Day 2: Model Calibration

1. Additional data considerations

2. Niche Modeling algorithms

3. Introduction to MAXENT

4. Basic MAXENT niche modeling

1. Additional data considerations

1.1 More on Sampling Bias

1.2 Dealing with Sampling Bias

1.3 Exercise

1.4 Variable selection

1.5 Exercise

1.1 More on Sampling Bias

How probable is it that all parts of this landscape were surveyed with equal frequency?

1.1 More on Sampling Bias

- Sampling bias occurs in **<u>geographic space</u>** but in turn often results in a bias in **<u>environmental space</u>**

- Leads to two related problems:

  1) Occurrence data that over-represent some conditions and under-represent other conditions where the species could occur may mislead our models in terms of what conditions are considered suitable for the species

  2) localities that are geographically “clumped” are likely to suffer from spatial autocorrelation → leads to issues with model evaluation

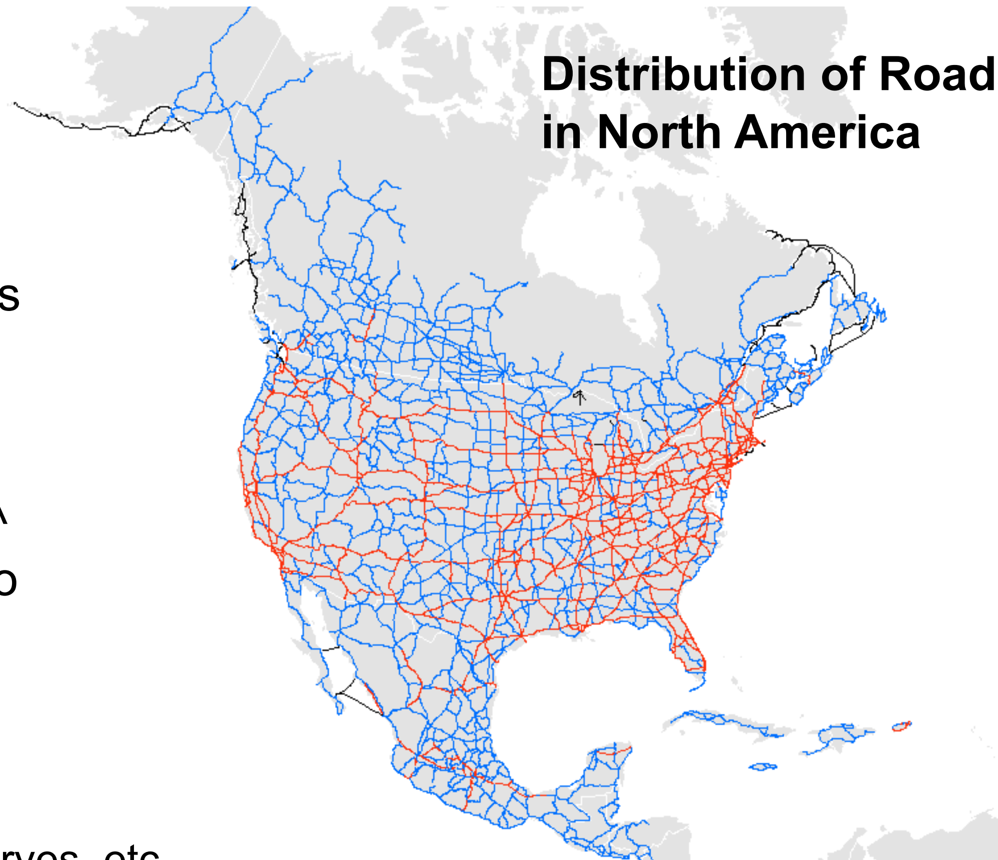
- Most methods for minimizing sampling bias directly address the first of these concerns and may or may not reduce SA

1.1 More on Sampling Bias

Bias in environmental conditions:

For example, surveys that are restricted to roads^ may not cover all parts of a species' range (some conditions where the species is found not represent)

Especially if one considered that the placement of roads^ is not random with respect to geography and thus environmental space



^ shipping lanes, reserves, etc.

http://www.naturalearthdata.com/tag/roads/

1.1 More on Sampling Bias

Spatial autocorrelation (SA):

→ Definition: Locations close to each other in geographic space exhibit more similar environmental values than those further apart

→ Problem in standard statistical terms: violates assumption that residuals are independent and identically distributed which can in term violate parameter estimates and increase the change of rejecting the null hypothesis (Type I error)

→ In niche modeling: if occurrence records used to build niche models are closer to occurrences used for evaluating the model than they are to background points used for evaluating the model (i.e. if sampling is spatially biased), then our model will tend to predict the testing presences but not the absences just because of SA alone (a bad model may look good from a statistical standpoint)

1.2 Dealing with Sampling Bias

Thinning Localities

→ based on geographic space and/ or environmental space
→ reduces over-representation of some environments

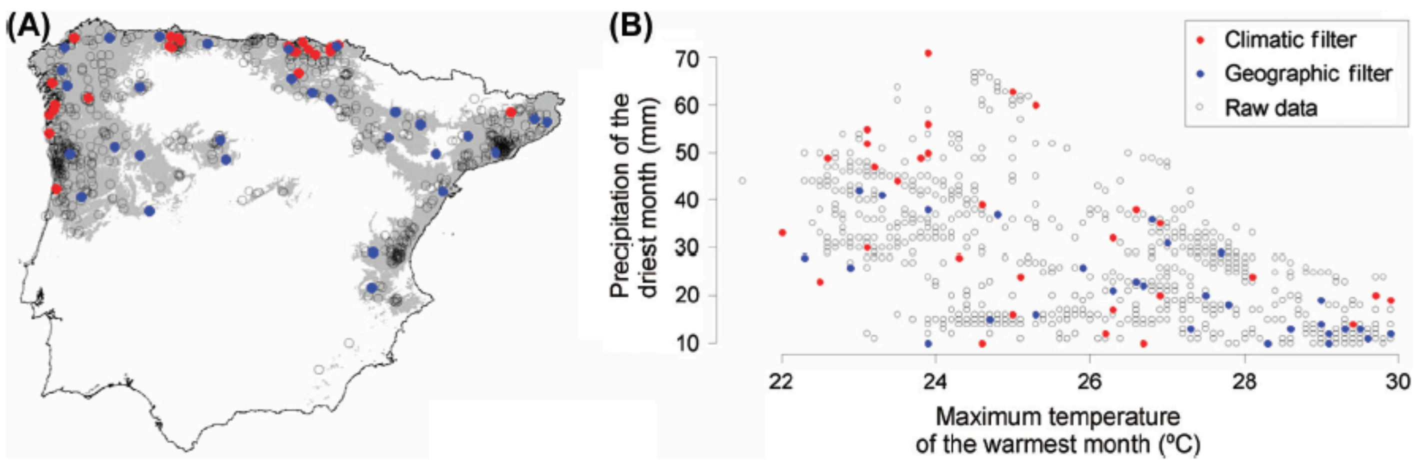


Figure 4. Differences between geographic (blue points) and climatic (red points) filters applied to the data sets in geographic space (A) and in environmental space (B), the latter shown for the two climatic variables used in the environmental filtering. Here, we provide examples for one experiment (bias of the raw data set: distance to populated areas; sample size for the filtered data set: 30 points).

