Day 3: Model Evaluation and Prediction

1. Model Evaluation
2. Model Predictions and Limitations
3. Presenting Results

1. Model Evaluation

1.1 Overview

1.2 Types of Model Evaluation

1.3 Evaluation Metrics

1.4 Thresholds

1.5 Exercise

1.6 Different Evaluation Approaches

1.7 Exercise

1.1 Overview of Model Evaluation

So you made a model. But is it…

…accurate?
…reliable?
…useful for making predictions to new data?
…insensitive to assumptions?

1.1 Overview of Model Evaluation

So you made a model. But is it…

…accurate?
…reliable?
…useful for making predictions to new data?
…insensitive to assumptions?

Model evaluation is a critical aspect of building and using niche models

1.2 Types of Model Validation

Some terminology:

Training or calibration data = data used fit the models

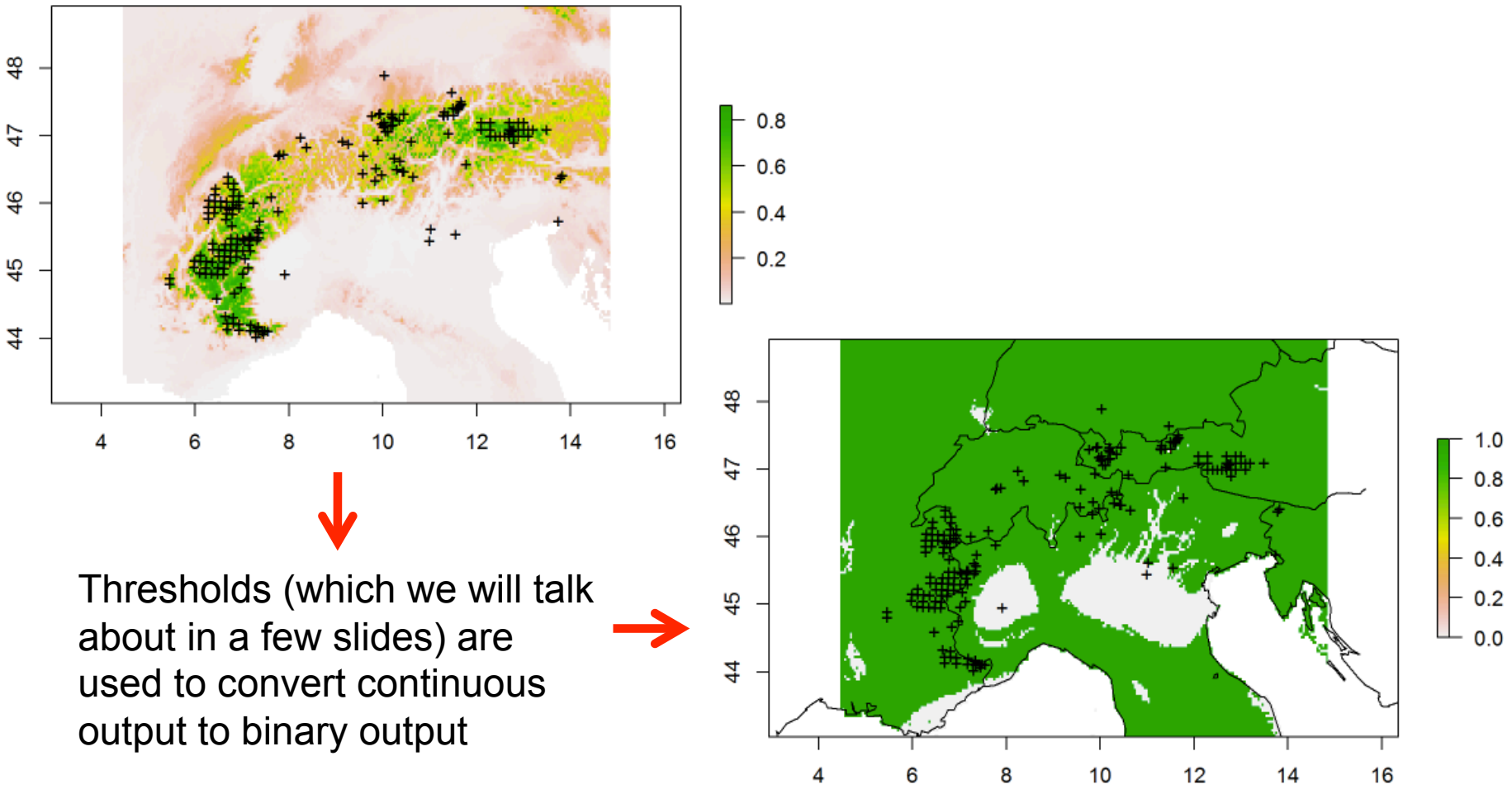
Testing of validation data = data used to test the performance of the model

1.2 Types of Model Validation

- **Resubstitution**
  - The same data used to build the model are then used to assess the model (i.e. how well does the model predict the sample data?)
  - Obviously not ideal…but sometimes only option
- **Internal Validation**
  - Data are permutated so that model is build from one subset of the data and tested on another
  - If the process is repeated multiple times, the sensitivity of the model to different data partitions can be assessed
  - But recall that bias in the sampling can affect these metrics…the training and testing data are not truly independent
- **External Validation**
  - Data used to test the model come from an independent survey designed to evaluate the model (e.g. randomized field test)
  - This is the gold standard but often logistically infeasible

1.3 Evaluation Metrics

Model output can be presented in a continuous or binary (suitable/not suitable) fashion…



Thresholds (which we will talk about in a few slides) are used to convert continuous output to binary output

1.3 Evaluation Metrics

Some evaluation metrics are **threshold-dependent**, requiring predictions for different places to be binary “yes”/”no”

Others are considered **threshold-independent** (though this usually means they evaluate the models across a range of thresholds)

<u>Threshold-Dependent</u>

e.g.
Commission/Omission
Sensitivity/Specificity
Kappa
True test statistic (TSS)

<u>Threshold-Independent</u>

e.g.
ROC/AUC

1.3 Evaluation Metrics

Most rely on what is known as a **confusion matrix**

This involves classifying presences and absences (or background sites) based on whether our model correctly predicts them (assuming we have binary predictions)

