Introduction to Maxent

Day 1 of 2

Outline

- Background on maximum entropy
- Basic Maxent concepts and jargon
- Intro Tutorial
- More Maxent features
- Exercises for additional features

Maximum entropy models

- Maximum entropy methods are very general ways to predict probability distributions given constraints on their moments
- Predict relative abundance distributions based on the number of individuals, species and total energy
- Predict community composition along ecological gradients based on traits
- Predict species' distributions based on environmental covariates
- Predict associations in food webs
- Probably much more to come…

Maxent for SDMs - Input

- **Presence locations**
  - More common than you think!
- Background locations (not pseudo-absences!)
- A gridded landscape (but you can work around this)
- **Environmental covariates for each landscape grid cell**
- Optional - A measure of sampling effort in each grid cell
- Optional – A landscape on which to project your predictions

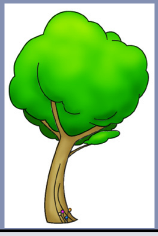
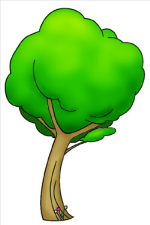
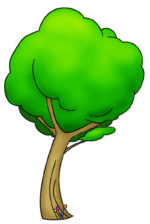
What is Entropy Maximization?

- You can think of Maxent as having two parts: a **constraint component** and an **entropy component**
- Constraint component
  - The data define moment *constraints* on the probability distribution
  - The temperature at the all the presence locations define the mean, variance, etc. of the temperature where the species occur
  - Maxent requires that the predicted distribution fulfills these constraints
- Entropy component
  - Many distributions could fulfill these constraints
  - Maximizing entropy is a method to choose among the many probability distributions that fit your data
  - Maxent starts by assuming the probability is perfectly uniform in geographic space and moves away from this distribution only to the extent that it is forced to by the constraints

What is Entropy Maximization

- Output
  - The output is a probability distribution that sums to 1
  - For species distributions this gives the **relative** probability of observing the species in each cell
  - Cells with environmental variables close to the means of the presence locations have high probabilities

Example of Maxent

T=10 	T=10	T=10
T=20 	T=20	T=20
T=30 	T=30	T=30

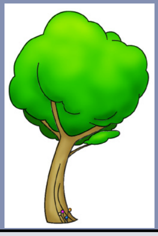
- •Mean annual temperature measured in each cell

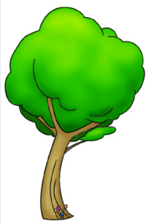
- •3 observed presences

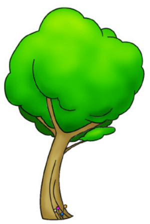
- •Mean temp of presence locations = (10+20+30)/3 = 20

- •What is the flattest distribution over all 9 grid cells that has a mean temp of 20?

Example of Maxent

T=10
p=1/9

	T=10
	p=1/9
		T=10
		p=1/9

T=20
p=1/9

	T=20
	p=1/9
		T=20
		p=1/9

T=30
p=1/9

	T=30
	p=1/9
		T=30
		p=1/9

- Constraint 1: avg. T= 20
- Constraint 2:

$$\sum_{i=1}^{i=9} p_i = 1$$

p_i =relative probability of observing the species in cell i.

- What is the flattest distribution over all 9 grid cells that has a mean temp of 20?

Example of Maxent

T=10
p=1/18
	T=10
	p=1/18
		T=10
		p=1/18

T=20
p=2/9
	T=20
	p=2/9
		T=20
		p=2/9

T=30
p=1/18
	T=30
	p=1/18
		T=30
		p=1/18

Constraints

$$\bar{t} = \sum_{i=1}^{i=9} p_i t_i = 20$$

$$\sum_{i=1}^{i=9} p_i = 1$$

- Other distributions are also consistent with these constraints, but they don’t have maximum entropy (i.e. they’re not as similar as possible to a uniform distribution).
- Can you think of some?

T=10
p=0
	T=10
	p=0
		T=10
		p=0

T=20
p=1/3
	T=20
	p=1/3
		T=20
		p=1/3

T=30
p=0
	T=30
	p=0
		T=30
		p=0

Maxent solution

General solution with uniform prior:

$$p_i = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots}}{\sum_i e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots}}$$

Prior information:

$$p_i = q_i \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots}}{\sum_i e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots}}$$

Note that these are the pretty solutions that don’t involve regularization (discussed shortly)

Priors

- In the context of species distributions we can interpret priors in two equivalent ways:
  - Our prior guess is that any cell is equally likely to contain the species. By applying constraints on temp, precip, etc. we want the most spatially uniform distribution possible.
  - Instead of thinking in terms of geographic space, we think in terms of environmental covariates. An equivalent statement to (1) is that we guess a species occurs at temperature x in proportion to the availability of value x in the landscape.
    - If 90% of cells are >30 degrees and 60% of occurrences are found in these cells, it doesn't mean that the species prefers temp >30 degrees, it actually suggests the opposite.

Priors

- Maxent uses a uniform prior (in geographic space) as the default, but others are possible
- Priors for sampling bias:
  - Downweight cells that have been searched but no presences
- Use priors to reflect other types of information?
  - Include other mechanisms – dispersal, biotic?
  - Incorporate known relationships between species' distributions and environmental covariates.

