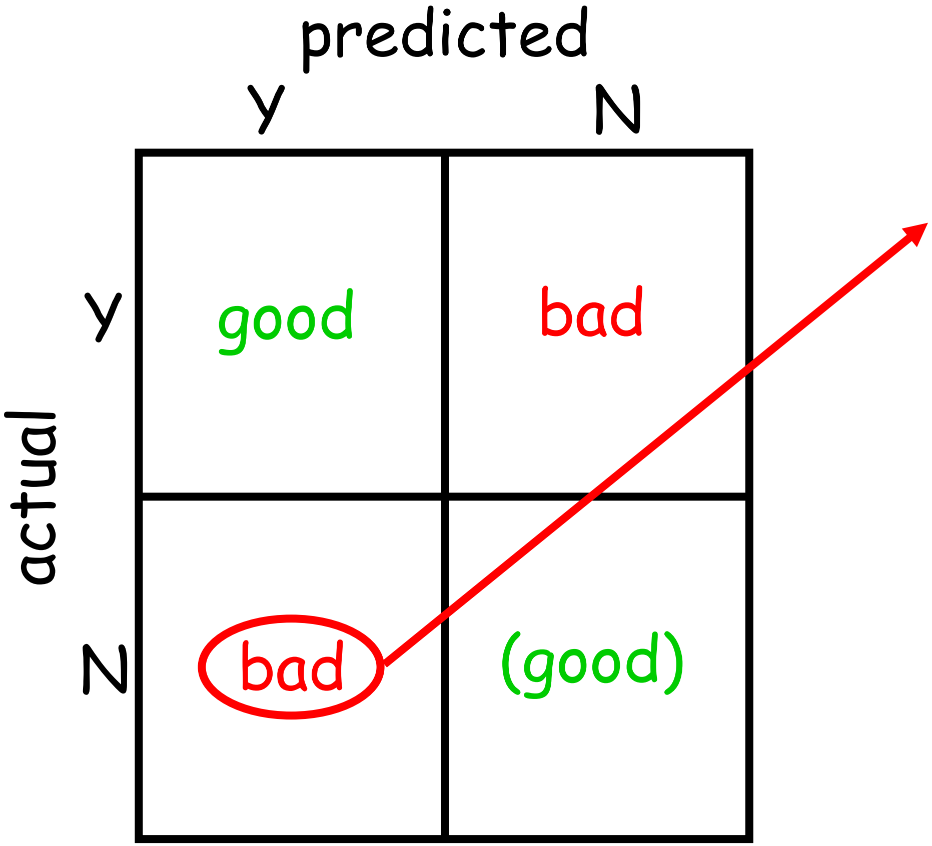
actual

	Y	N
Y	good	bad
N	bad	(good)

species occurs in places where it shouldn't ("not habitat"):

- bad model?
- dispersal subsidy?

Habitat models: misclasses



actual \ predicted	Y	N
Y	good	bad
N	bad	(good)

species does not occur where it should ("habitat"):

- bad model?
- dispersal limits?
- disturbance?
- biogeography?
- or simply rare?