The “mistakes” that the model makes are known as **commission** and **omission** errors

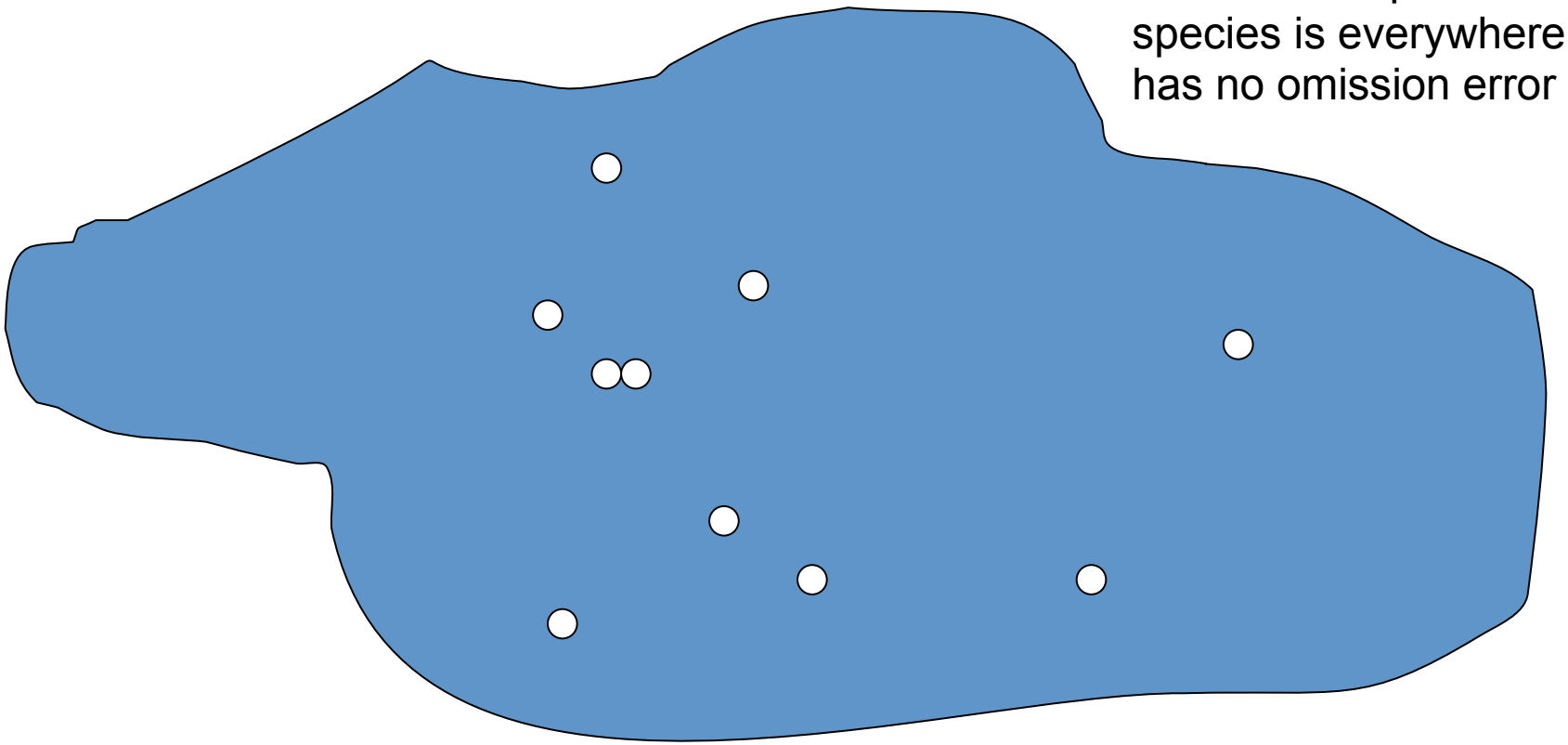
	Actually Present	Actually Absent
Predicted Present	*a*	*b*
Predicted Absent	*c*	*d*

a, d = correct predictions
b = **commission error** (false positives or overprediction of the model)
c = **omission error** (false negatives or underprediction of the model)

1.3 Evaluation Metrics

Problems with using omission or commission…

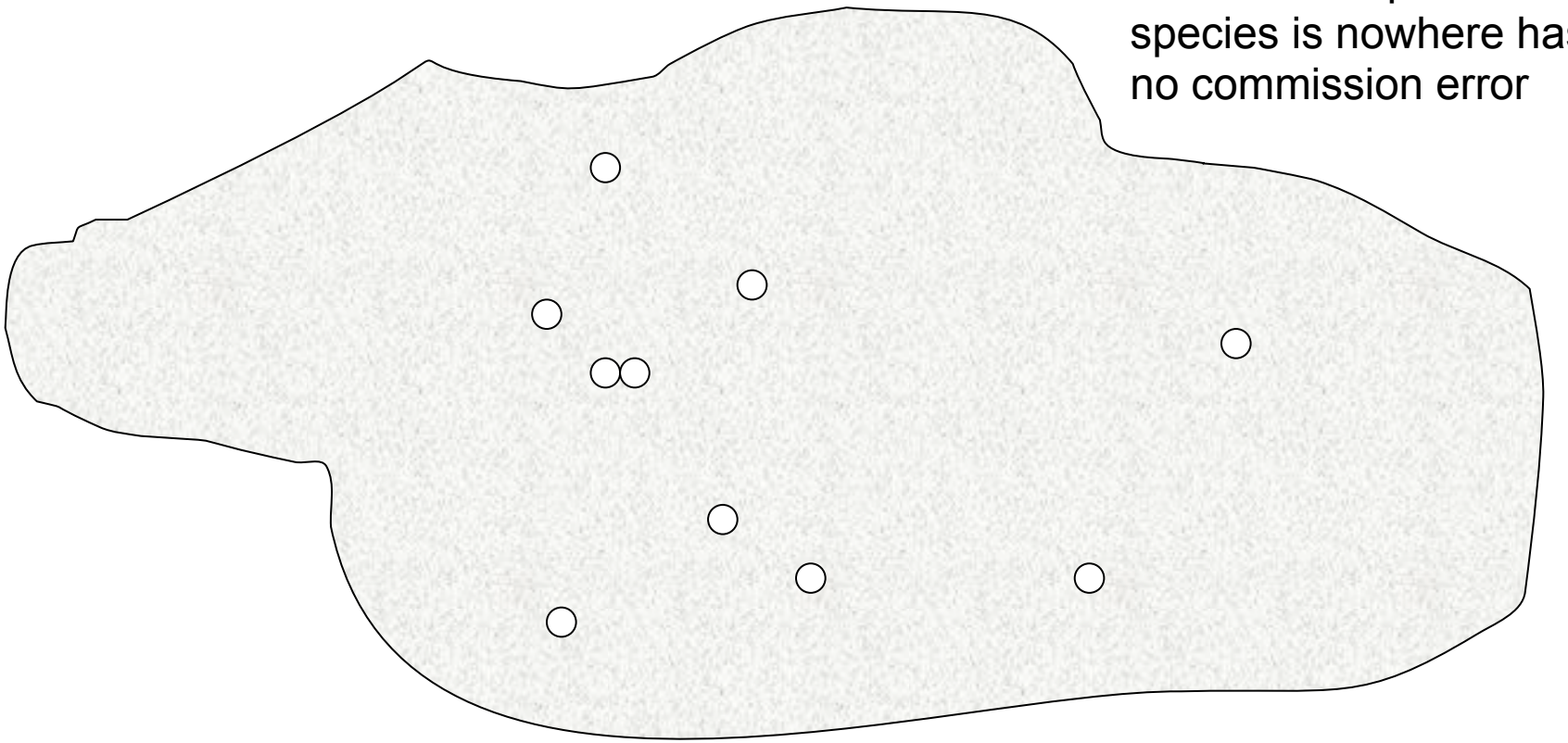
A model that predicts the species is everywhere has no omission error



