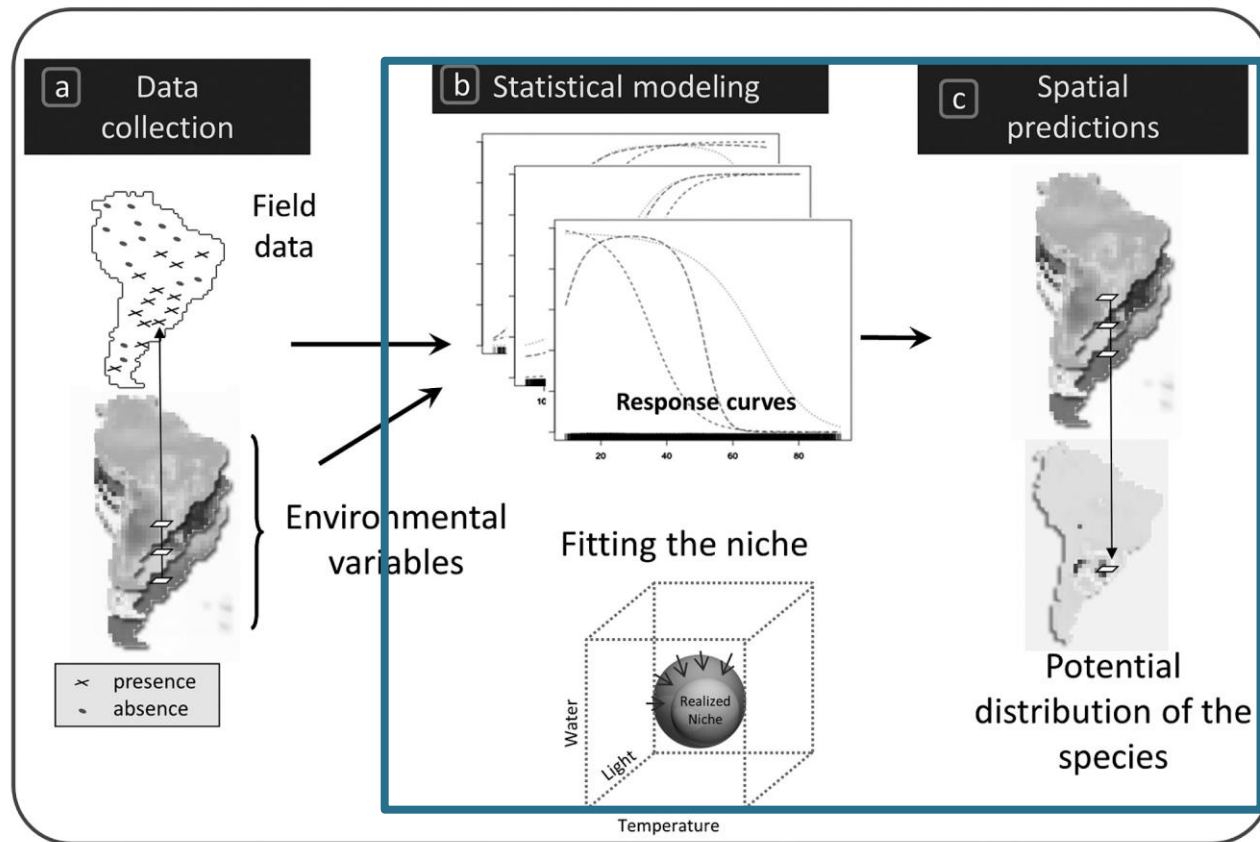


SDM Crash Course #4: Common SDM Algorithms



What are we doing today?

From Niche to Distribution: Basic Modeling · 43



SDM Algorithms:

Regression-Based
Brief overview
Machine Learning
In-depth

Disclaimer

“All models are wrong, but some are useful”