Model fitting

- We need to find the distribution with that's as flat as possible which fulfills the constraints
- The fitting proceeds using a random walk in parameter space (the coefficients for each predictor) by proposing a new parameter value and accepting it if it increases the *gain*
- The gain is like a deviance from GLMs; it's a function that we want to maximize to get the best model
  - A gain of 2 means that an average presence location has a relative probability exp(2) = 7.4 time higher than an average background site
- The random walk starts with a uniform distribution and continues until the increase in gain falls below some specified convergence level

Model fitting

- But we don’t necessarily want to fit the constraints *exactly*, they just need to be close
    - You don’t expect that the optimal temp for your favorite plant is 17.45 °, right? Maybe 17-18°.
- The constraints are relaxed using *regularization* (β)
- Regularization adds some junk to the gain function so that it doesn’t fit the observed presences too tightly
- A larger regularization parameter means more junk
    - You can adjust the regularization multiplier to relax the constraints more.
- Regularization is also used to select variables
    - When changes in a parameter don’t appreciably change the gain, the feature associated with that predictor gets tossed.

Why Use Maxent?

- Presence only data

- Explore complex relationships with environment (But don't retain ALL the complex relationships)

- Unbiased predictions, based on constraints (kind of)

- Generative, and not discriminative model – good for small data sets

- Many species to process

- Its predictions compare favorably to presence/psuedo-absence models when tested against real pres/abs data (e.g., Elith et al 2006)

Why Not Use Maxent?

- Presence/Absence data, *where you believe the absences at the spatial scale that you're modeling*
- Statistical properties not well understood – prediction variance limited to resampling methods
- Software is a bit of a black box and allows only some customization

Things we're still working on (reasons for cautious interpretation)

- Statistical uncertainty
- Appropriate priors
- Variable selection
- Incorporating sampling bias
- Evaluating models

Reasonable methods exist to deal with these issues, but its not clear (to me) that they're the best ways

Outline

- Background on maximum entropy
- Basic Maxent concepts and jargon
- Intro Tutorial
- More Maxent features
- Exercises for additional features

How do we interpret the predicted probability distribution?

- A continuous, not a binary surface
- To get a binary presence/absence surface, you need to choose a threshold (which can be a bit arbitrary)
- The **raw** output sums to 1. They just measure the relative likelihood of presence in 1 cell compared to another
- The **cumulative** output is the empirical cumulative distribution function; if the value is 40 in cell x then 40% of cells have a lower value than cell x
  - This can be interpreted in terms of the omission rate. A cumulative value of 40 means that (roughly) 40% of presences would be predicted as absences if we used this value as a threshold to create a binary surface.
- The **logistic** output

Feature types

- **Linear** - continuous variables should be close to their observed values (their mean at occurrence localities)
- **Quadratic**- variance of continuous variables should be close to observed values
- **Product** - covariance of two continuous variables should be close to observed values
- **Threshold** - proportion of model that has values above a threshold for a continuous variable should be close to observed proportion
- **Hinge** - linear feature truncated at threshold
- **Binary** - the proportion of each category in a categorical feature should be close to the observed proportions

Testing vs. Training

- There are two ways to evaluate a model
  - Fit
    - How well does it explain the data used to fit the model?
    - Maxent calls this *training* data
  - Predict
    - How well does Maxent predict independent data
    - Hold out data when fitting models
    - Maxent calls this *test* data
- Maxent provides you will a bunch of metrics that are either based on test or training data
- The best case scenario is having independent presence/ absence data
  - But if you had this, you wouldn't be using Maxent in the first place

Outline

- Background on maximum entropy
- Basic Maxent concepts and jargon
- Intro Tutorial
- More Maxent features
- Exercises for additional features

Intro tutorial

- Go through the first 25 pages of the tutorial to get familiar with Maxent's buttons
- The tutorial doesn't explicitly ask you to create each model, but you should do it anyways
- If you get done early, start the exercises posted on the web site

Outline

- Background on maximum entropy
- Basic Maxent concepts and jargon
- Intro Tutorial
- More Maxent features
- Exercises for additional features

Post tutorial notes

- Managing the files
  - Based on species name
  - Overwrites
- ***maxentResults.csv*** **-** listing the number of training samples used for learning, values of training gain and test gain and AUC. Test gain and AUC are given only when a test sample file is provided or when a specified percentage of the samples is set aside for testing. If a jackknife is performed, the regularized training gain and (optionally) test gain and AUC for each part of the jackknife are included here.
- ***maxent.log*** **-** records the parameters and options chosen for the run, and some details of the run that are useful for troubleshooting.
- ***mySpecies.html*** - the main output file, containing statistical analyses, plots, pictures of the model, and links to other files. It also documents parameter and control settings that were used to do the run.
- ***mySpecies.asc*** **(or *mySpecies.grd*) -** containing the probabilities in ESRI ASCII grid format (or in DIVA-Gis format if -H switch is used)
- ***mySpecies.lambdas*** **-** containing the computed values of the constants *c1, c2,* ... (described below)
- ***mySpecies.png*** **-** is a picture of the prediction
- ***mySpecies_omission.csv*** **-** describing the predicted area and training and (optionally) test omission for various raw and cumulative thresholds
- various plots for jackknifing and response curves, in the *plots* subdirectory.