Research

Ecography 37: 1084–1091, 2014
doi: 10.1111/j.1600-0587.2013.00441.x
© 2014 The Authors. Ecography © 2014 Nordic Society Oikos
Subject Editor: Catherine Graham. Accepted 12 September 2013

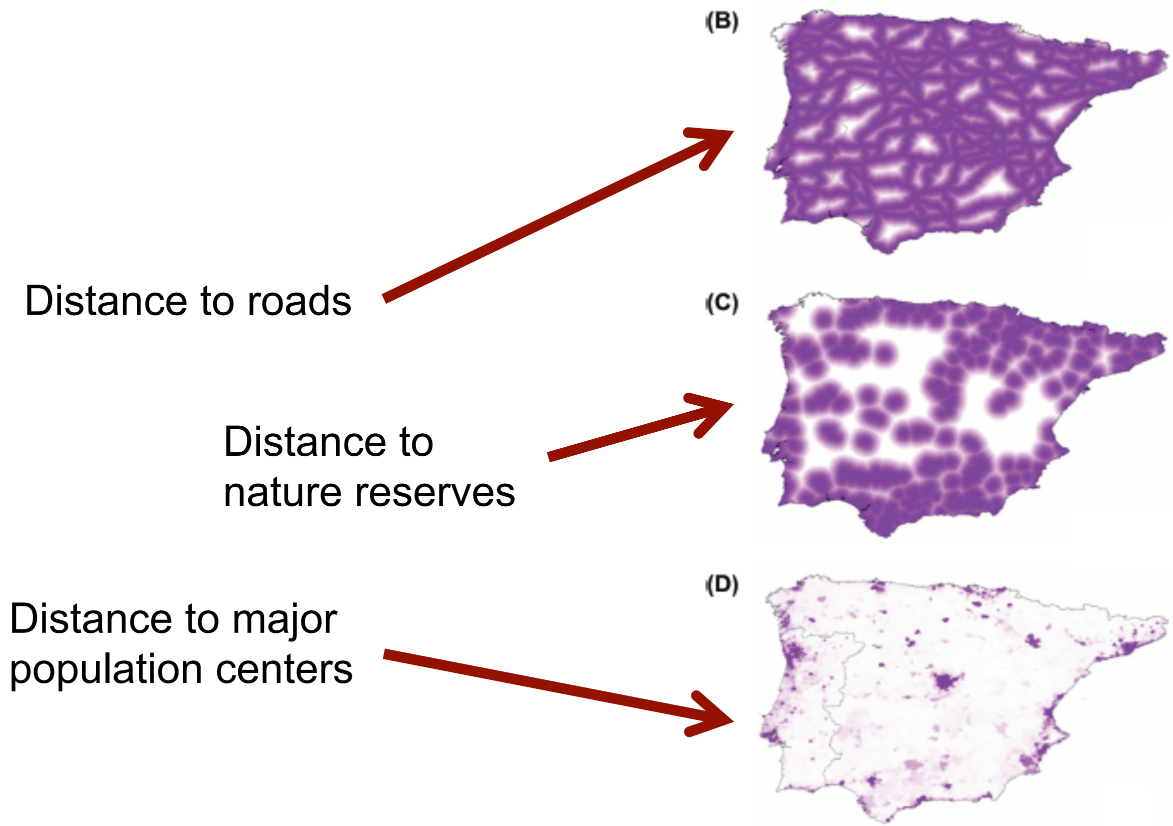
Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models

Sara Varela, Robert P. Anderson, Raúl García-Valdés and Federico Fernández-González

1.2 Dealing with Sampling Bias

Use of a Bias Grid for selecting background points

→ choose background points with the same bias as the occurrence records
→ reduces variation in background points relative to occurrences
→ avoids overpredicting presences



1.2 Dealing with Sampling Bias

Related: The extent from which background points are sampled matters

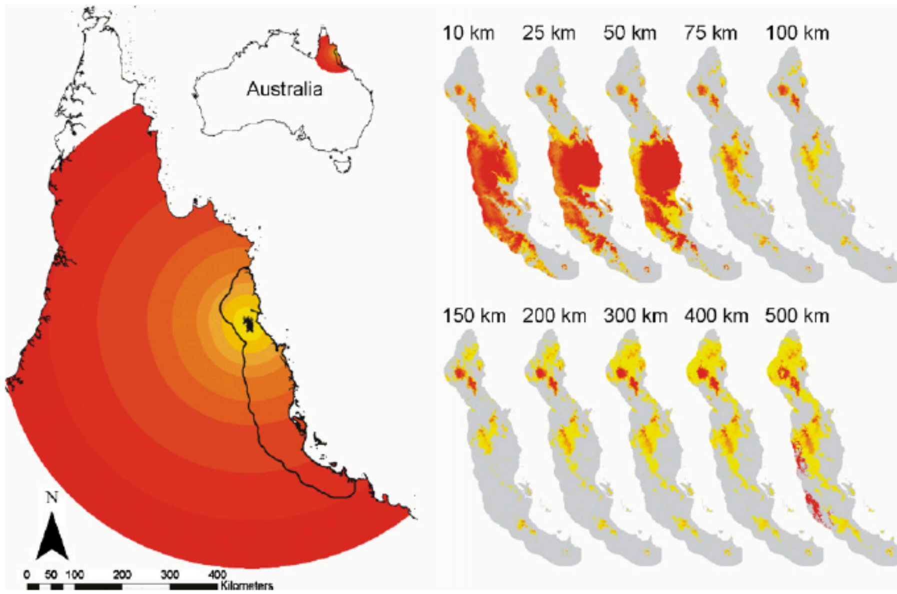
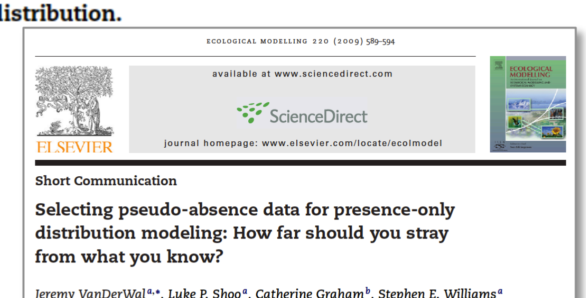


Fig. 1 – Backgrounds from which pseudo-absences were drawn for C. *hosmeri* and predicted distributions in the Australia Wet Tropics region (outlined in black) given the different background sizes. Increasing background size corresponds to darkening of buffering bands surrounding the occurrence points of the species (represented as x symbols here) in the left image; these regions represent increasing distances from 10 to 500 km from the occurrence points. Warmer colors on the right images infer greater predicted suitability for the C. *hosmeri*. Grey areas fall below the threshold of suitability and are assumed to not be part of the distribution.


ECOLOGICAL MODELLING 220 (2009) 589–594
available at www.sciencedirect.com
ScienceDirect
journal homepage: www.elsevier.com/locate/ecolmodel

Short Communication

Selecting pseudo-absence data for presence-only distribution modeling: How far should you stray from what you know?

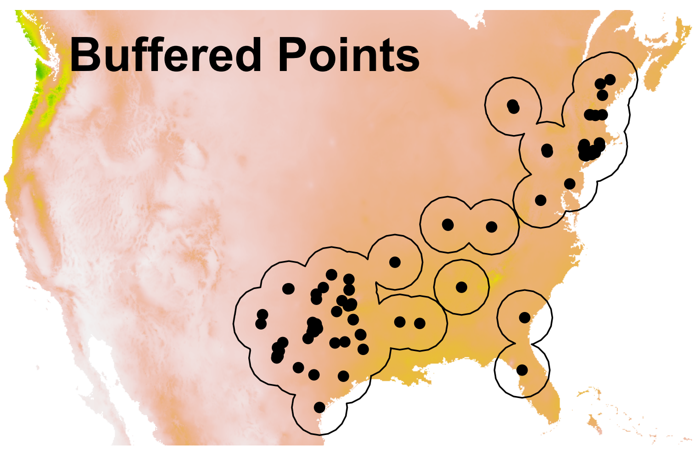
Jeremy VanDerWal[a,], Luke P. Shoo[a], Catherine Graham[b], Stephen E. Williams[a]*

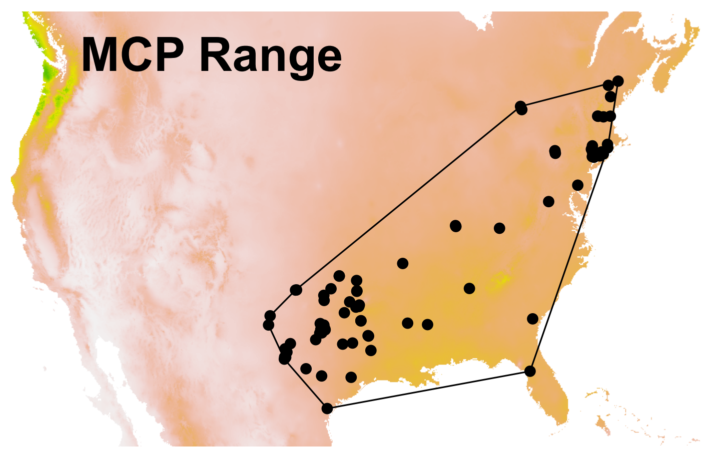
1.2 Dealing with Sampling Bias

In addition to considering sampling bias, we ideally want to sample background sites from areas the species has been able to “test the waters”

Examples

- **Buffered Points**: Area immediately around sites; for instance, within the dispersal distance of the species
- **Range Extent**: Based on the idea that the species has at least covered this extent during colonization and had a chance to sample sites over distances consistent with it’s range
- **Ecotone Boundaries**: Assumes that the species could have reasonably tested all areas in the ecological zones it occupies.