1.3 Evaluation Metrics

Problems with using omission or commission…

A model that predicts the species is nowhere has no commission error



1.3 Evaluation Metrics

From here we can define the **sensitivity** and **specificity** of the model

	Actually Present	Actually Absent
Predicted Present	*a*	*b*
Predicted Absent	*c*	*d*

a, d = correct predictions
b = **commission error** (false positives or overprediction of the model)
c = **omission error** (false negatives or underprediction of the model)

Sensitivity (ability to correctly identify true presence)

= a / (a+c)

Specificity (ability to correctly identify true absence)

= d / (b+d)

Altogether, commission, omission, sensitivity and specificity are the simplest evaluation metrics…they are all threshold-dependent

1.3 Evaluation Metrics

Other threshold-dependent metrics

Table 2. Measures of predictive accuracy calculated from a 2×2 error matrix (Table 1). Overall accuracy is the rate of correctly classified cells. Sensitivity is the probability that the model will correctly classify a presence. Specificity is the probability that the model will correctly classify an absence. The kappa statistic and TSS normalize the overall accuracy by the accuracy that might have occurred by chance alone. In all formulae $n = a + b + c + d$

Measure	Formula
Overall accuracy	$\frac{a+d}{n}$
Sensitivity	$\frac{a}{a+c}$
Specificity	$\frac{d}{b+d}$
Kappa statistic	$\frac{\left(\frac{a+d}{n}\right) - \frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}{1 - \frac{(a+b)(a+c)+(c+d)(d+b)}{n^2}}$
TSS	sensitivity + specificity – 1

	Actually Present	Actually Absent
Predicted Present	*a*	*b*
Predicted Absent	*c*	*d*