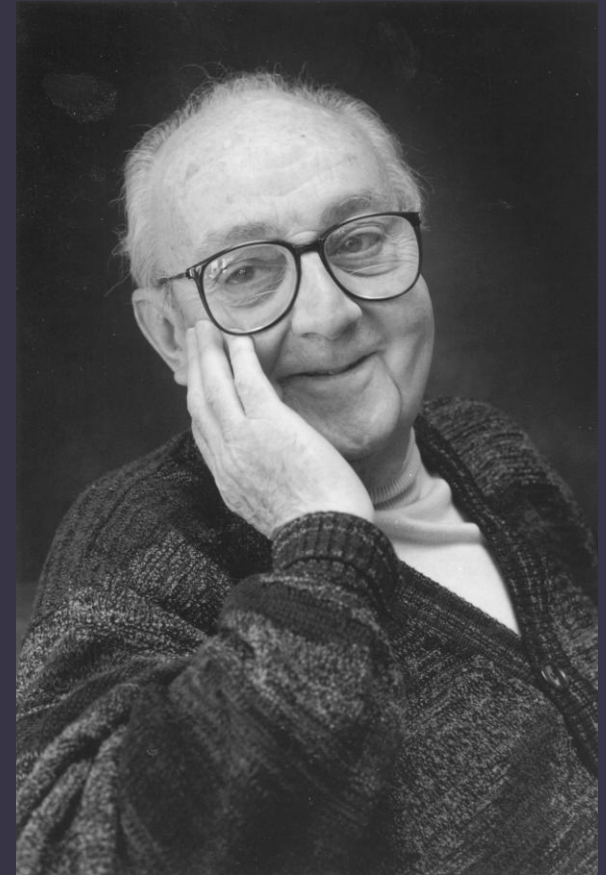
All input is flawed in some manner

Different models make different assumptions

Distributional dynamics are *complex*, to put it lightly

However, we can do our best to reflect reality in our models

Doing so is vital for 1) integrity purposes, & 2) for applicability



What is the aim?

From the characteristics of a site, we are trying to predict the probability that :

- a) A given species occupies that site
- b) A site is suitable for > 0 population growth for a given species
- c) A site is a part of a given specie's fundamental niche
- d) A site is part of the geographical range of a species



What is the aim?

From the characteristics of a site, we are trying to predict the probability that :

- a) A given species occupies that site
- b) A site is suitable for > 0 population growth for a given species
- c) A site is a part of a given specie's fundamental niche
- d) A site is part of the geographical range of a species

Answer: The interpretation varies, and some people get quite heated about it, but the math is the same

How do these models imply probability?



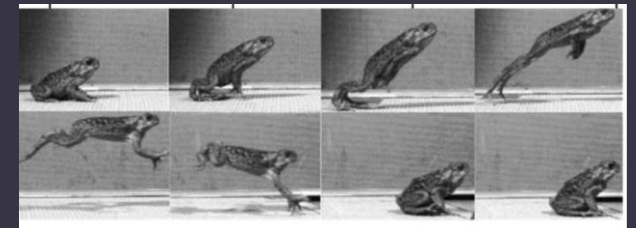
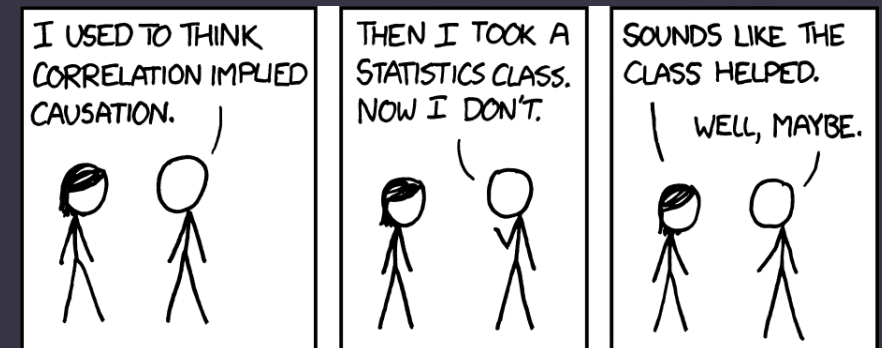
Types of Models

Correlative: Produces a function that encapsulates relationships between spatially explicit variables and species occurrence

- Pro: So many models available, no physiology data needed
- Con: So many models available, strictly associational (makes extrapolation into new spaces/time less robust)

Mechanistic: Produces a function using known physiological performance over a gradient of environmental conditions to predict suitability

- Pro: Directly informed by a species physiological performance
- Cons: A full mechanistic model would require A TON of experimentally collected data over many conditions, not there for most organisms