Model Evaluation

Area under ROC curve (AUC)

- Receiver Operating Characteristic (ROC)
- Contingency Table for a given threshold:

		Actual Value (Data)	
		Presence (pos)	Absence (neg)
Predicted Outcome (Model)	Presence (pos)	True Positive (TP)	False Positive (FP)
	Absence (neg)	False Negative (FN)	True Negative (TN)

Model Evaluation

Area under ROC curve (AUC)

- Sensitivity- True Positive Rate (TPR)

		Actual Value (Data)	
		Presence (pos)	Absence (neg)
Predicted Outcome (Model)	Presence (pos)	True Positive (TP)	False Positive (FP)
	Absence (neg)	False Negative (FN)	True Negative (TN)

Model Evaluation

Area under ROC curve (AUC)

- Specificity- True Negative Rate (TNR)

		Actual Value (Data)	
		Presence (pos)	Absence (neg)
Predicted Outcome (Model)	Presence (pos)	True Positive (TP)	False Positive (FP)
	Absence (neg)	False Negative (FN)	True Negative (TN)

Model Evaluation

Area under ROC curve (AUC)

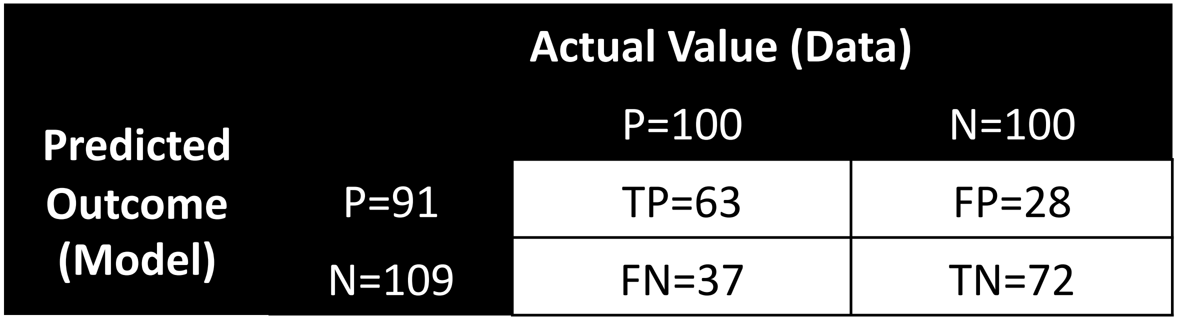
- Specificity- True Negative Rate (TNR)
- ROC is Sensitivity by (1- Specificity)=(FPR)

		Actual Value (Data)	
		Presence (pos)	Absence (neg)
Predicted Outcome (Model)	Presence (pos)	True Positive (TP)	False Positive (FP)
	Absence (neg)	False Negative (FN)	True Negative (TN)

Model Evaluation

Area under ROC curve (AUC)

- An example:



		Actual Value (Data)	
Predicted Outcome (Model)		P=100	N=100
	P=91	TP=63	FP=28
	N=109	FN=37	TN=72

- TPR = 63/100 = .63
- FPR = 28/100 = .28

Model Evaluation

Area under ROC curve (AUC)