1.2 Dealing with Sampling Bias

Other Suggestions:

→ use locality information from closely related taxa as background points; surveys of related taxa more likely to have the same sampling bias as focal taxa (Phillips et al. 2009)

→ to reduce impact of spatial autocorrelation on evaluation statistics, ensure each presence point in the evaluation dataset has a non-presence point that is equally distant from the closest occurrence record used to build the model (Hijimans 2012)

→ directly correct for spatial autocorrelation in models (e.g. include autocorrelation term) (Dormann et al. 2007)

1.2 Exercise (60 minutes)

Exercise D2.1

Using the scripts provided on the website:

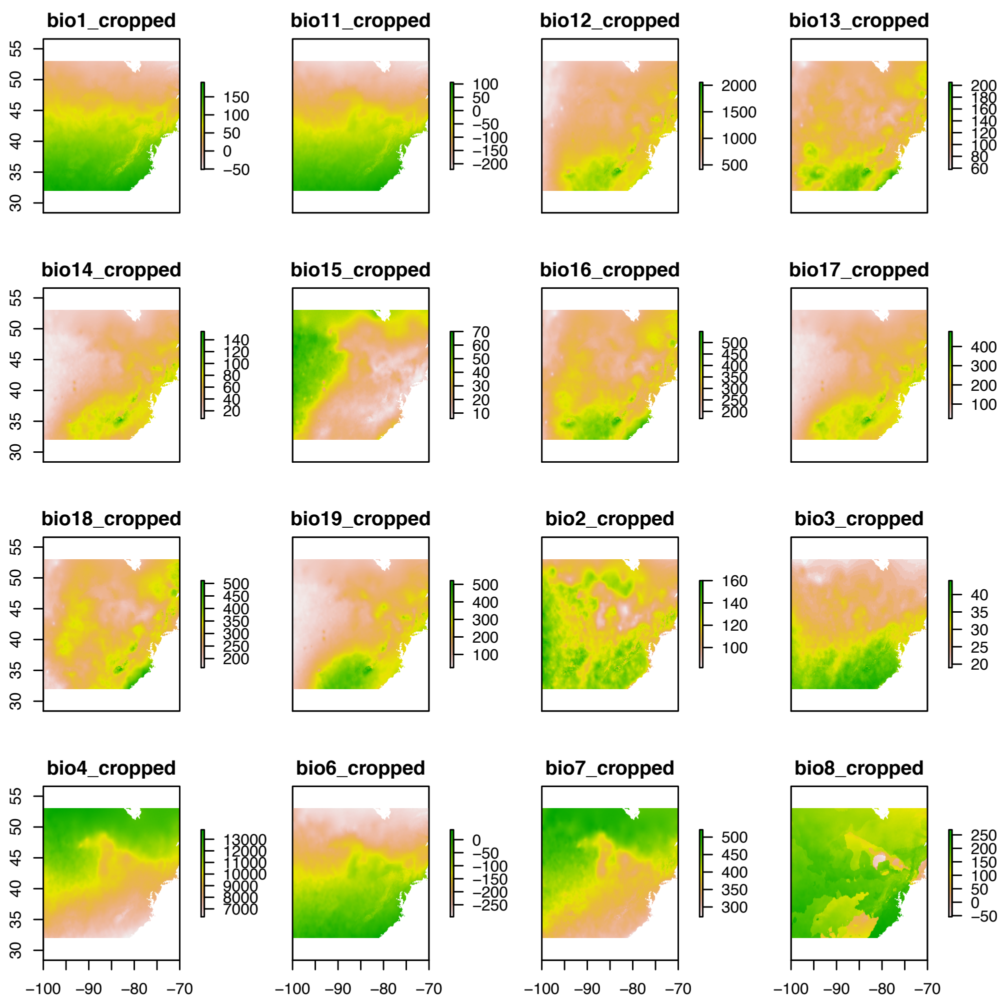
1) Examine the distribution of our frog localities in geographic and environmental space for any two of the environmental variables.

2) Thin the locality records based on geography and then environment. Compare the thinned localities in both cases to the full dataset in environmental space.

As a class:

3) What methods (if any) do we want to use for dealing with any sampling bias?

1.2 Variable Selection



The problem:

16 Bioclim layers
+ 10 Soil layers
+ 1 land cover layer
+ …etc.
= >27 variables in the model!

What if we only have 30 locality records? Imagine building a linear model with 30 data points and 27 predictors! (NOTE: in MAXENT, these get converted into potentially many more features)

Rule of thumb: minimum of ten data points for every variable

Some variable selection necessary! But how to choose?

1.2 Variable Selection

How to choose?

→ **Above all: variables should be biologically relevant!**

→ Remove correlated variables (more important for regression-type methods; less important for MAXENT but will make it hard to tease apart effects of specific variables)

→ Transform your variables (i.e. using PCA) to reduce dimensionality and correlations between variables

→ Use statistical model selection approaches (AIC, Random Forests, tuning in the niche modeling algorithm itself)

Exercise (30 minutes)

Exercise D2.2

Using the scripts provided on the website:

1) Using both the information I gave you about the system and a correlation analysis, decide on at least three biologically relevant variables for inclusion in your niche models