Note that all methods discussed so far assume we have absence data

→ common practice is to use background points when we don't

Allouche et al. 2006

1.3 Evaluation Metrics

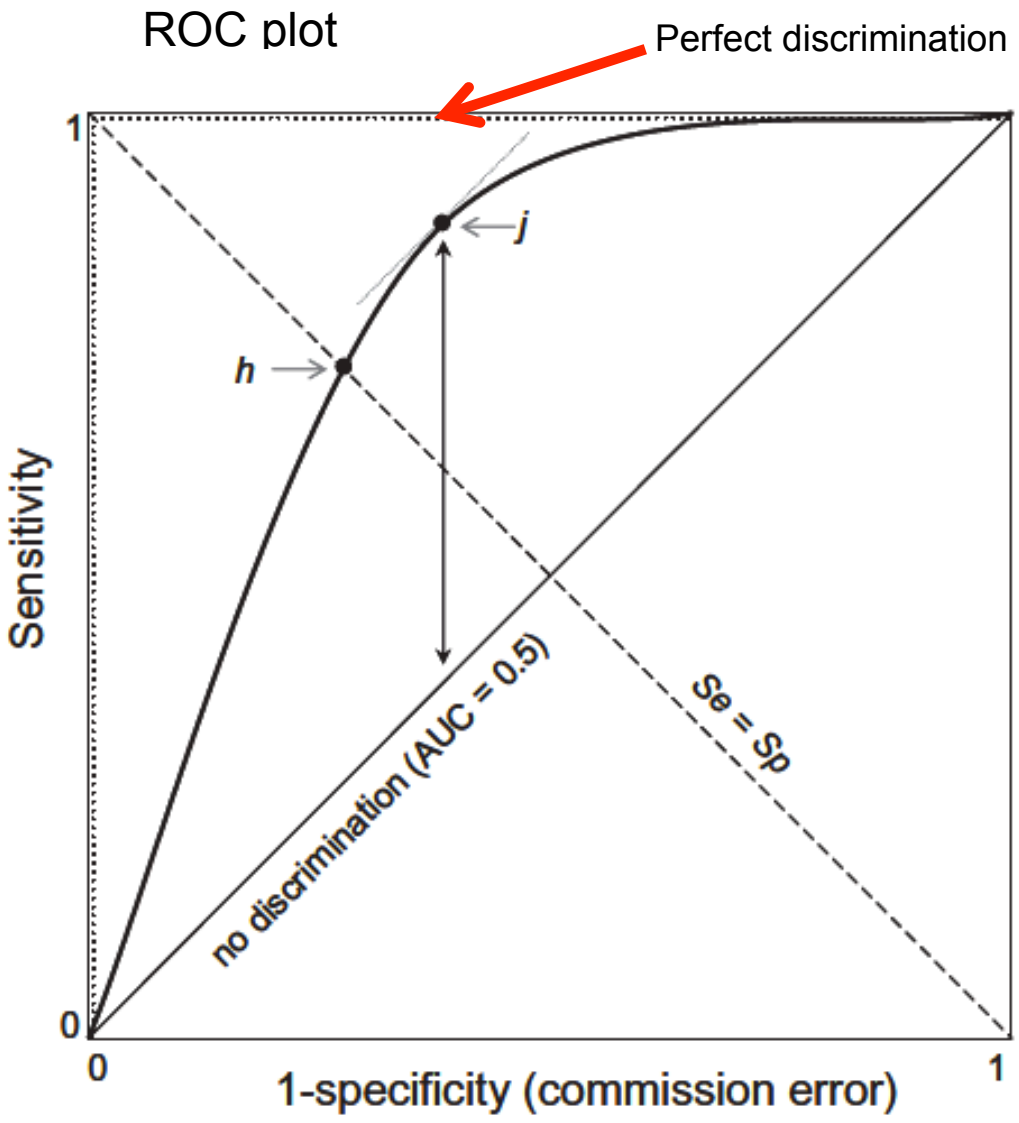
Threshold-independent metrics

AUC = “Area under the curve’; by curve we mean the **Receiver Operator Characteristic (ROC)**

ROC is a plot of false positives (1-specificity) versus true positives (sensitivity) over a range of thresholds from 0 to 1

When absence data are not available, the ROC is derived by plotting sensitivity against the proportion of background locations predicted as presence

1.3 Evaluation Metrics



A high AUC score indicates that our model is insensitive to the threshold used

A perfect model will have AUC = 1

A model that performs no better or worse than random will have AUC = 0.5

Figure from Jiménez-Valverde 2012

1.3 Evaluation Metrics

What values will a good model have?

TSS: ranges from -1 to 1
<=0, model is no better than random
>0.70, considered “excellent”

Kappa: <0.4, considered “poor”
0.4 to 0.75, considered “good”
>0.75, considered “excellent”

AUC: = 0.5, model is no better than random
>0.75, considered “good”
>0.90, considered “excellent”

***** All of this is “rule of thumb”**
***** Major caveat: Just because a model is statistically “good” does not mean it is biologically useful; conversely, a model that does not obtain a high evaluation score may still be informative**

1.4 Thresholds

NOTE: also relevant to making final predictions for sites

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

Table from Peterson et al. 2011

1.5 Exercise

Exercise D3.1

1) Load your model from yesterday. Get the AUC score for the model based on a resubstitution approach.

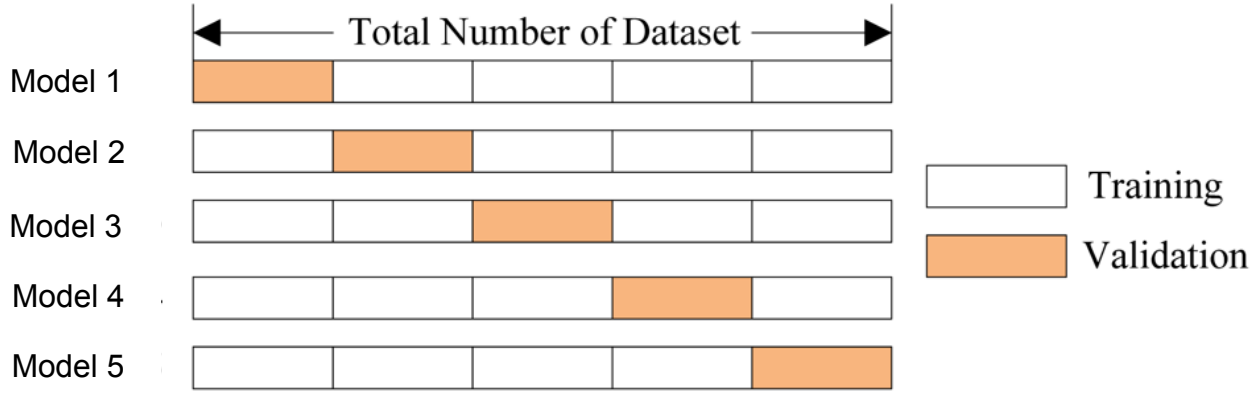
2) Use two common thresholds to convert the continuous model output to binary predictions and calculate TSS and accuracy.

1.6 Different Evaluation Approaches

K-fold Cross Validation

→ Randomly split data into k folds
→ Build k niche models, each time holding out one of the k folds during the calibration stage
→ Evaluate each model with the withheld data
→ Use mean, min, range of AUC values (or other statistic) as indication of model performance
→ The number of folds (k) one should use will depend on dataset size but usually 5 or 10 folds are used
→ Note that all observations are used during model evaluation once and only once

5-fold Cross Validation



1.6 Different Evaluation Approaches

Jackknifing

→ Randomly remove p observations from the dataset and fit the model using n-p observations
→ Evaluate the model with the withheld data
→ Repeat many times (100-1000), sampling without replacement
→ When n is large, not all observations will necessarily be used for model evaluation; when n is smaller, some observations will be used multiple times during model evaluation
→ Similar to k-fold cross-validation in that you get mean, min, range of evaluation statistics