A

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

B

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

C

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

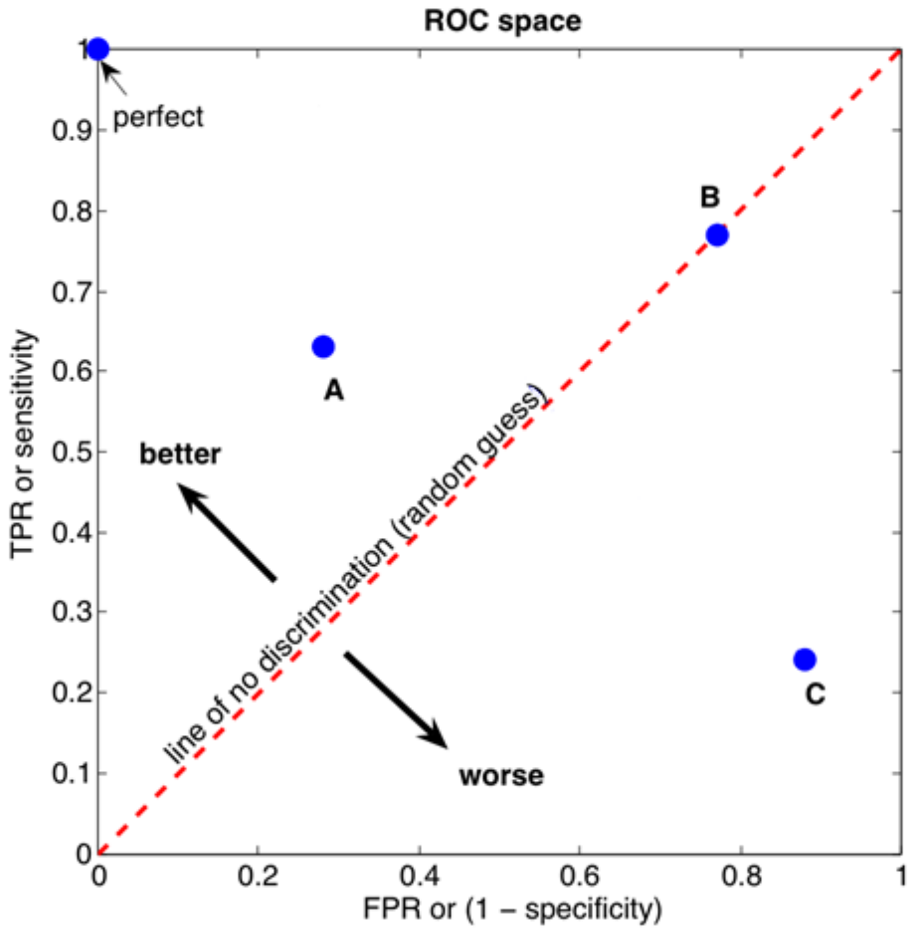
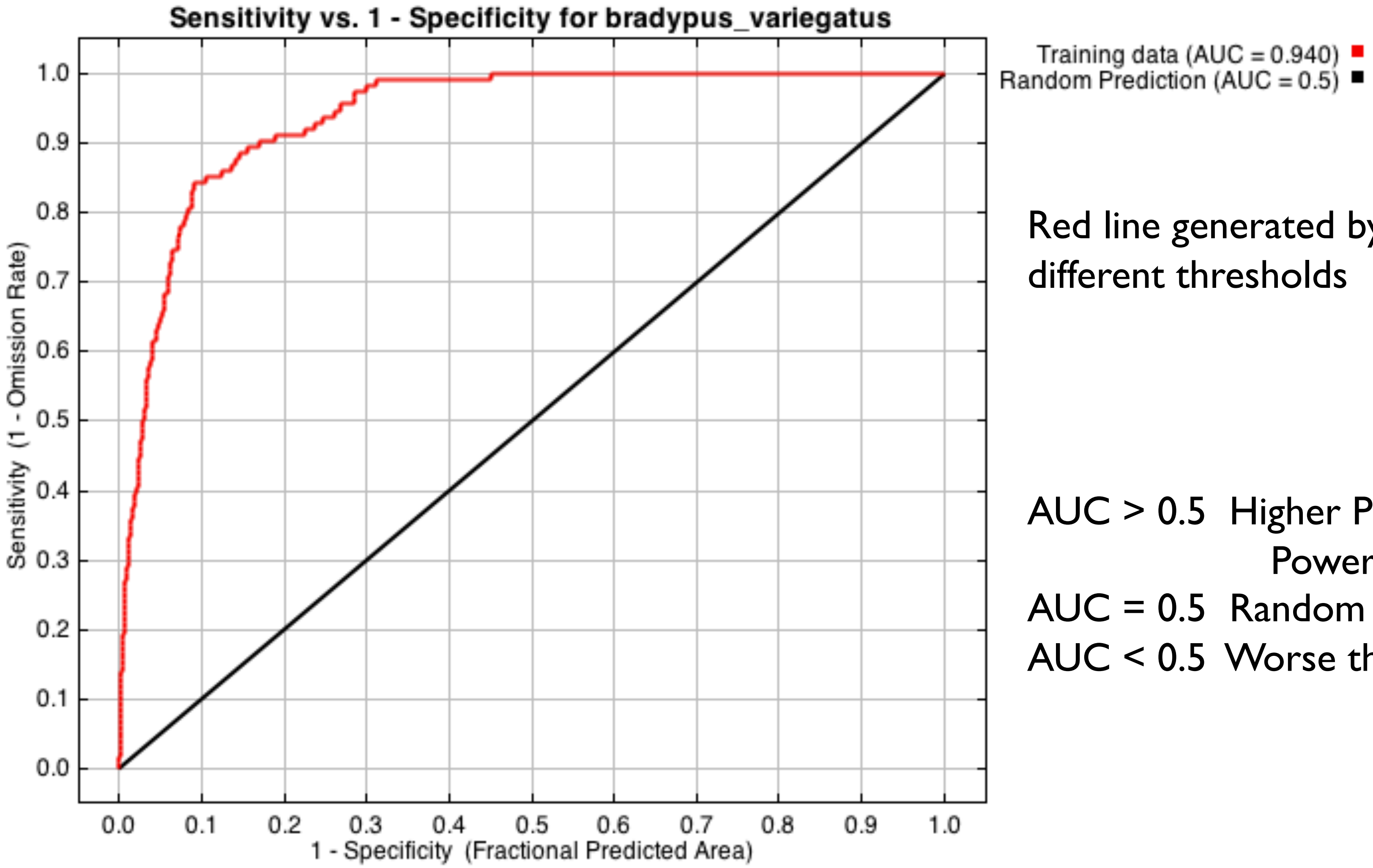


Image from Wikipedia

Model Evaluation

Area under ROC curve (AUC)



Red line generated by using different thresholds

AUC > 0.5 Higher Predictive Power

AUC = 0.5 Random Chance

AUC < 0.5 Worse than Random

Model Evaluation

BUT……………

Recall that we don't have true absences, otherwise we wouldn't be using Maxent in the first place.

AUC calculated by Maxent discriminates between presences and background points.

Background points are treated as pseudo-absences only for the evaluation procedure (and not for the model fitting)

This isn't great, particularly if your background samples are a poor representation of absences, as could happen when little sampling, but there aren't really alternatives

Estimating robustness of predictions

Resampling methods to validate model

- **Bootstrap** – resampling the original data and refit
- **Jackknife -** called 'subsample' to avoid confusion with jackknife for predictors– withhold 1 or more points and refit model
- **Cross validation** (k-fold) – most common - break sample in to k subsets and run the model k times, withholding the kth subset in each one. Use the kth subset for testing. Usually average fit statistics (e.g. AUC) over the k replicates
- **Permutation** – not implemented internally, but you can rearrange points yourself - randomly rearrange which sites are presences and absences and refit

- For more info on resampling: http://en.wikipedia.org/wiki/Resampling_(statistics)#