2. Niche Modeling Algorithms

2.1 Presence-only

2.2 Presence-absence

2.3 Presence-background

2.4 Presence-pseudoabsence

2.5 Multi-model approaches

2.1 Presence-only methods

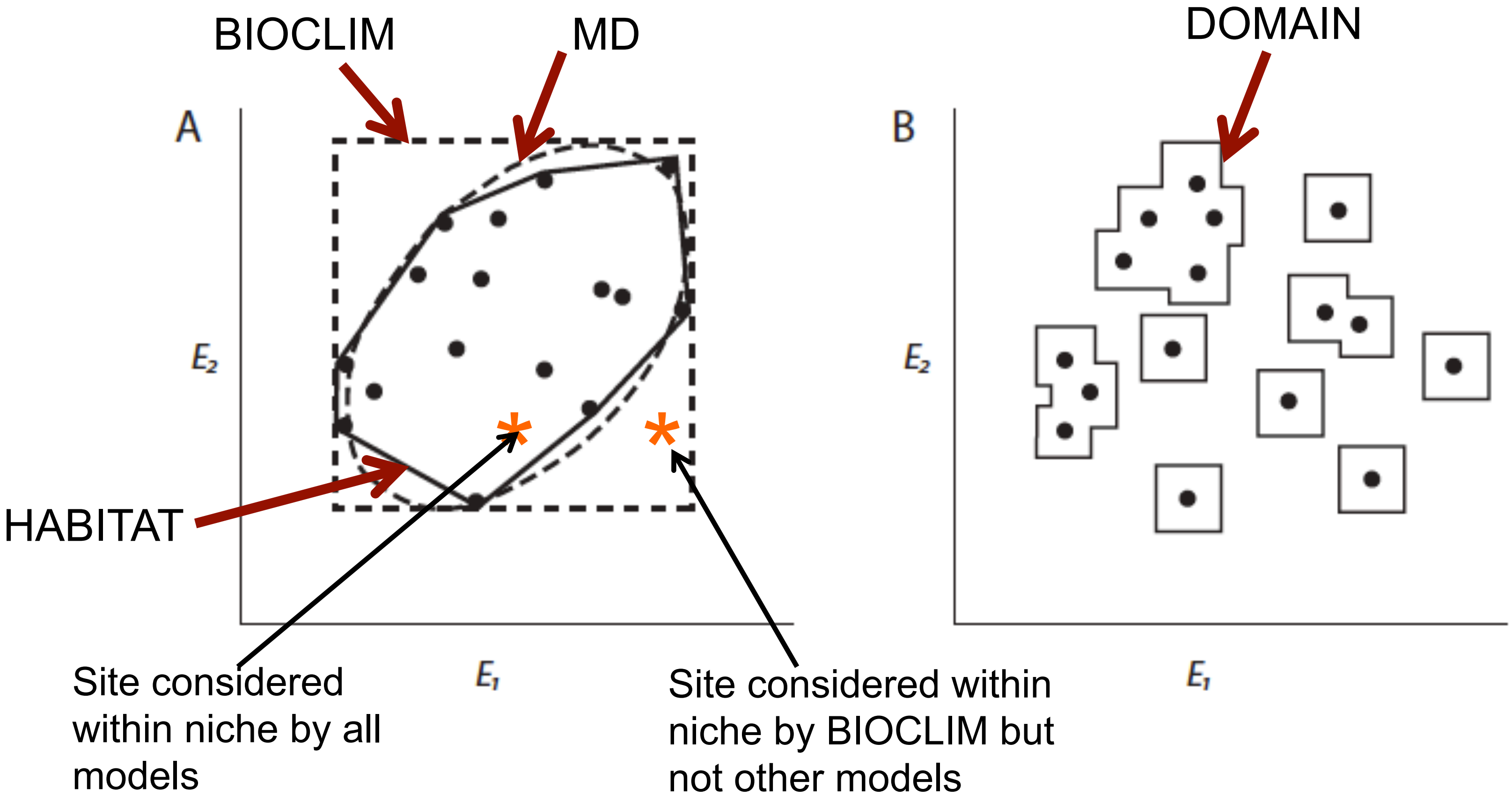
- Require only occurrence records
- Involve fitting “environmental envelopes” that encompass conditions associated with occurrence localities in E-space or calculating environmental distances between locations of interest and the occurrence records

2.1 Presence-only methods

- **Envelop Methods**
  - BIOCLIM: Rectangular envelop that spans the minimum and maximum range of values occupied by a species across all variables
  - HABITAT: Minimum convex polygon around occurrence records in E-space
- **Distance Methods**
  - Mahalanobis distance (MD): Difference between location of interest and mean of occurrence records in multidimensional E-space
  - DOMAIN: Uses Gower distances to evaluate the proximity of a location of interest to any occurrence record

NOTE: This is not an exhaustive list…..

2.1 Presence-only methods



NOTE: in all cases we had to make a decision as to "threshold" to use to draw the polygons (e.g. for BIOCLIM: do we use full range of values of occurrence records or the 90^{th} percentiles? For DOMAIN, what constitutes being "proximate" to an occurrence record?)

Figure from Peterson et al. 2011

2.2 Presence-absence methods

- Best for predicting area occupied; less suitable for predicting potential distribution (e.g. species' distribution modeling but maybe not the best for niche modeling)

- Include regression methods, classification and regression trees and machine learning techniques

- Caution: true absences are hard to collect! (think detectability issues)

- Sometimes these method are used when only pseudoabsences or background data are available → be careful doing this! Need to correct estimates of occurrence probability…

2.2 Presence-absence methods

- **Regression Methods**
  - GLM: like least-squares regression but shape of relationship is described by a “link function” and different distributions can be used to describe the error (e.g. normal, binomial etc.)
  - GAM: extensions of GLM (also use link functions) but make fewer assumptions regarding the shape of the function so complex, non-linear relationships can be fit
  - Multivariate adaptive regression splines (MARS): fit a series of linear segments to the data

2.2 Presence-absence methods

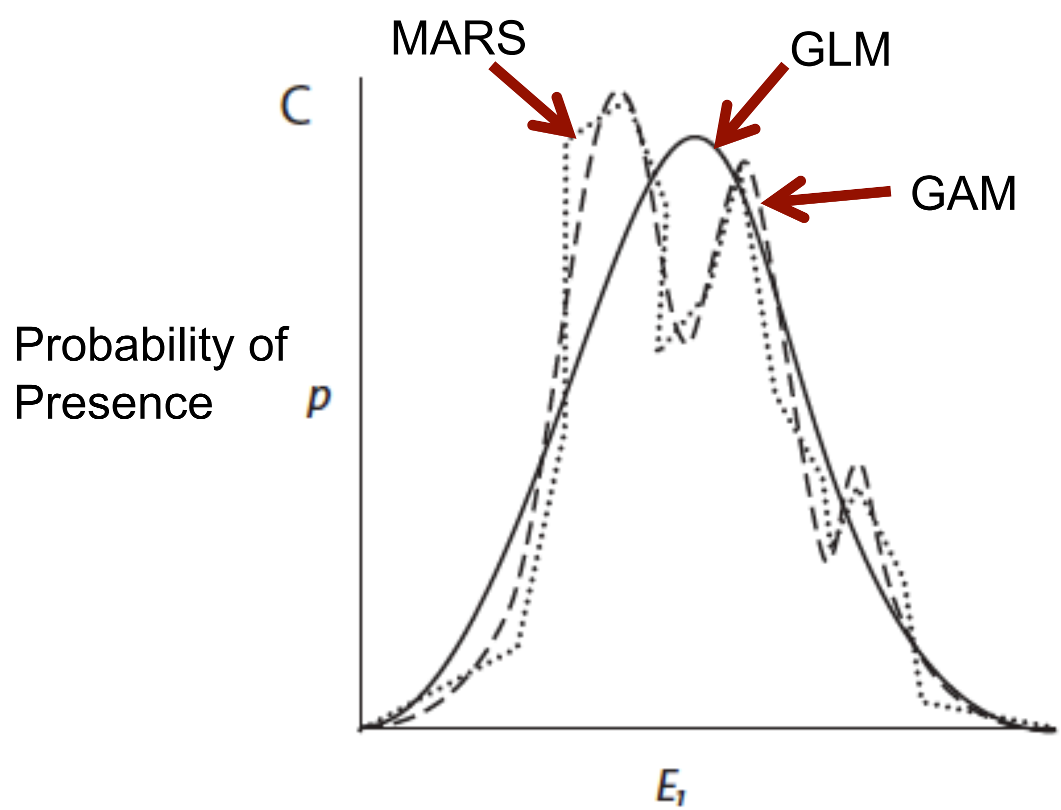
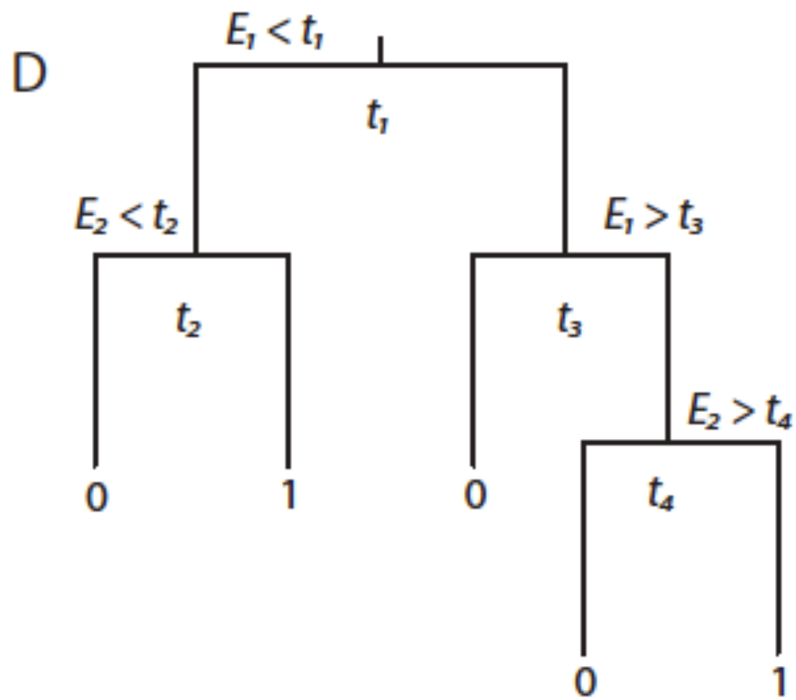