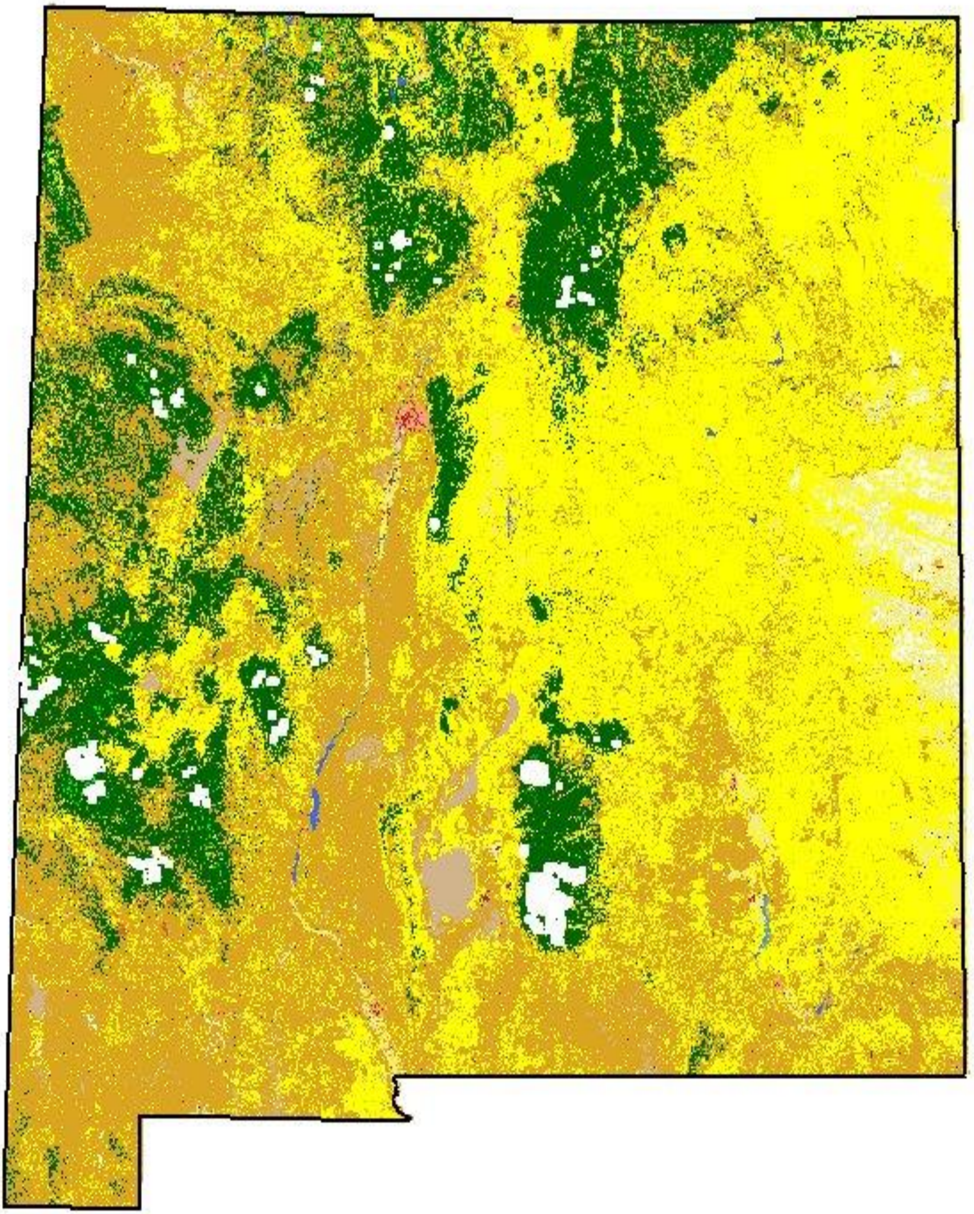
Habitat models: interpretation

Regression analysis:

- Bias?
  - Check residuals to see if they are correlated with predictors
  - Check residuals to see if they are correlated with response
- Spatial errors?
  - Check residuals for autocorrelation