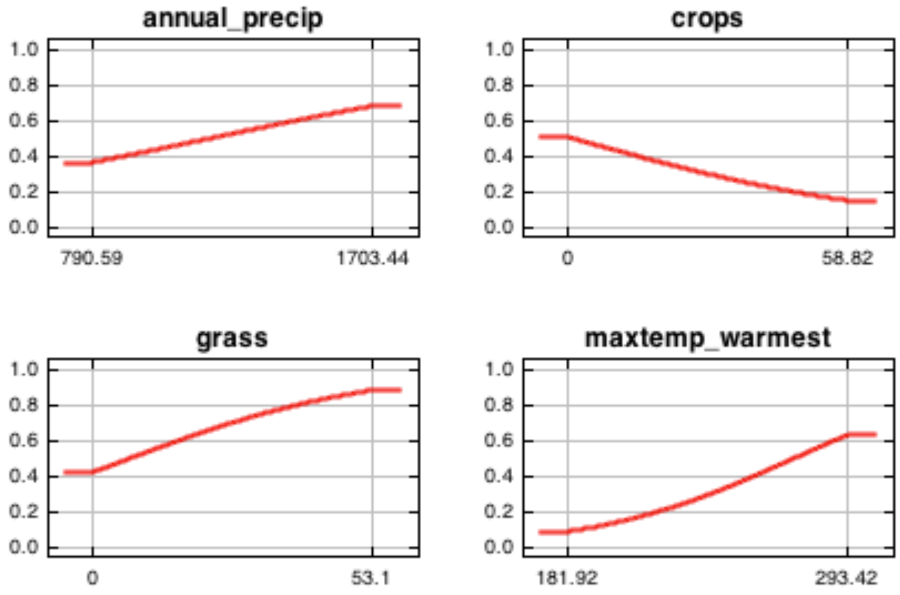
Variable selection

- **Jackknife**
  - Withhold 1 predictor and refit model
  - Withhold all predictors but 1 and refit the model
- **Response curves**
  - Why do we need to look at response curves and not just the coefficients associated with each variable?
  - Under what circumstances could we just look at the coefficients?
- **Lambdas file**
  - Try to look for nonsense to toss feature types or predictors
  - ad hoc
- **Permutation**
  - not implemented
  - for the i^{th} replicate, randomly select values of the j^{th} predictor, fit the model, and compare observed model fit to permuted fits. Can get significance of predictor this way

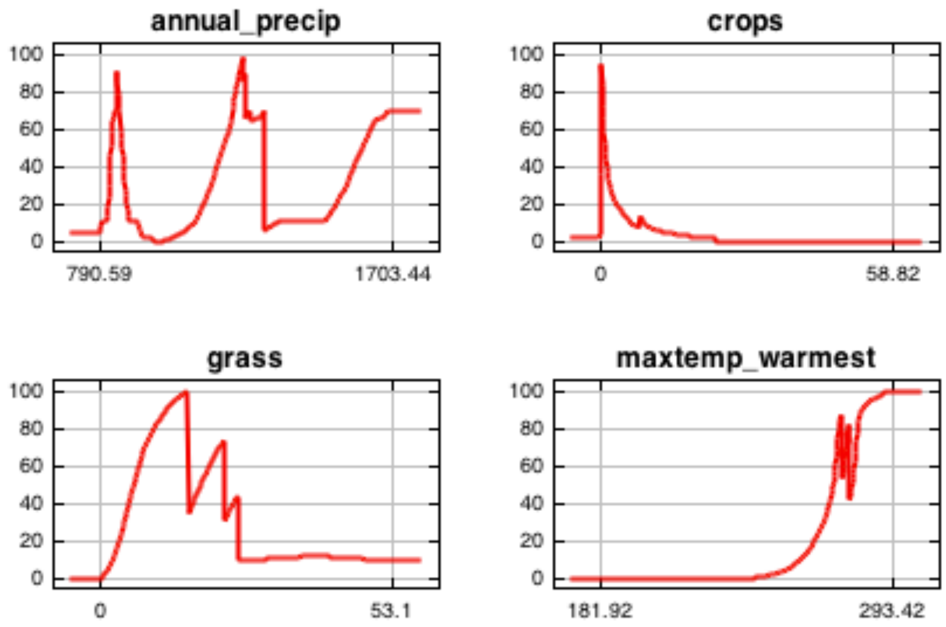
Response Curves

- Picking up the details or overfitting?
- Results for Japanese barberry from IPANE plots