Figure from Peterson et al. 2011

2.2 Presence-absence methods

- **Classification and Regression Trees**
  - RANDOM FORESTS: repeatedly partition the data into groups based on combinations of the environmental variables so that you minimize the diversity of categories (presence or absence) within groups
  - Boosted regression trees (BRTs): fit many classification trees and then combine them
- **Machine-Learning Approaches**
  - Include genetic algorithms, artificial neural networks, support vector machines

2.2 Presence-absence methods



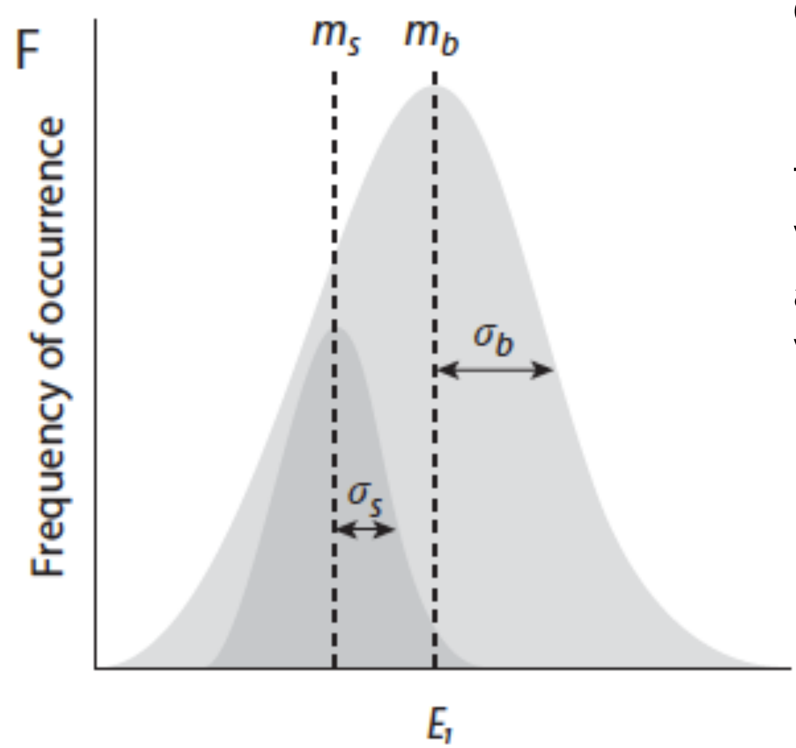
Environmental variables (E_1 and E_2) are divided at different split points (t_1, t_2 etc) to classify suitability along each branch

Figure from Peterson et al. 2011

2.3 Presence-background methods

- Incorporate information about the environment at large but don’t require absence information
- MAXENT (which we will talk about in-depth)
- Ecological Niche Factor Analysis (ENFA): an ordination technique that compares environmental conditions at occurrence records to the distribution of values across the background

2.3 Presence-background methods



ENFA compares species occurrence points to background points both in terms of differences in the mean value for each variable (marginality) and differences in variance (specialization)

Figure from Peterson et al. 2011

2.4 Presence-pseudoabsence methods

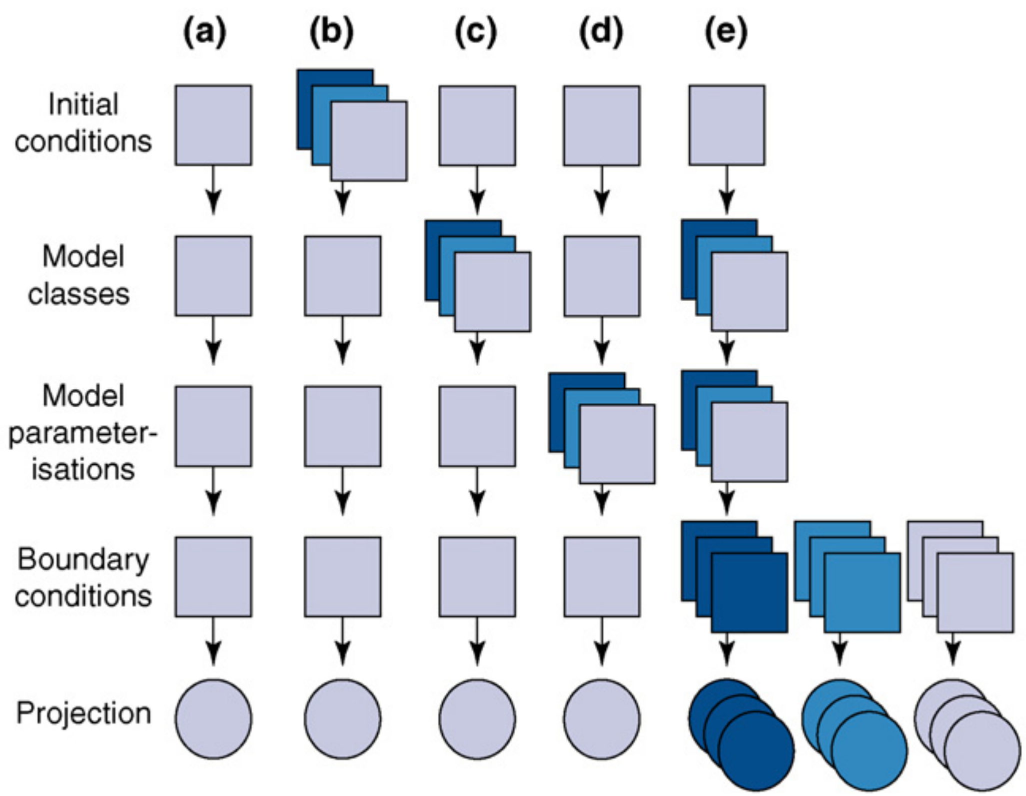
- Data are similar to presence-background methods except, unlike background points which may include some locations (grid cells) where the species is present, “pseudo-absences” are drawn from locations lacking presence records

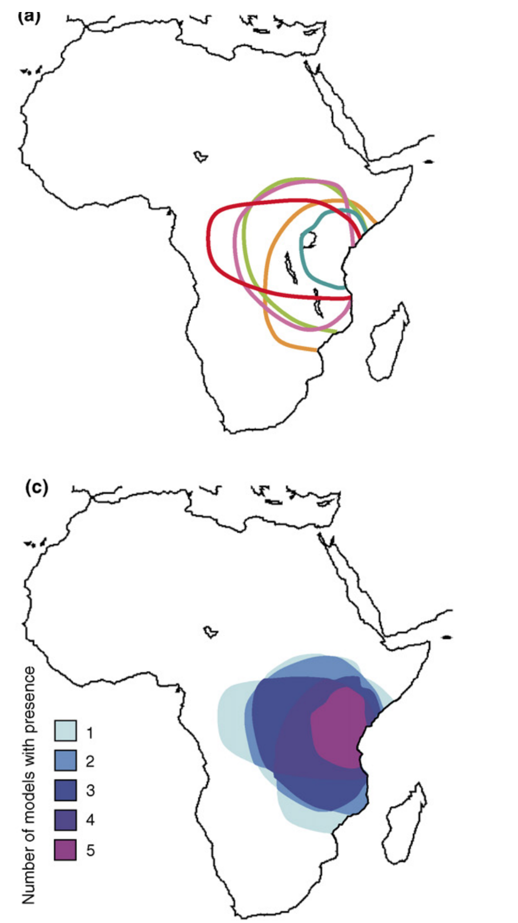
- GARP: combines climate envelops and GLMs in a genetic algorithm framework

2.5 Multi-model approaches

When in doubt, try everything?

→ Build many models using different approaches and look for a consensus (e.g. ESEMBLE forecasting)





Araújo and New, *TREE*, 2007

2.5 Multi-model approaches

Caution needed when using ensemble forecasting!