Combination Approach

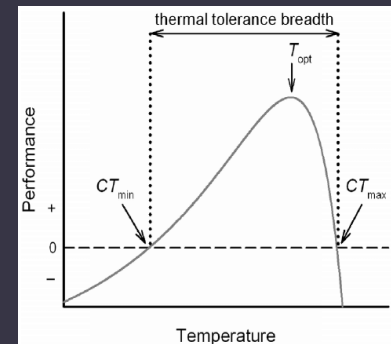
Correlative models informed by physiology:

Almost no spp. has the information to be modeled fully mechanistically across their entire extent

However, some forms of physiological performance (namely temperature-dependent performance) are fairly common

Utilizing mechanistic information from these can allow for more robust spatiotemporal extrapolation

MaxEnt



Basic Regression Models

Generalized Linear Models:

Parametric linear or higher-order polynomials used to fit model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

Valavi et al., 2022 found that GLMs fit with Lasso (L1) regularization perform as well as MaxEnt

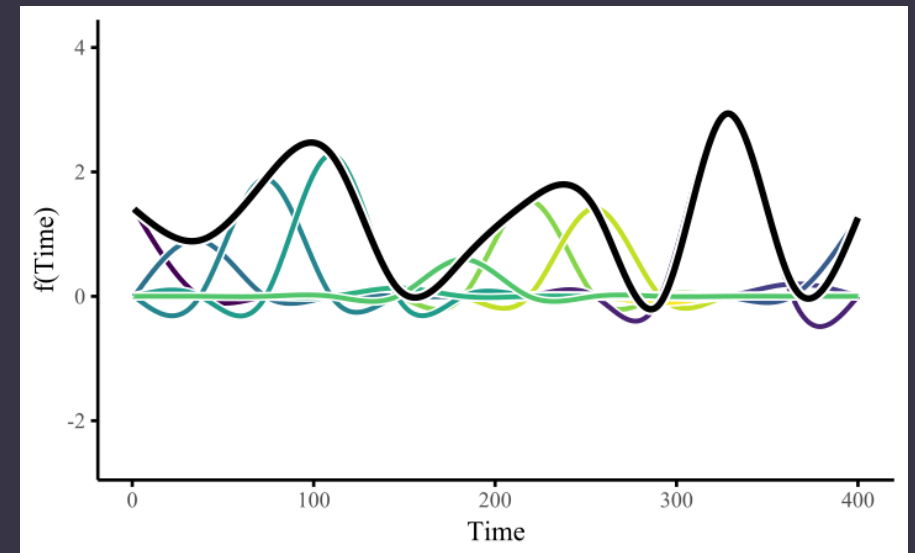
Predictive performance of presence-only species distribution models: a benchmark study with reproducible code

Roosbeh Valavi ✉ Gurutzeta Guillera-Arroita, José J. Lahoz-Monfort, Jane Elith

Generalized Additive Models:

Non-parametric smoothing functions to fit models

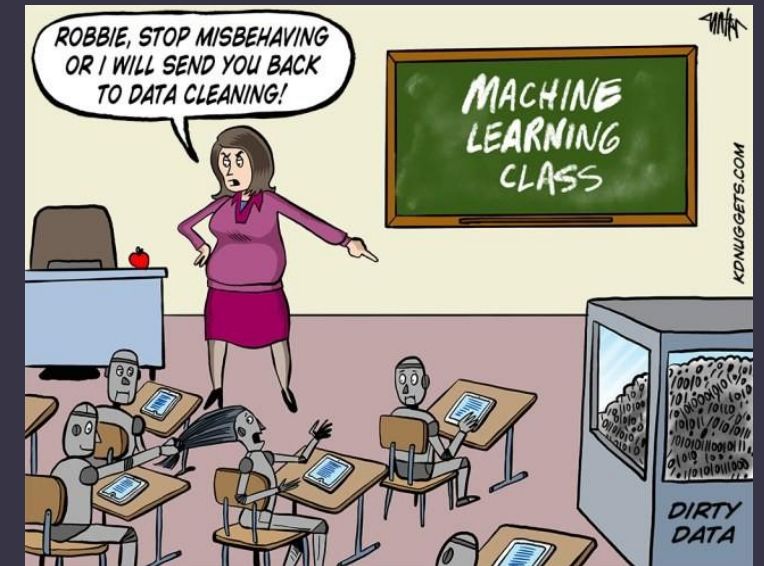
Captures non-linear relationships better than GLM