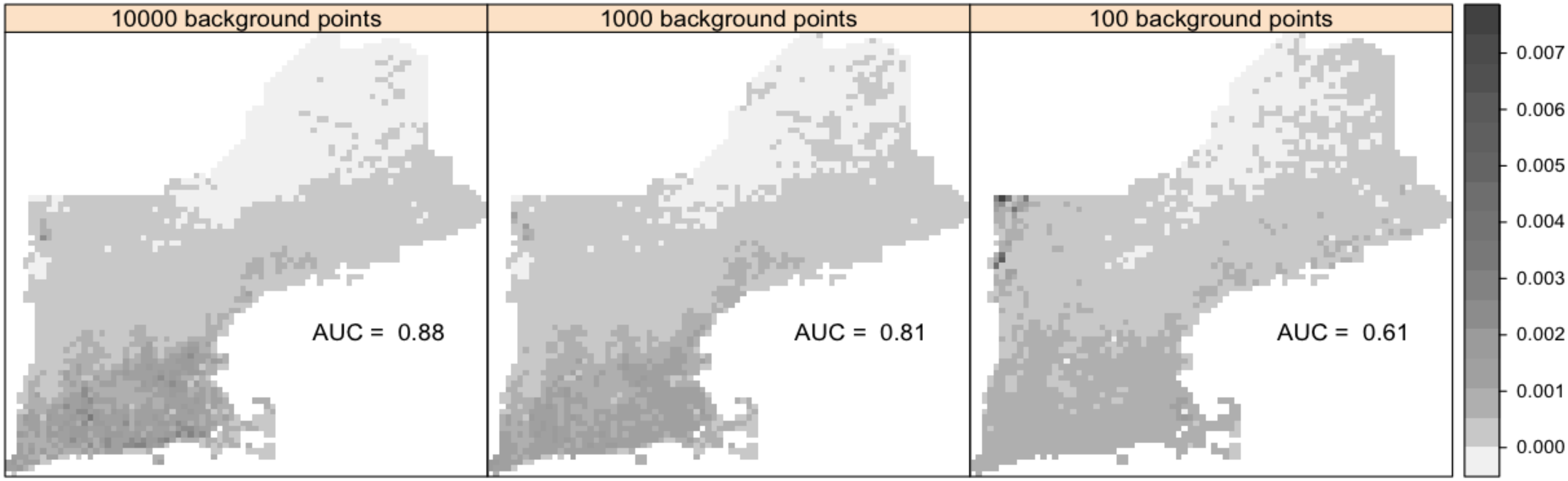
Linear features



All features (default)

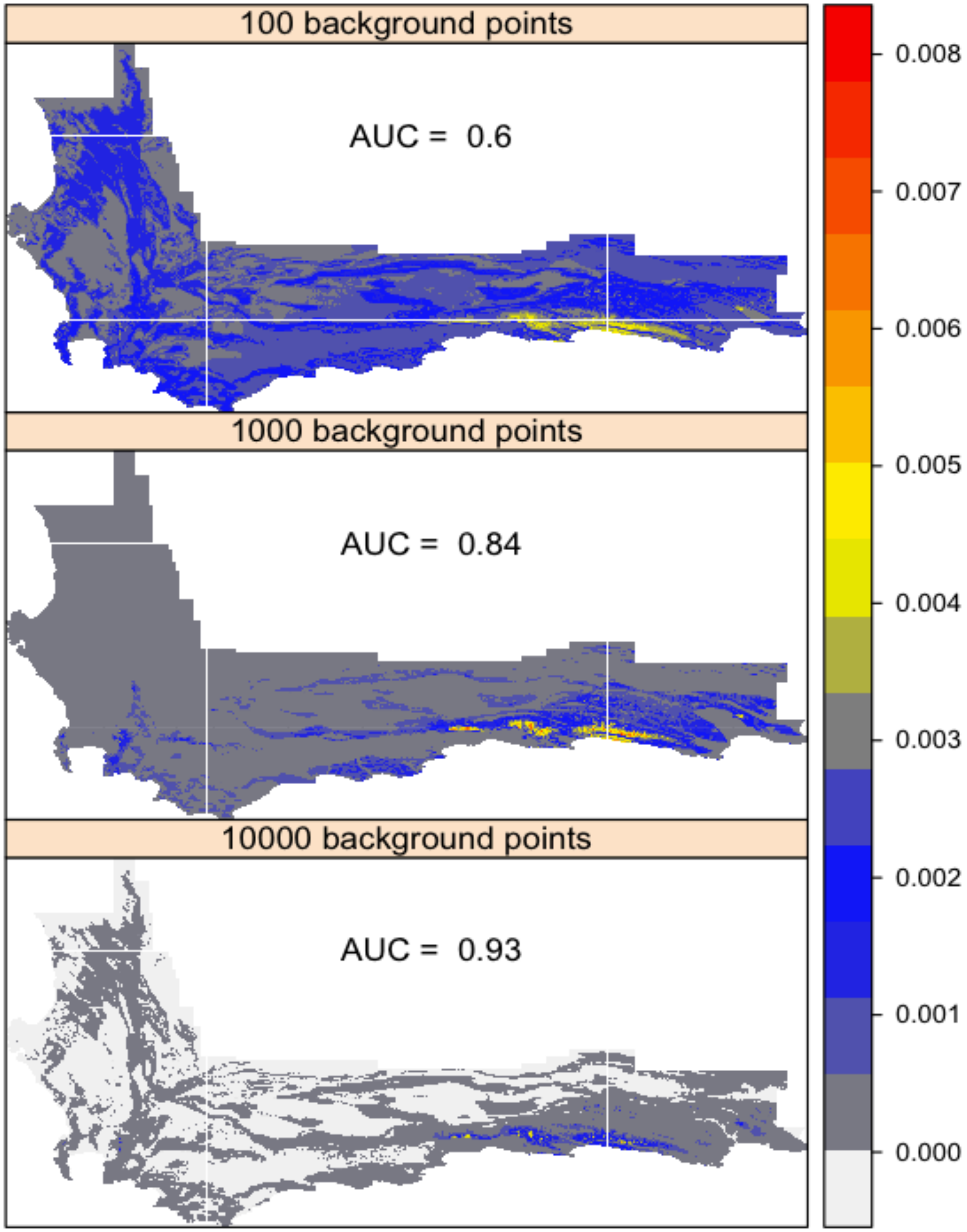


Background points



Note that I had to normalize these surfaces of raw output to sum to 1 and make them comparable because MaxEnt only makes the predictions sum to 1 over the training data. So using fewer background points assigns a larger probability to each cell, which I then normalize. Using fewer background points seems to make it harder to distinguish the high suitability sites.