- When ensemble forecasting involves combining different modeling methods (i.e. when using the BIOMOD package), note that each has its own set of parameters and assumptions that you need to consider and evaluate

- Also, different methods may estimate different things: be sure you are comparing apples to apples

- A caveat: bunch of crappy models do not make a good one (ENSEMBLE modeling can't help you if you have poor-quality data)! And when taking the consensus of many models, if you have many bad ones, they may overwhelm the signal from the few good ones

AN IMPORTANT NOTE: even when working with one modeling approach, expect to run multiple iterations, not necessarily for purposes of generating consensus predictions, but to look at the sensitivity of your results

3. Introduction to MAXENT

3.1 Overview

3.2 Key Settings

3.3 Parameter Tuning

3.1 Overview of MAXENT

MAXENT stands for Maximum Entropy

Widely used software in the literature

Advantages:

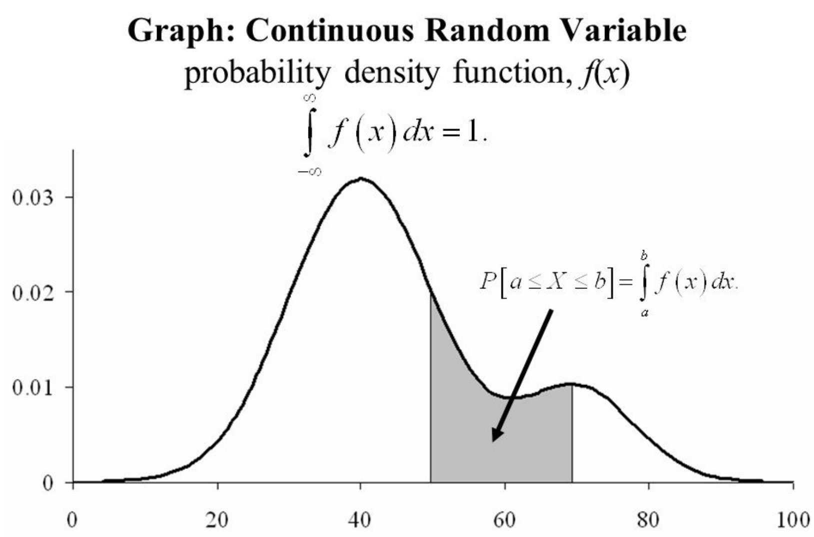
- Only requires presences (will draw background data for final presence-background dataset) but can also handle presence-absence data
- Works well with small sample sizes
- Consistently performs well when compared to other methods
- Can deal with correlated variables
- Can model complex interactions among variables

Disadvantages:

- Largely black-box (hard to understand / get under the hood)
- Despite best efforts, model overfitting can still be a problem
- Sensitive to background chosen → must be biologically justified
- Nature of the model (exponential) means no upper bound and thus extrapolation is problematic
- Model evaluation is difficult

3.1 Overview of MAXENT

Recall: A **probability density function** describes the probability of a continuous random variable being close to a given value



So for instance, we can think about the probability density of a set of environmental conditions across a landscape (some combinations of conditions are more probable than others)

3.1 Overview of MAXENT

Now let

z = vector of environmental covariates (variables)
L = background cells
$y=1$ for locations where a species is present

$f(z)$ = probability density of covariates across L
$f_1(z)$ = probability density of covariates across presence sites

3.1 Overview of MAXENT

We want to know what is the probability of presence (y =1) given a certain set of environmental values (z)?

$Pr(y=1 \mid z)$

3.1 Overview of MAXENT

We want to know what is the probability of presence (y =1) given a certain set of environmental values (z)?

Pr(y=1 | z)

Recall: Baye's Theorm tells us that the probability of A given B is equal to the likelihood of B given A times the probability of A divided by the probability of B or

P(A|B) = P(B|A) P(A) / P(B)

3.1 Overview of MAXENT

We want to know what is the probability of presence (y =1) given a certain set of environmental values (z)?

$Pr(y=1 \mid z)$

Recall: Baye's Theorm tells us that the probability of A given B is equal to the likelihood of B given A times the probability of A divided by the probability of B or

$P(A|B) = P(B|A) P(A) / P(B)$

When working with our probability density functions, this becomes:

$Pr(y = 1| z) = f_1(z)Pr(y = 1)/f(z)$

3.1 Overview of MAXENT

We want to know what is the probability of presence (y =1) given a certain set of environmental values (z)?

Pr(y=1 | z)

Recall: Baye's Theorm tells us that the probability of A given B is equal to the likelihood of B given A times the probability of A divided by the probability of B or

P(A|B) = P(B|A) P(A) / P(B)

When working with our probability density functions, this becomes:

$P(y = 1| z) = f_1(z)P(y = 1)/f(z)$

We can't know P(y=1) or prevalence from presence-only data
But MAXENT can estimate $f_1(z)$ and f(z) from the data….so we can get part way to a solution

3.1 Overview of MAXENT

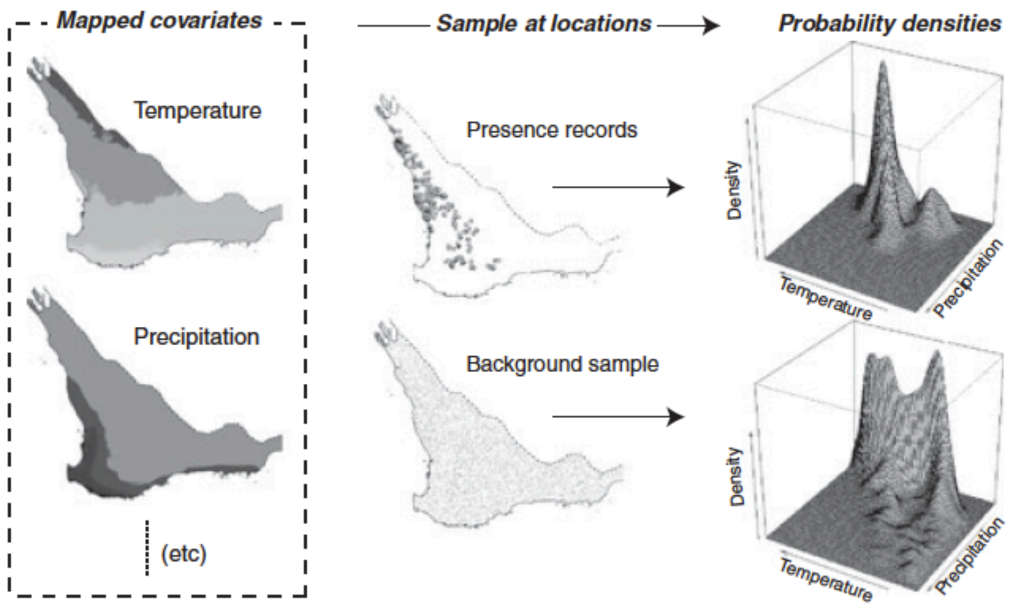


Figure 1 A diagrammatic representation of the probability densities relevant to our statistical explanation, using data presented in case study 1. The maps on the left are two example mapped covariates (temperature and precipitation). In the centre are the locations of the presence and background samples. The density estimates on the right are not in geographic (map) space, but show the distributions of values in covariate space for the presence (top right) and background (bottom right) samples. These could represent the densities $f_1(\mathbf{z})$ and $f(\mathbf{z})$ for a simple model with linear features.

MAXENT first estimates $f_1(z)$ such that it is as close to $f(z)$ as possible but constrained by information from the presence records (for instance mean of the distribution might be constrained to match mean of presences)

3.1 Overview of MAXENT

We want to know what is the probability of presence (y =1) given a certain set of environmental values (z)?

$P(y = 1| z) = f_1(z)P(y = 1)/f(z)$

How MAXENT specifically estimates $f_1(z)$ and $f(z)$ and what it does about the problem of prevalence are controlled by the parameters we set.