Machine Learning Approaches

Machine Learning: Subset of AI that focuses on using algorithms to learn from data to make predictions

- *Supervised Learning*
 - Every function trained on a random subset of the input data and tested against remainder iteratively
 - Stops when improvement is no longer being made
- *Consensus*
 - Functions make individual predictions
 - Final prediction is weighted proportion of those



MaxEnt, Roughly

Occum's Razor: A simpler explanation should be preferred in the absence of falsifying data

- Simple: Fewer assumptions
- Making an assumption requires explanation be done, which detracts from the explanatory power of the hypothesis



Principle of Maximum Entropy: A good fitting probability distribution best represents the current knowledge of a system when it implies the largest entropy (in other words, when the informational *entropy* is *maximized*)

- In layman's terms, our explanation has no more information than absolutely necessary to explain our input data

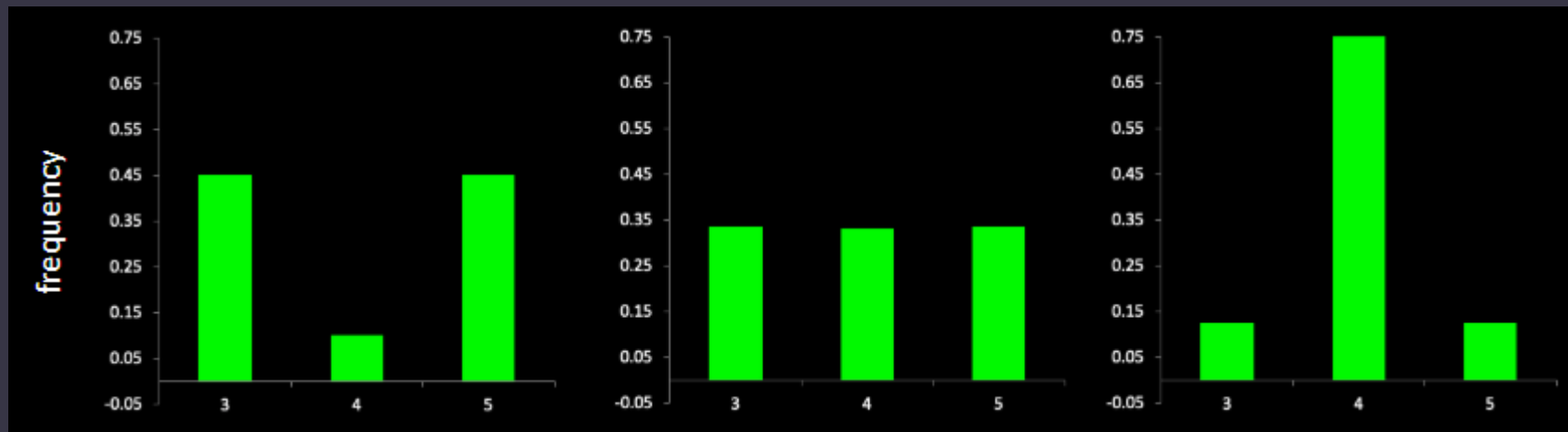


MaxEnt: Data Concept

Goal: Select models that fit constraint but maximize system entropy

Constraints:

1. $\bar{x} = 4$
2. $x_i \in [3,4,5]$

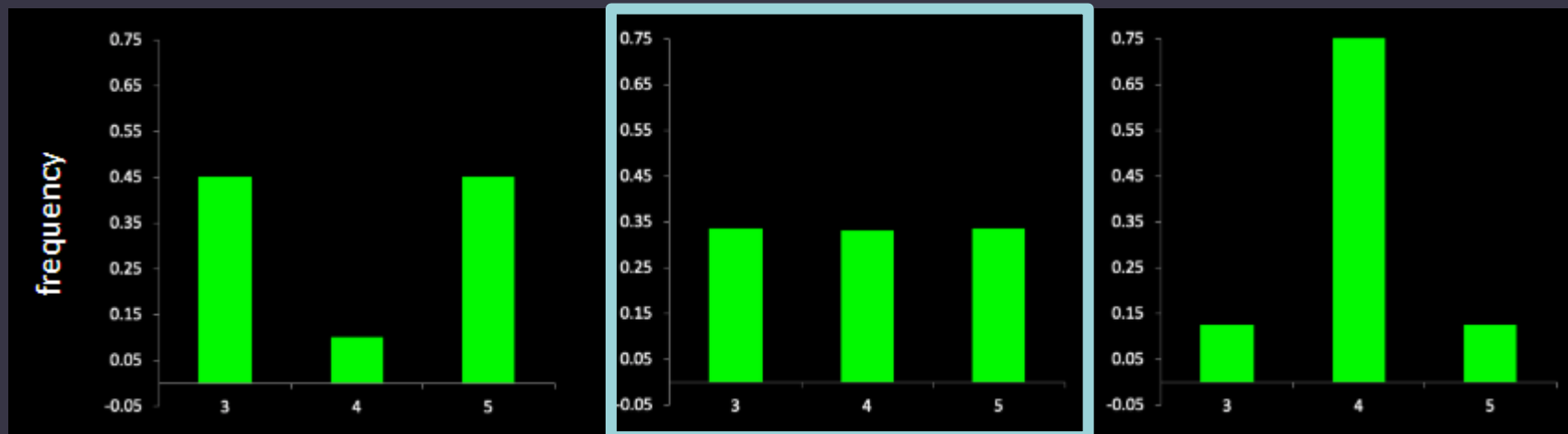


MaxEnt: Data Concept

Goal: Select models that fit constraint but maximize system entropy

Constraints:

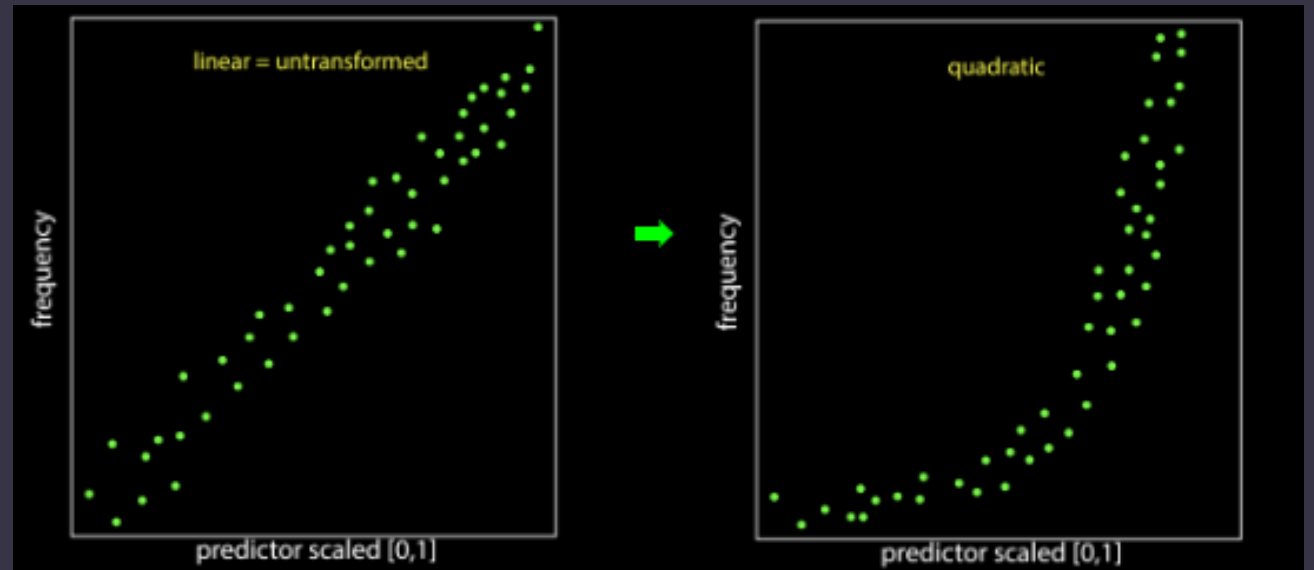
1. $\bar{x} = 4$
2. $x_i \in [3,4,5]$



MaxEnt: Transformations

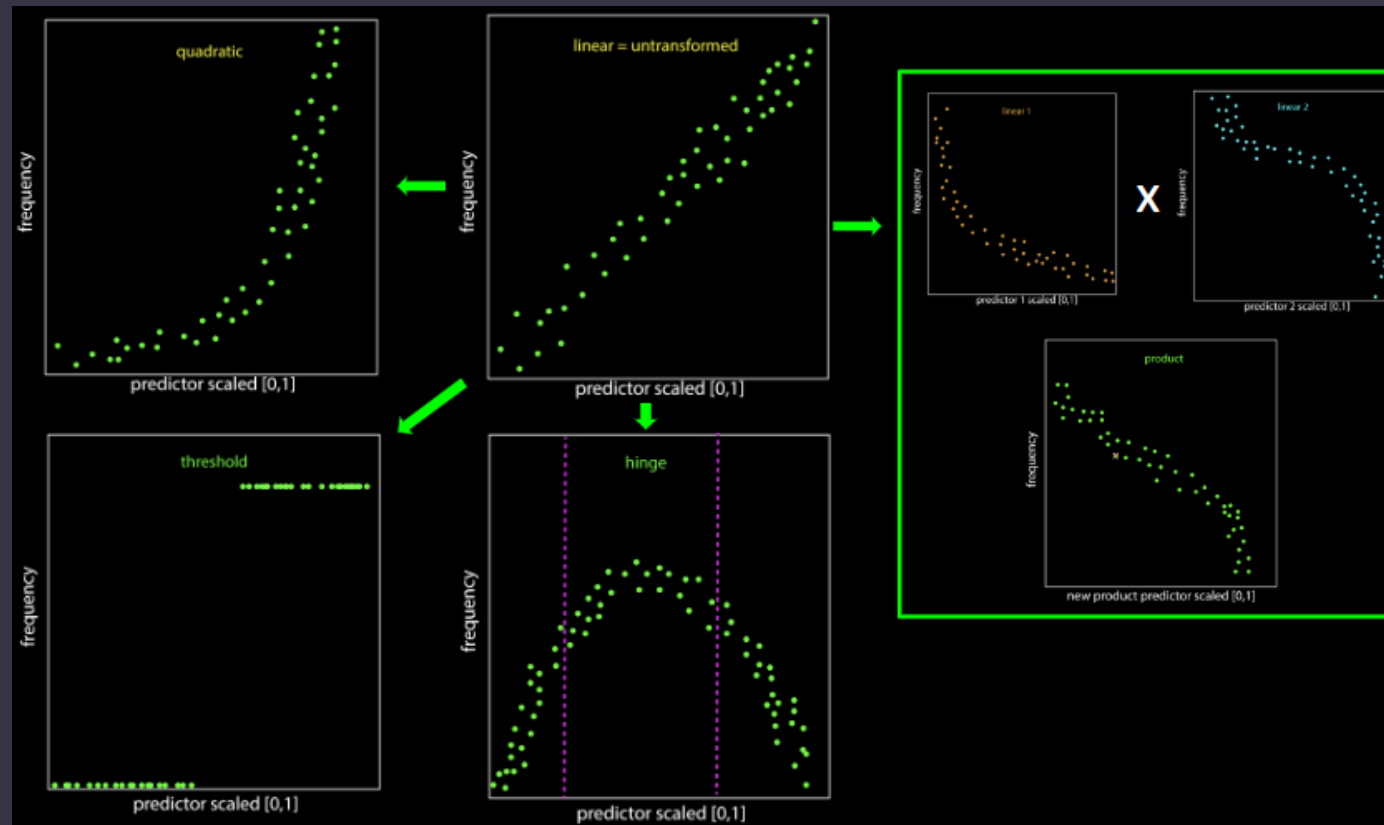
Two Options:

1. Fit using many different model types
2. Transform our predictors and fit one (or multiple) models to our transformations (features)



MaxEnt: Feature Classes

Which features are used in the MaxEnt model is known as feature class selection



MaxEnt: Regularization

Another means of tuning MaxEnt models is through tweaking the regularization multiplier

This is the degree to which the model is penalized for complexity (many features, sharp responses) via shrinking coefficients

Default is 1, range is 0-5

Higher == More penalization (less complex)

Lower == Less penalization (more complex)



MaxEnt: In a Nutshell



MaxEnt is a ML algorithm that works by fitting the data s.t. informational entropy is maximized

This is achieved by tuning 1) feature class (predictor transformation) and 2) regularization multiplier (complexity penalization)