Background points



Coefficients

Full model

annual_precip	0.00
crops	0.00
deciduous	0.00
developed	-0.01
evergreen	0.00
grass	0.00
maxtemp_warmest	1.01
mintemp_coldest	0.80
precip_seasonality	0.00
precip_warm_quarter	0.00
shrubland	0.00
crops^2	-14.21
developed^2	-0.86
evergreen^2	-0.77
grass^2	-3.53
mintemp_coldest^2	-1.39
annual_precip*deciduous	0.90
annual_precip*grass	0.57
crops*developed	0.97
crops*evergreen	1.70
crops*precip_seasonality	-4.37
crops*shrubland	4.21
deciduous*developed	-0.67
deciduous*evergreen	-0.11
deciduous*shrubland	-5.84
developed*evergreen	-0.55
developed*mintemp_coldest	-2.90
developed*precip_warm_quarter	-0.02
developed*shrubland	-0.86
evergreen*mintemp_coldest	-1.10
evergreen*precip_seasonality	0.76
grass*mintemp_coldest	0.54
grass*shrubland	0.38
maxtemp_warmest*precip_seasonality	-0.72
mintemp_coldest*precip_seasonality	0.95
(-126.03499984741211<minte	

LQP Model

annual_precip	0.00
crops	-0.16
deciduous	0.84
developed	0.89
evergreen	-0.21
grass	1.07
maxtemp_warmest	0.91
mintemp_coldest	0.00
precip_seasonality	-0.93
precip_warm_quarter	0.00
shrubland	-0.53
crops^2	-15.76
deciduous^2	-1.71
developed^2	-3.82
evergreen^2	-4.97
grass^2	-28.12
maxtemp_warmest^2	0.26
mintemp_coldest^2	-6.58

Linear Model

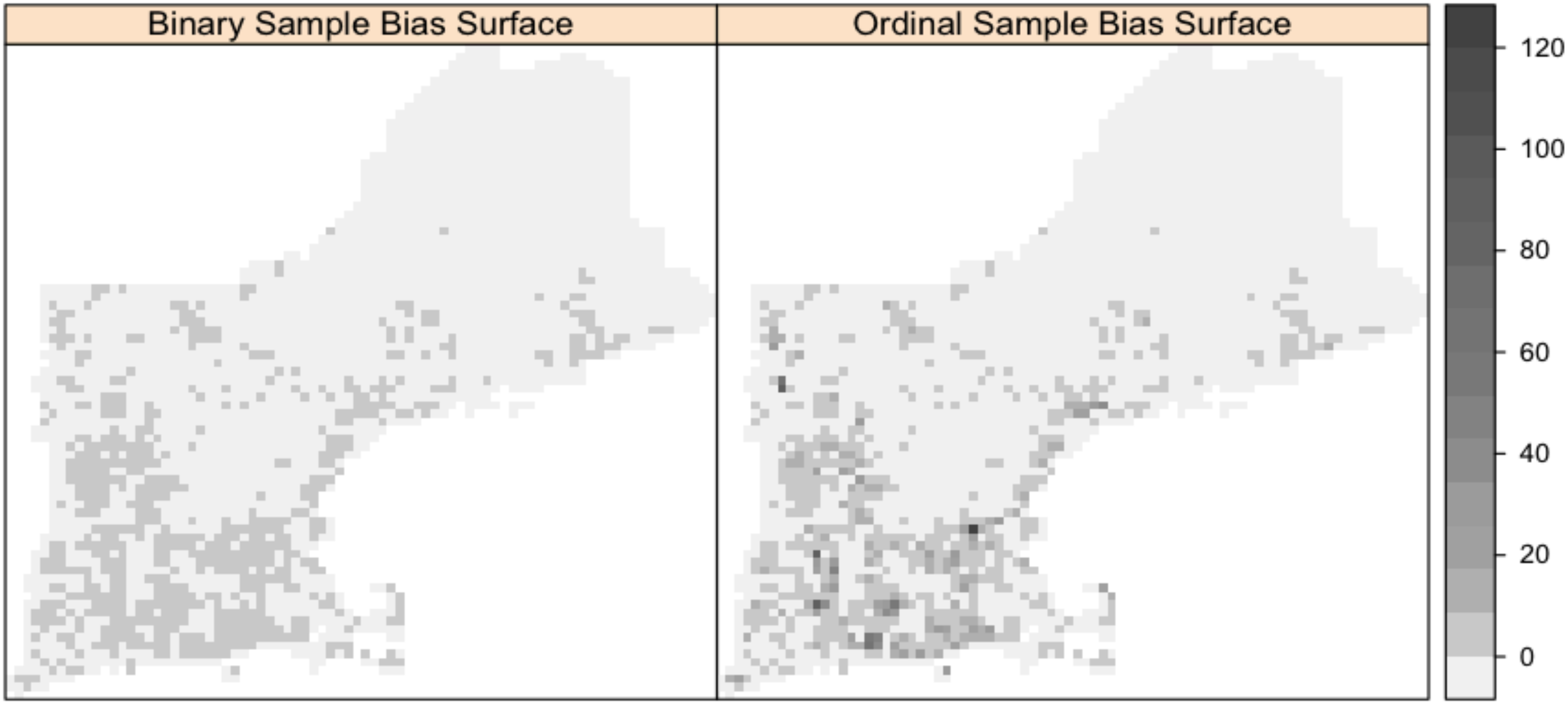
annual_precip	1.31
crops	-1.73
deciduous	1.03
developed	0.66
evergreen	0.19
grass	2.32
maxtemp_warmest	2.80
mintemp_coldest	2.72
precip_seasonality	-1.70
precip_warm_quarter	-1.47
shrubland	-9.10