1.6 Different Evaluation Approaches

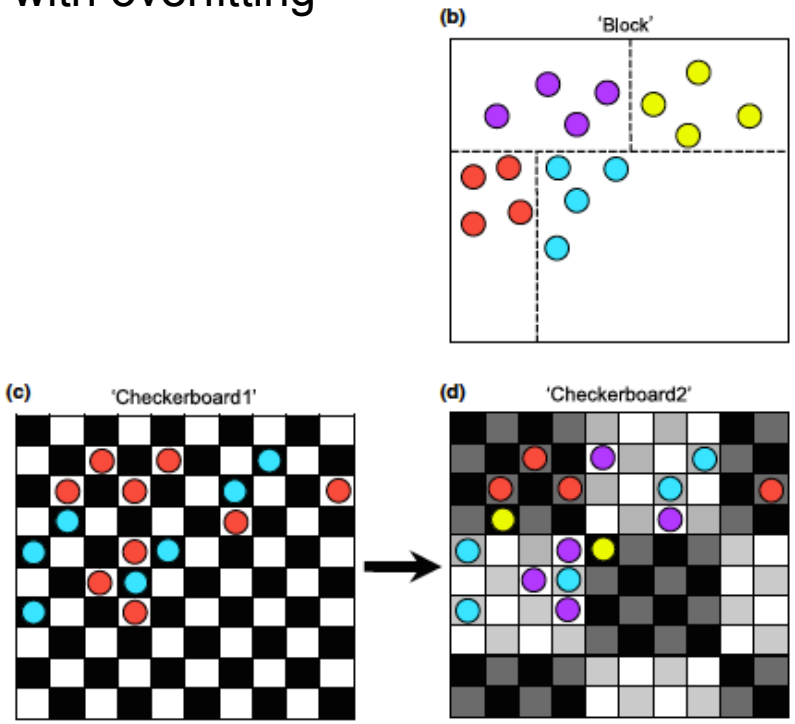
Bootstrapping

→ Randomly split data into 2 partitions (e.g. 80:20)
→ Build models with the larger partition, evaluate with the smaller partition
→ Repeat many times (500-1000), sampling with replacement
→ Again you will get information about the potential distribution of the test statistic

1.6 Different Evaluation Approaches

Spatially Structured Partitioning

→ Rather than randomly splitting the data for testing and evaluation, we split into different spatial subsets
→ Models are calibrated holding out one subset each round to serve as the testing data
→ Helps remove some of the effects of sampling bias on evaluation and helps detect issues with overfitting



Three different approaches for creating spatial subsets

Muscarella *et al*. 2014

1.6 Different Evaluation Approaches

Null Model Approach

→ Method of Raes and ter Steege 2007
→ Build model with all localities and calculate your test statistic (observed value)
→ Choose random points across the study region, build a model based on these points and calculate the test statistic using some subset of your locality records
→ Repeat the random model building 100-1000 times to get a distribution of random model test statistics
→ Compare your observed test statistic to the above distribution; if observed is in the 95^{th} percentile tail of the distribution, model performs better than random
→ Useful if sample sizes are small
→ Speaks more specifically as to whether the localities are “useful”
→ As originally formulated, relies on resubstitution validation but can combine with data-splitting approach to convert to internal validation approach

1.7 Exercise

Exercise D3.2

1) Try rebuilding the model to test the effects of different features and settings of the regularization parameter.

2) Used 5-fold cross-validation to evaluate the final tuned model.

2. Model Predictions and Limitations

2.1 Maxent Output

2.2. Exercise

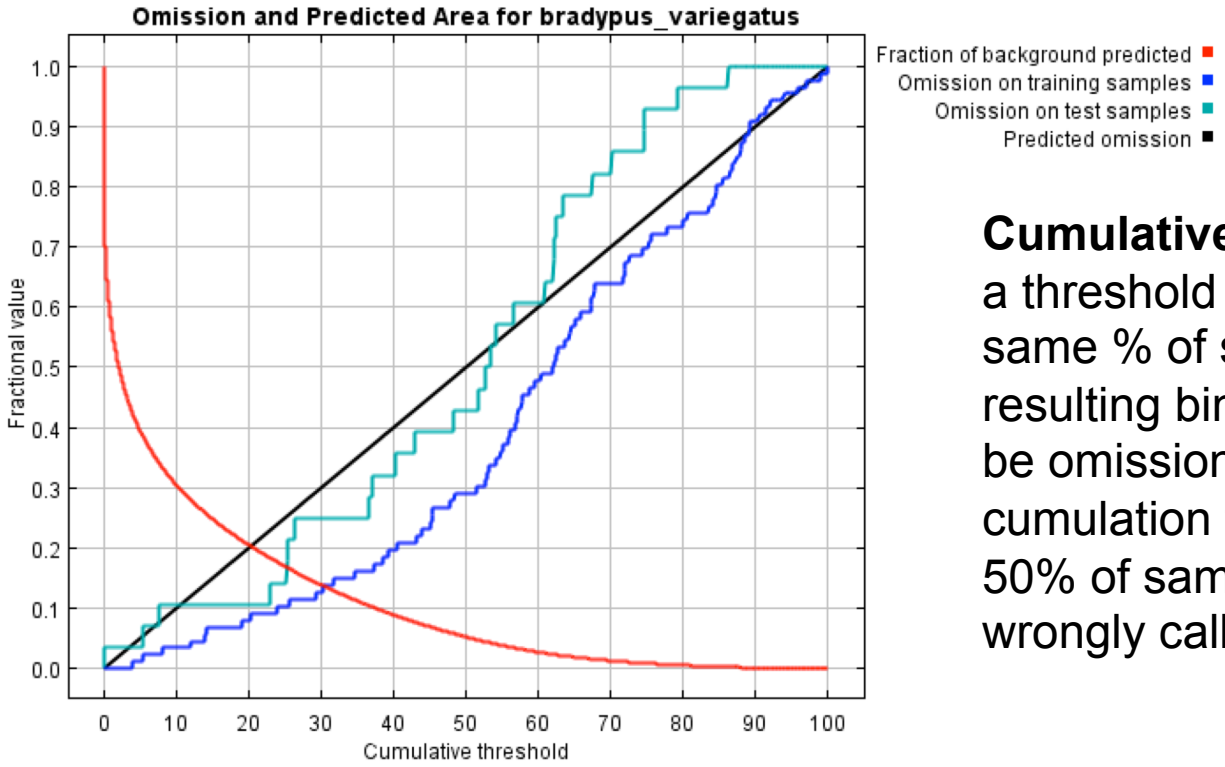
2.3 Dealing with Extrapolation

2.4 Exercise

2.1 MAXENT Output

The Omission and Predicted Area plot

- How the fraction of background predicted as suitable and omission based on the training presence data (and testing data if you used it) vary with the cumulative threshold
- Observed omission should match predicted omission