3.2 Key Settings

Features

Recall we’re thinking about our equation:

$P(y = 1| z) = f_1(z)P(y = 1)/f(z)$

Thinking about features is relevant when thinking about how we get estimates of $f_1(z)$ and $f(z)$

Features are behind-the-scene transformations of our original environmental covariates that specify the types of relationships we can have been the environment and the probability of presence

The types of features we select determines the types of constraints on $f_1(z)$

3.2 Key Settings

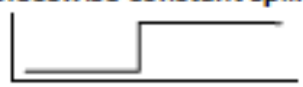
Features

Feature class	Description in relation to environmental variable	Constraint imposed on estimated distribution $\hat{P}$	Ecological interpretation of the constraint
Linear (L)	Variable itself	The mean of variable under $\hat{P}$ should be close to its mean in the sample locations	The mean of the sample indicates average conditions for species presence
Quadratic (Q)	Square of variable	If used with L, variance of variable under $\hat{P}$ is close to its variance in the sample	The variation in that variable in the sample indicates the tolerance of the species for variation from suitable conditions
Product (P)	Product of 2 variables	If used with linear features for the 2 variables, that the covariance of the variables under $\hat{P}$ should be close to its covariance in the sample	The effect of one variable on species presence varies with the value of the other variable – i.e. there are interactions between the variables.

Elith et al. 2011

3.2 Key Settings

Features

Threshold (T)	A step function that allows a different response below the threshold (the "knot") to that above it. Equivalent to a piecewise constant spline. 	The proportion of $\hat{P}$ that has values of this variable above the knot should be close to that proportion in the sample	Many threshold features can be used on the same variable, with different thresholds. These can add together to model an arbitrary stepped response to the variable.
Hinge (H)	Similar to the threshold feature, but the response above the knot (forward hinge; below left) or below the knot (reverse hinge; below right) is linear with a positive or negative coefficient (slope). Equivalent to a piecewise linear spline. 	The mean of the variable above the knot under $\hat{P}$ should be close to its mean above the knot in the sample locations	A model using only hinge features fits a piecewise linear response. If hinge features are used, linear features are redundant (a linear feature can be created from a hinge, with the knot at one extreme of the feature space).
Category (C)	A binary indicator showing membership in one class of a categorical variable. For a k-class categorical variable there will be k categorical features	The proportion of $\hat{P}$ that has values in this class should be close to that proportion in the sample	

Elith et al. 2011

3.2 Key Settings

Features

Note the distinction between the number of feature classes and the number of features:

A single feature class results in many features depending on the number of datapoints and environmental variables

For instance, if thinking just about the threshold and hinge feature classes, if we have 100 presence records and and 19 environmental variables, MAXENT will consider:

3 types of piecewise functions: threshold, forward hinge, reverse hinge) x
99 datapoint pairs x
19 predictor variables
= 5643 features! All of which it has to decide if it will retain in the final model

3.2 Key Settings

Regularization

In estimating the ratio $f_1(z)/f(z)$, MAXENT uses regularization to set constraints on how tightly the empirical sample must be fit

This is essentially an adjustment of model coefficients to balance model fit with model complexity (prevents overfitting to the presence data); forces many coefficients to zero thus retaining only the most critical

The regularization parameter for each type of feature achieves this adjustment. Default values of these parameters have been tuned using a large empirical dataset.

The beta multiplier setting in the program controls the strength of the feature-specific regularization parameters influence.

3.2 Key Settings

Output Type

The estimate of the ratio $f_1(z)/f(z)$ is referred to as the **raw** output

To get all the way to $P(y=1|z)$, MAXENT needs to "guess" at $P(y=1)$, which is what we ask it to do when we specify that we want the **logistic** output (default setting)

The way in which this is done relies on the **prevalence parameter**, τ

For practical purposes the logistic output scales monotonically with the raw output: the rank order of predictions for different sites will be the same but the actual logistic values depend on τ

MAXENT also can be instructed to produce **cumulative** scores…not often used

3.2 Key Settings

Probability of Presence/ Prevalence

The ratio $f_1(z)/f(z)$ describes an exponential-family model

τ parameter is used to transform this raw output to a logistic model

τ is defined as the probability of presence of the species at ordinary occurrence points

For instance, the default setting of 0.5 assumes we have a 50% chance of the species being present in suitable areas

For something that is very rare, may want to use a smaller value; for something that is more common, may want to use a higher value (would be ideal if we had empirical observations to support our choice)

3.2 Key Settings

Clamping (for the prediction stage)

→ At the prediction stage: Any grid cell in any prediction raster that has values that are above or below values used during training (i.e. higher or lower than values at the occurrence and background sites) are given the max or min value of the training data

→ Likewise derived features from the prediction dataset (i.e. combinations of variables etc.) are clamped

3.3 Parameter tuning

How to know what feature classes or regularization coefficients to use?

Default settings left alone in most empirical studies but this may not be optimal

We can run several preliminary models with different feature classes and regularization values and compare them to decide on a final set of parameter values. This is called “tuning”.

Models with different feature classes or regularization coefficients are compared using the evaluation metrics we will discuss tomorrow and/or AIC.