Habitat models

Mexican spotted owl distribution in NM

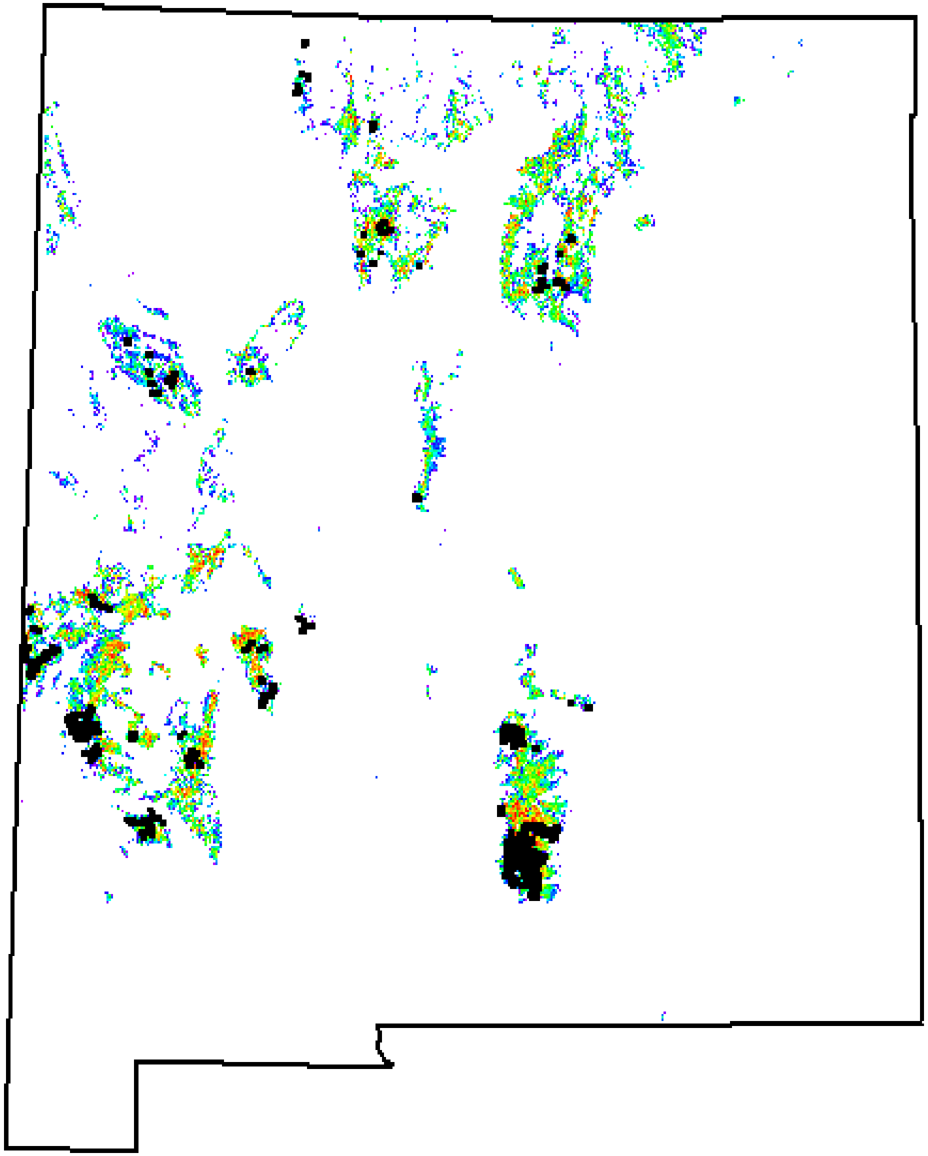


forests
range

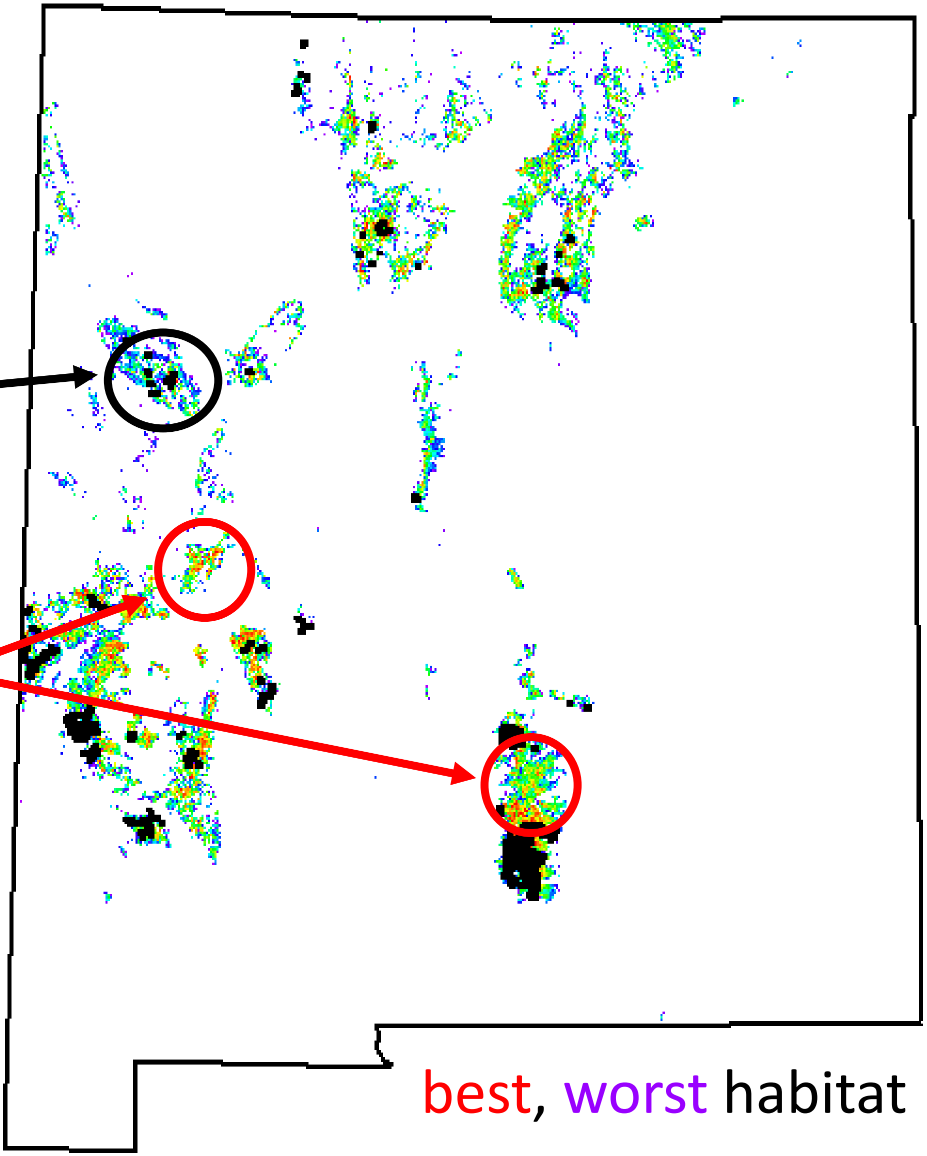
Habitat models

Mahalanobis distance (cells most similar to the observed owl locations)

best, worst habitat



Habitat models



occupied, marginal habitat

unoccupied, good habitat

best, worst habitat

Habitat models: interpretation

Map predicted habitat …

- Visual inspection of model errors is usually revealing:
  - suspect classification?
  - dispersal subsidy or constraints?
  - other confounding factors?

Habitat models: interpretation

- Partial regression logic:
  - If we know where potential habitat is, then misclasses can tell us a lot about other factors:
    - False positives in isolated patches?
    - False negatives in patches near sources?
- This will be the basis for inferential models and habitat management

Habitat models: ensembles

- Averaging models provides an estimate of consensus “best habitat”
- Locations where the models do *not* agree provide insight into the assumptions of each model (we can learn from these disagreements)

Habitat models: interpretation

		predicted	
		Y	N
actual	Y	good	bad
	N	bad	(good)

the only cases that are *really* interesting are the model failures

Maps are useful!

- Nothing about habitat modeling *requires* GIS
- Mapping habitat is useful for management (siting)
- Mapping habitat models is immensely useful for interpreting and evaluating models