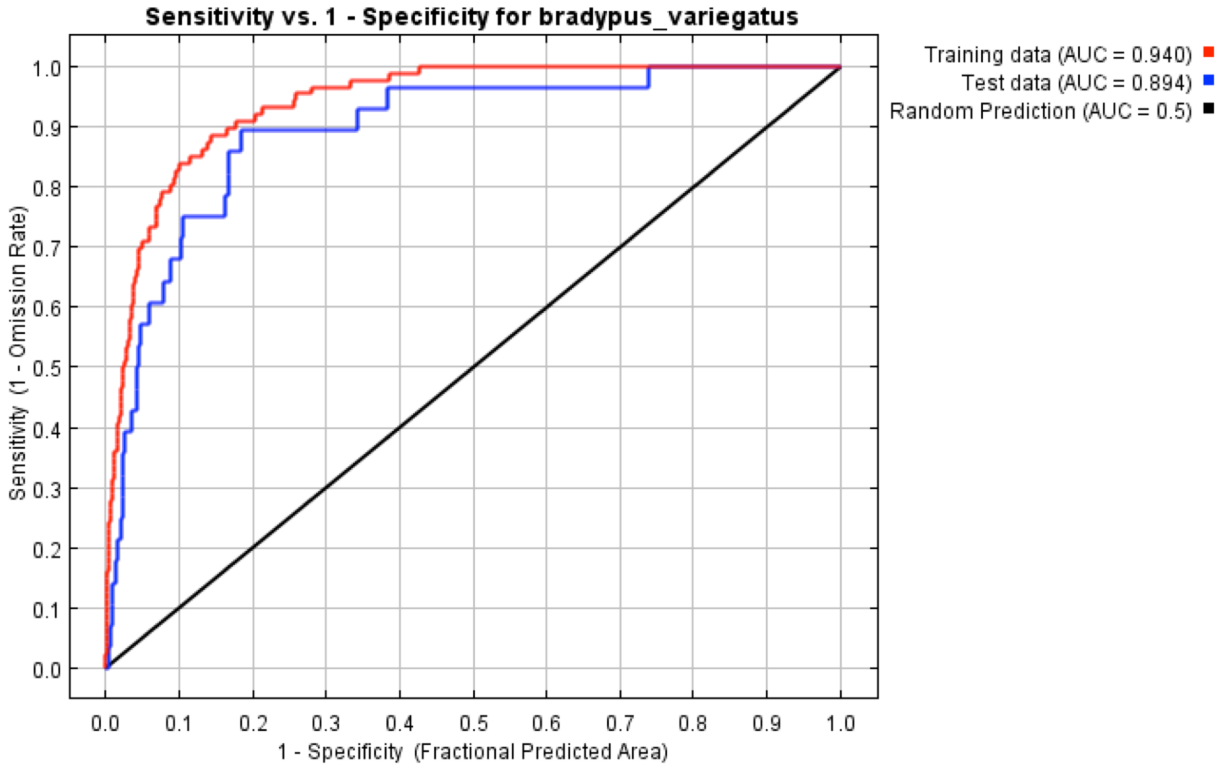


Cumulative threshold = if we set a threshold of this much, then the same % of samples drawn from the resulting binary distribution would be omission errors (so a cumulation threshold of 50 means 50% of samples expected to be wrongly called as absences)

2.1 MAXENT Output

The ROC plot

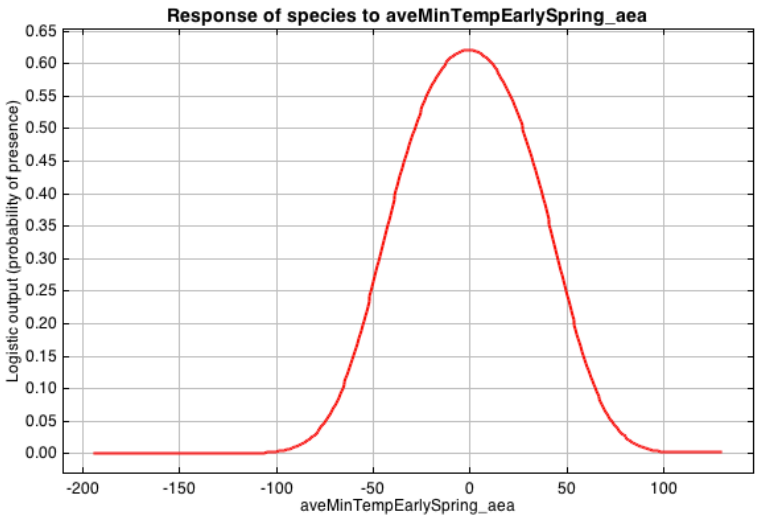
- Based on training (and testing if provided) data
- Also gives AUC
- The closer to the upper left hand side of the graph the testing ROC plot is, the better the model is at predicted the testing presences

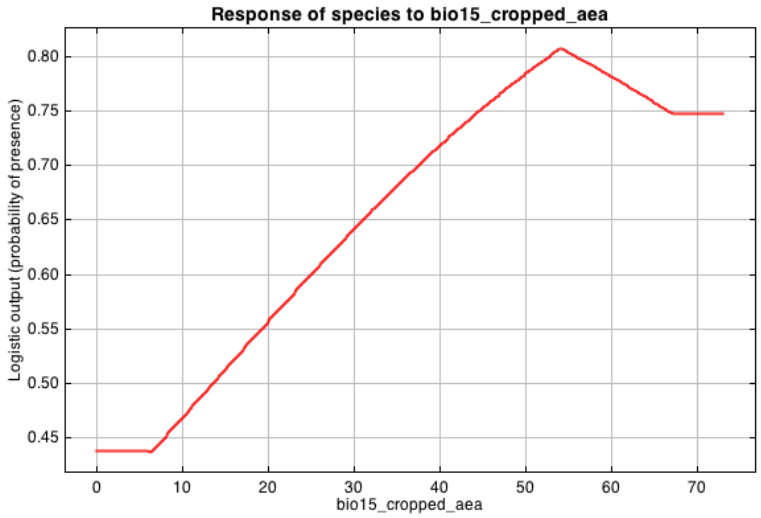


2.1 MAXENT Output

Response Curves (one for each variable in the model)

- Predicted probability of presence (log) of the species as a function of the variable in question with all other variables held at their average value at known presences
- Gives a sense of how a given variable affects the suitabiltiy of sites but is not independent of the other variables; if you have highly correlated variables, these plots may be very affected
- Remember clamping!!