MaxEnt is by far the most prevalent algorithm in the broader SDM literature

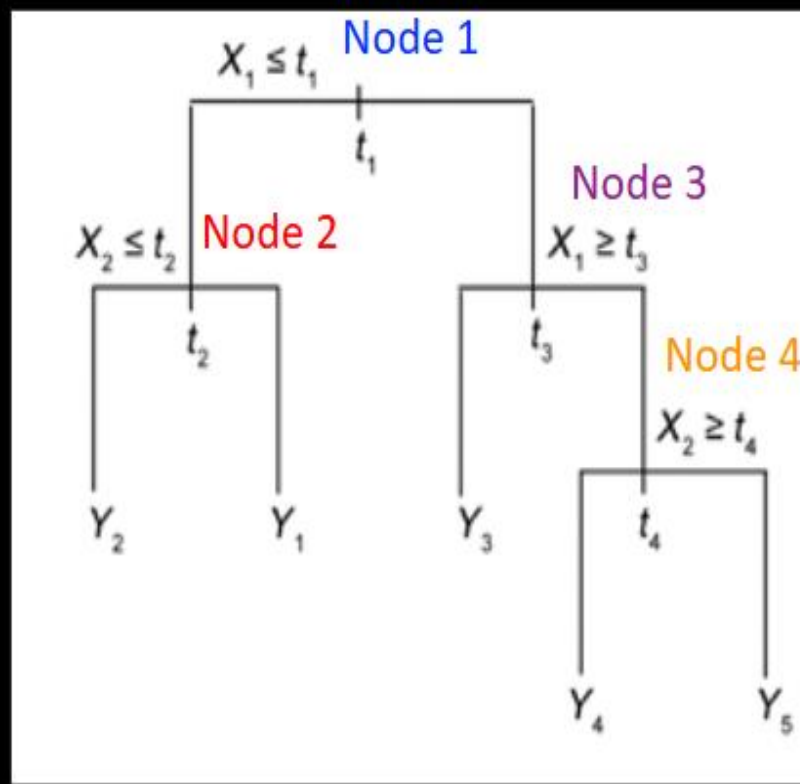
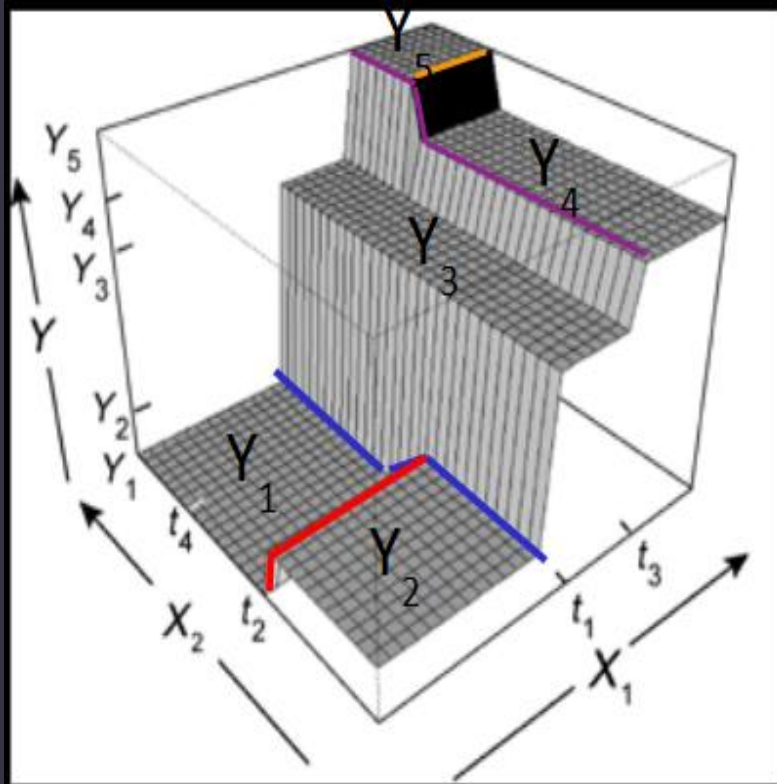
This paper gives a good explanation of the underlying statistical properties

A statistical explanation of MaxEnt for ecologists

Jane Elith ✉, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, Colin J. Yates