Coefficients

Predictor	Model with all features	Model with linear, quadratic, product features	Model with only linear features
annual_precip	0.00	0.00	1.31
crops	0.00	-0.16	-1.73
deciduous	0.00	0.84	1.03
developed	-0.01	0.89	0.66
evergreen	0.00	-0.21	0.19
grass	0.00	1.07	2.32
maxtemp_warmest	1.01	0.91	2.80
mintemp_coldest	0.80	0.00	2.72
precip_seasonality	0.00	-0.93	-1.70
precip_warm_quarter	0.00	0.00	-1.47
shrubland	0.00	-0.53	-9.10

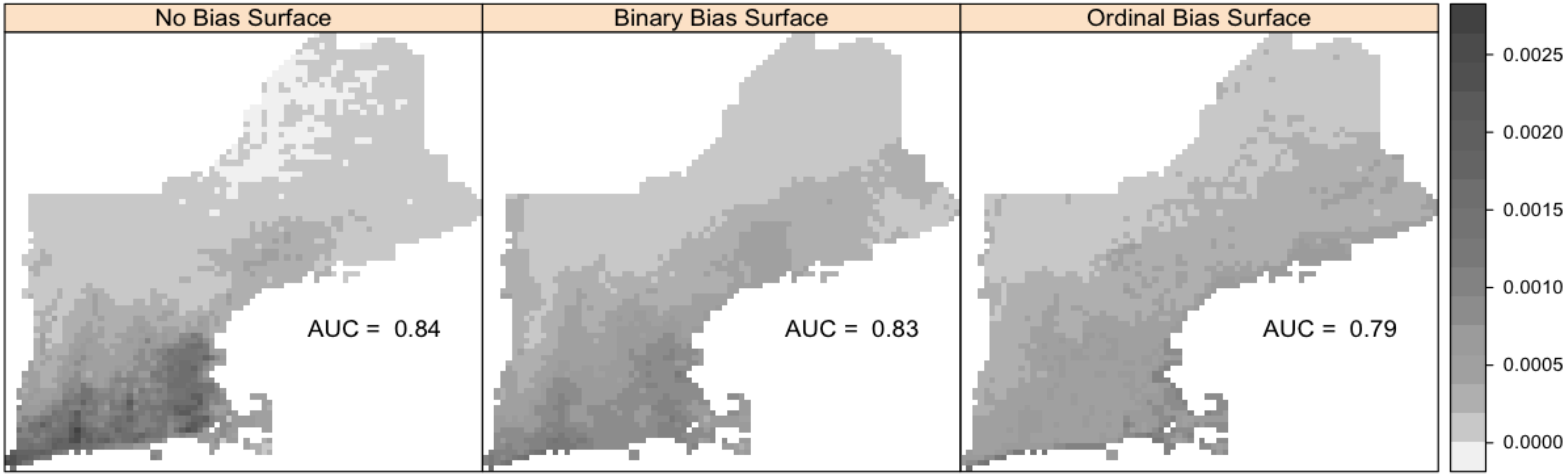
Bias surfaces

Two ways of constructing a bias surface. These are not predictions, just sampling intensity



Bias Surfaces

These are the predictions. Models use only linear features and show raw output.



Summary

- Showing you how Maxent predictions vary under different assumptions IS NOT MEANT TO DISCOURAGE YOU from using it
- Any modelling strategy will will provide variable results with varying assumptions
- The point is that you need to check your assumptions, which is not currently common practice in many, if not most, published papers that use Maxent
- Maxent is a perfectly good strategy when you have presence only data, which most data is on the scale that environmental grids are obtained
- Right now there aren't really alternatives for presence only data, but there will be shortly and we'll then be able to make comparisons

My suggestions

- Only linear features unless you expect other types are important independently
  - Include ONLY fancy features for the variables that matter
- Don't accept crazy response curves
  - You won't get as many with linear models
- Background choice depends on the question you're asking
  - See Elith et al 2011 for a good example
- You probably need a bias surface – check
  - Not sure what the best methods for this are yet
- Variable importance metrics provided are heuristic - only large differences mean much
  - It may turn out that these are good, but I haven't seen good proof
- Explore changes in regularization for your species

More of my suggestions

- Don’t threshold predictions – it’s too arbitrary
- Use logistic output but don’t interpret it as probability of presence
  - This transformation might be ok, but I think it’s based on tenuous assumptions about prevalence
- Don’t buy too much into the AUC because its treating background as pseudo-absences
  - Not really alternatives, unfortunately
- The defaults are good for batch processing large numbers of species (this is how they were obtained) but if you’re looking at just a few species, you need to customize your model
  - Phillips and Dudik 2006 argue for default settings when high bias or few samples
- Ask Phillips on Monday – maybe I’m full of crap!

Outline

- Background on maximum entropy
- Basic Maxent concepts and jargon
- Intro Tutorial
- More Maxent features
- Exercises for additional features
  - Download ‘Exercises Day 4.doc’ from the website
  - Write your answers to the questions in the document and email it to me as ‘homework 3’.