2.1 MAXENT Output

Measures of Variable Importance

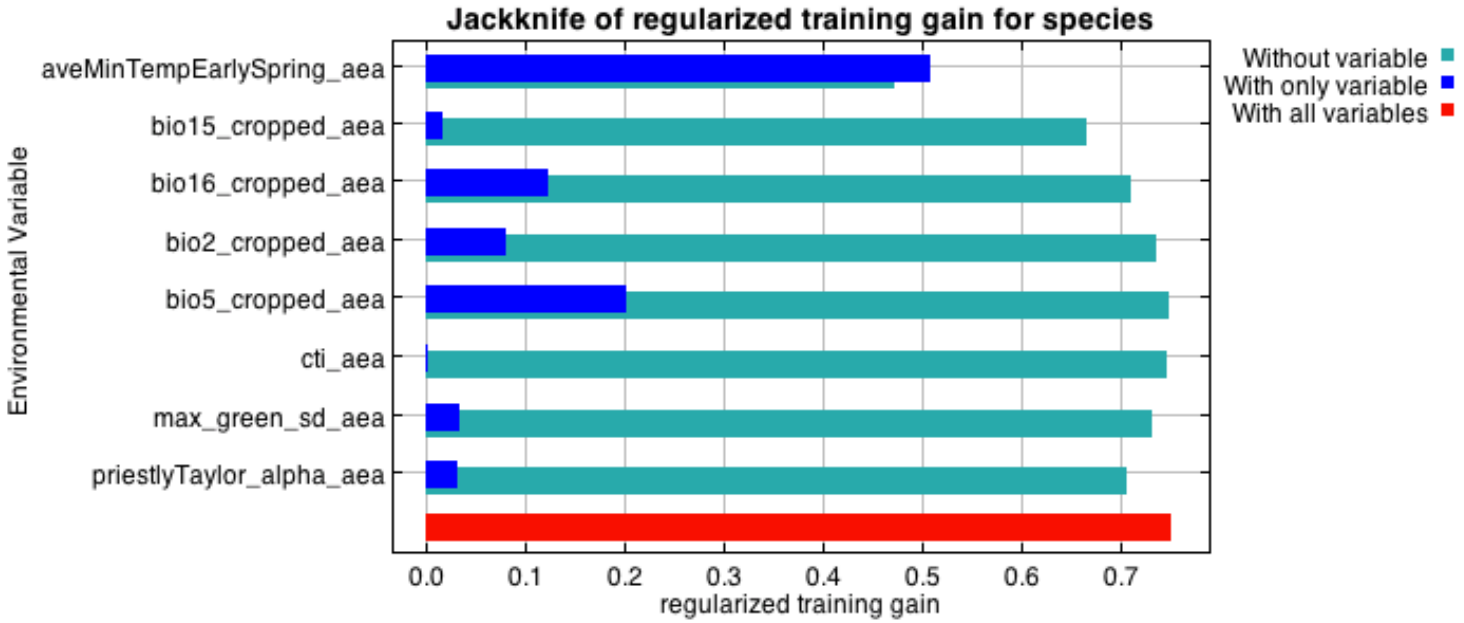
- Maxent continuously increases regularization gain by changing the coefficient of a single feature; it keeps track of the variable upon which that feature depended and at the end of calibration, calculates the percentage of times the variable contributed to an increase in gain; this is the **Percent Contribution** → depends on path during optimization
- **Permutation Importance** measures the drop in gain when values of each variable are randomly permuted across the training presence and background points → depends only on final model

Variable	Percent contribution	Permutation importance
aveMinTempEarlySpring_aea	62.9	44
bio5_cropped_aea	11.7	1.2
bio16_cropped_aea	9	25.6
bio15_cropped_aea	5.9	6.3
priestlyTaylor_alpha_aea	4.5	18.2
bio2_cropped_aea	3.4	2.6
max_green_sd_aea	2.4	1.9
cti_aea	0.1	0.1

2.1 MAXENT Output

Measures of Variable Importance

- If MAXENT's jackknife default is used, it will run multiple models, asking about the drop in regularization gain when each variable is excluded and amount of gain when models are based on each variable in isolation



2.1 MAXENT Output

Warning! The default outputs in dismo do not provide you will the files that the default settings in the MAXENT GUI provide. You must tell dismo what you want!

2.3 Exercise

Exercise D3.3

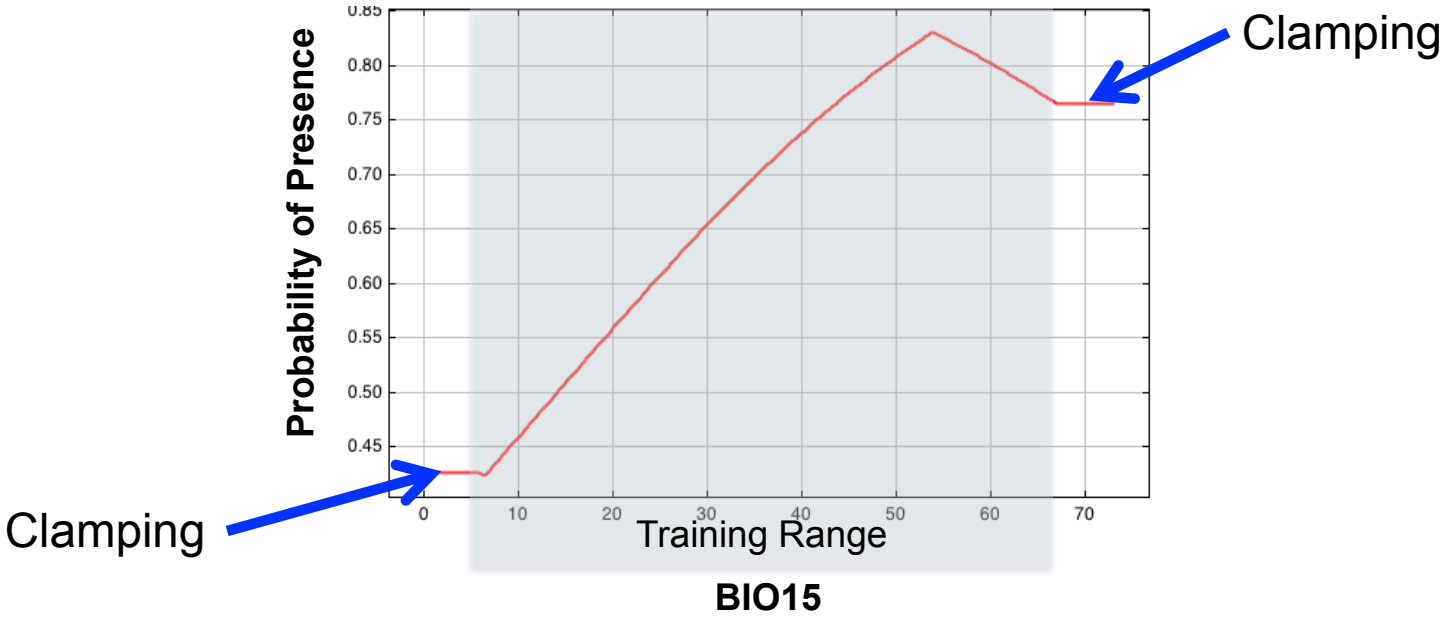
1) Examine the variable importance plots generated by MAXENT for your tuned model.