4. Basic MAXENT niche modeling

4.1 A quick look at the GUI

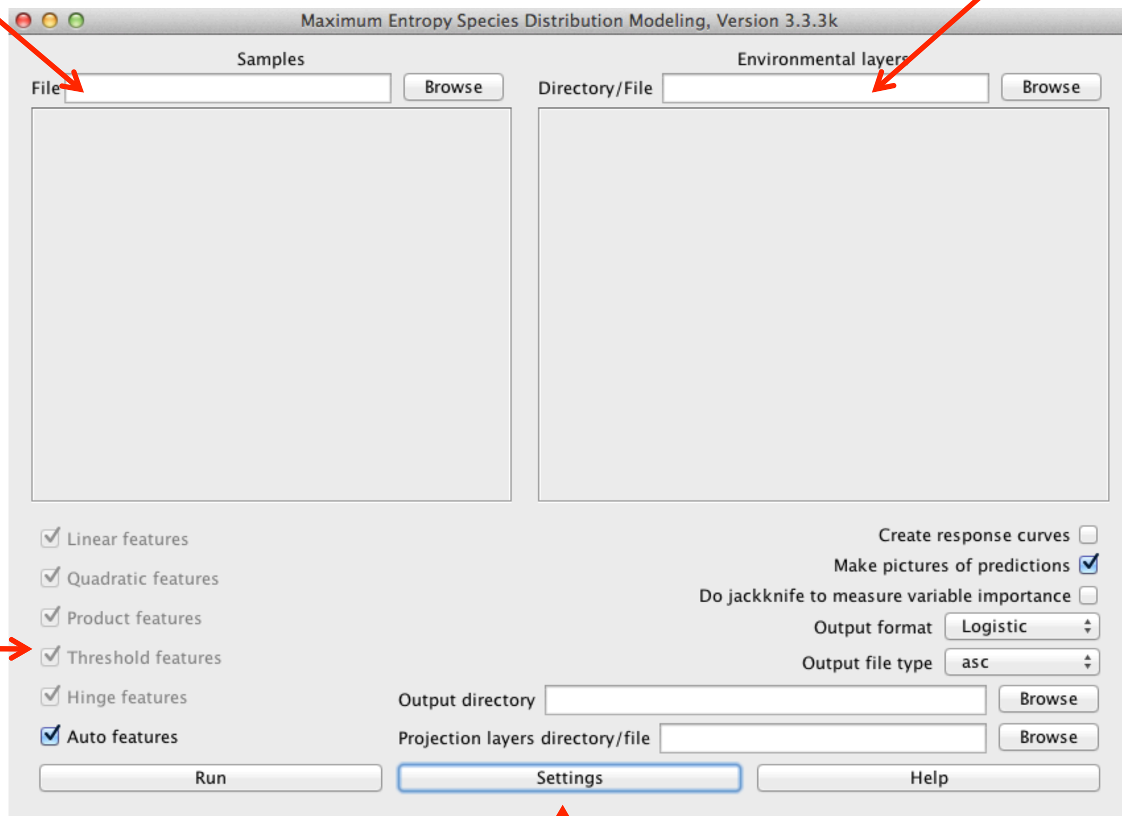
4.2 MAXENT modeling in R

4.3 Exercise

4.1 The MAXENT GUI

Load occurrence records

Load environmental rasters



Set features

Set type of output

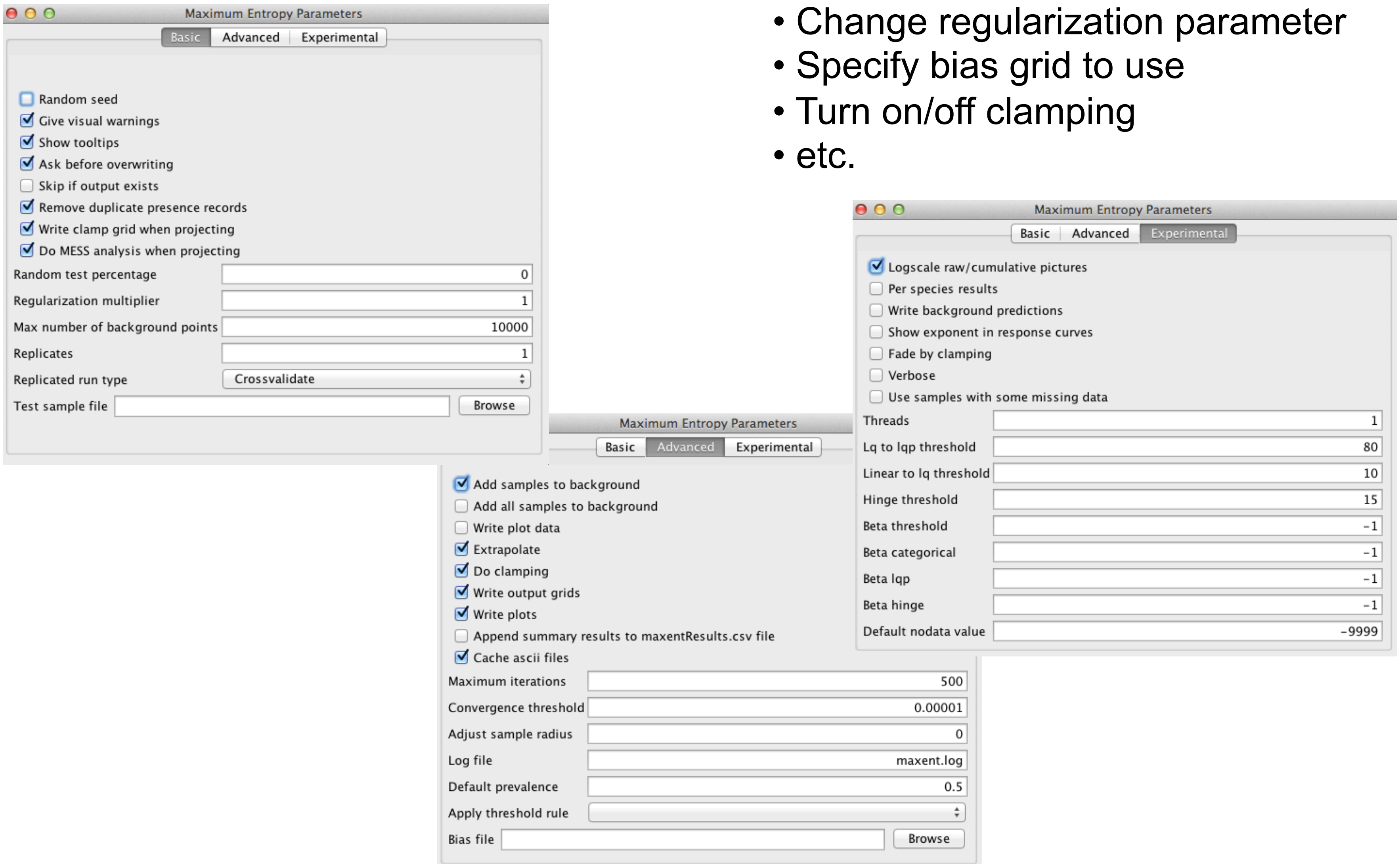
All other parameters

Note: you can open the GUI by double-clicking the jar file in your dismo package install

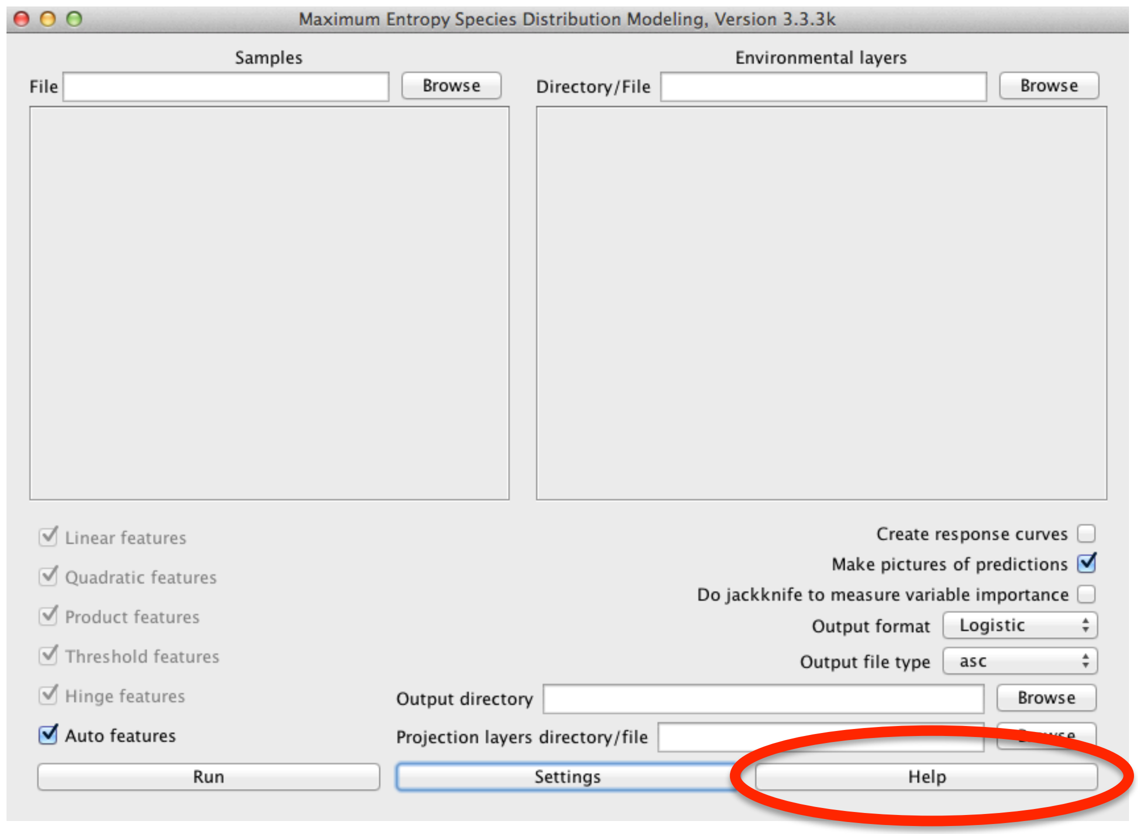
4.1 The MAXENT GUI

Setting Menus

- Change regularization parameter
- Specify bias grid to use
- Turn on/off clamping
- etc.



4.1 The MAXENT GUI



Note: The Help file accessible from the GUI has good explanations for all the parameters and the syntax given in this file is the same as that you would use to change default parameters in R

4.2 MAXENT modeling in R

R packages for generating MAXENT niche models

- dismo (we'll be using this)
- biomod2
- sdm

* All of these packages can be used to generate ENM using other algorithms

4.2 MAXENT modeling in R

Time for a Demo!

4.3 Exercise

Exercise D2.3

Time to build our first model!

1) Use the dismo package to build a MAXENT niche model in R using your final occurrence records and raster layers.

2) Create a simple plot of model predictions across the native range of the species.

3) Save your model.

Homework Checklist

1) Finish up the exercises from today

Tomorrow we will:

→ Learn about model evaluation

→ Project and interpret our models