2) Examine how the probability of presence varies with each variable.

2.4 Dealing with Extrapolation

In statistics in general, we want to be careful about extrapolating results beyond the range of the original observations

MAXENT deals with this by clamping…but we still don't know what relationships with the environment might look like outside of the training data range



2.4 Dealing with Extrapolation

Multivariate Environmental Similarity Surfaces (MESS)

- A simple method for assessing where environments are novel

If we have a set of predictor variables (V_1, V_2…), we can ask how similar a given point, P, is with respect to predictor values in our training dataset:

The MES of a point P is calculated as follows:

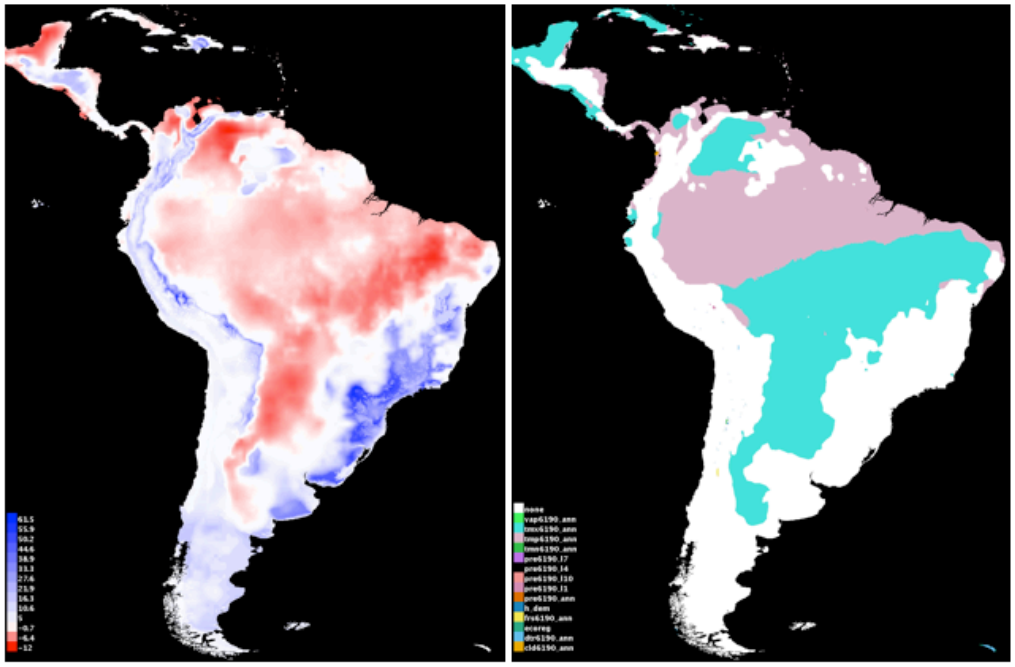
1. Let min_i be the minimum value of variable V_i over the reference point set, and similarly for max_i.
2. Let p_i be the value of variable V_i at point *P*.
3. Let f_i be the percent of reference points whose value of variable V_i is smaller than p_i.
4. Then the similarity of *P* with respect to variable V_i is:
   - $(p_i - min_i) / (max_i - min_i) * 100$ if $f_i = 0$
   - $2 * f_i$ if $0 < f_i \leq 50$
   - $2 * (100 - f_i)$ if $50 \leq f_i < 100$
   - $(max_i - p_i) / (max_i - min_i) * 100$ if $f_i = 100$
5. Finally, the multivariate similarity of *P* is the minimum of its similarity with respect to each variable.

Elith *et al*. 2010

2.4 Dealing with Extrapolation

Multivariate Environmental Similarity Surfaces (MESS)

- MAXENT (at least the GUI) can plot the MES scores across our desired prediction extent to get a MESS map showing novel conditions (anywhere values are negative) → avoid making predictions here?
- It will also output a map showing which the most dissimilar variable driving the negative MES scores



2.4 Dealing with Extrapolation

Multivariate Environmental Similarity Surfaces (MESS)

- Note this approach is limited: will not identify changes in correlations between variables outside of the training range! And recall MAXENT is often using those correlations

- Does not handle categorical variables (landcover etc)

- Other options?

2.5 Exercise

Exercise D3.4

1) Follow the scripts to manually generate a MESS map.

2) Use the MESS map to plot areas of North America where our model would definitely have to extrapolate to make predictions.

3. Presenting Results

What goes in the Methods Section:

- List of variables and their sources, their resolution, the projection used, any manipulations you did to the layers
- Estimate of error in the occurrence records
- List any parameter settings that you changed and specify “default” settings for the rest
- Any assumptions made during data preparation and choice of algorithm/ parameters

What Results to report:

- Evaluation metrics (AUC) in text
- Most important variables (if relevant to questions) with their percent importance

Considerations for Figures:

- Mind basic mapping rules: Include a scale bar, north arrow, names of major landscape features, an inset with a broader extent if you study region is small
- ROC plots, variable gain/loss plots etc. probably don’t belong in the main text

3. Presenting Results

Exercise D3.5

1) Generate a prediction surface based on your final model.

2) Use the MESS map to mask out areas where we may want to avoid making predictions.

3) Optional: See if you can make a final predictions map that is nicer looking than this!

That’s a wrap!
You now have the tools to
Go Forth and Model!