Decision Trees: Variance Partition

Classification and Regression Trees: Partition variance in environmental factors correlated with probability of presence



Prediction of probability of occurrence: For site A with conditions $[X_1, X_2]$ choose a path at each node conditional on the probability of drawing $[X_1, X_2]$ randomly

Decision Trees: Variance Partition

| Predictors | | | | | Response |
|---------------|-----------------|---------------------|-------------------|-----------|----------|
| Isothermality | Temp. Wettest Q | Precip. Seasonality | Precip. Coldest Q | Elevation | Presence |
| 31 | 190 | 25 | 113 | 88 | 1 |
| 34 | 85 | 22 | 218 | 693 | 1 |
| 41 | 125 | 64 | 289 | 53 | 1 |
| 33 | 131 | 45 | 438 | 37 | 1 |
| 34 | 68 | 14 | 125 | 12 | 0 |
| 38 | 60 | 21 | 250 | 91 | 0 |
| 31 | 37 | 15 | 147 | 1006 | 0 |
| 25 | 74 | 22 | 208 | 928 | 0 |

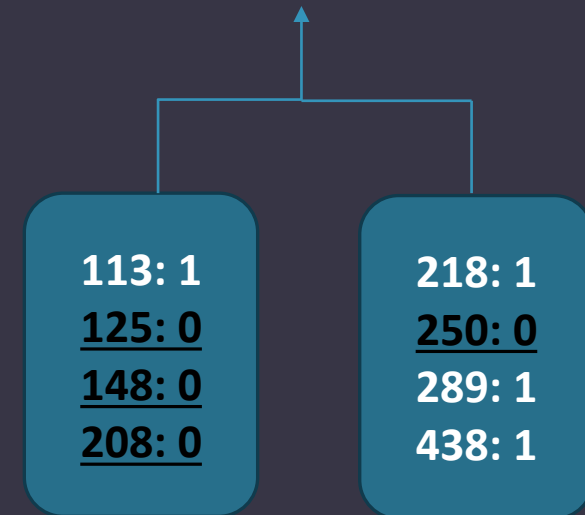
Decision Trees: Variance Partition

| Precip. Coldest Q | Presence |
|----------------------|----------|
| 438 | 1 |
| 289 | 1 |
| 250 | 0 |
| 218 | 1 |
| 208 | 0 |
| 147 | 0 |
| 125 | 0 |
| 113 | 1 |

Mostly present; >217

Mostly absent; <208

Precip Coldest Q > 217 || < 208

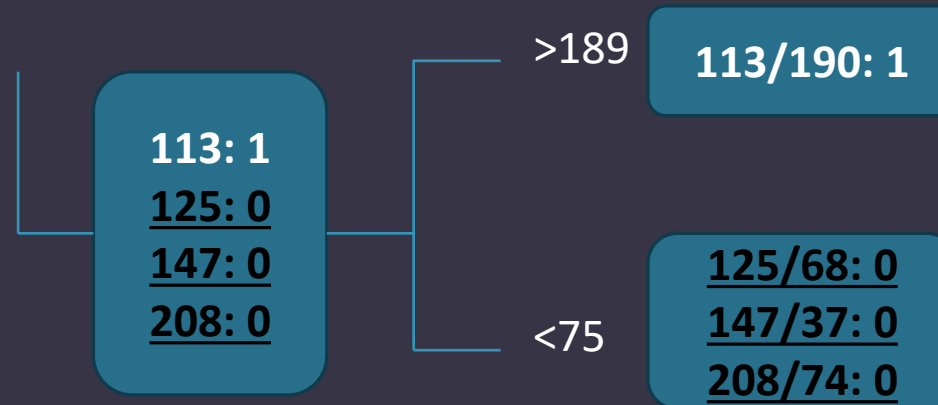


Decision Trees: Variance Partition

| Precip. Coldest Q | Temp. Wettest Q | Presence |
|----------------------|--------------------|----------|
| 113 | 190 | 1 |
| 438 | 131 | 1 |
| 289 | 125 | 1 |
| 218 | 85 | 1 |
| 208 | 74 | 0 |
| 125 | 68 | 0 |
| 250 | 60 | 0 |
| 147 | 37 | 0 |

How could we further partition data to more accurately predict occurrence?

Answer: adding another split

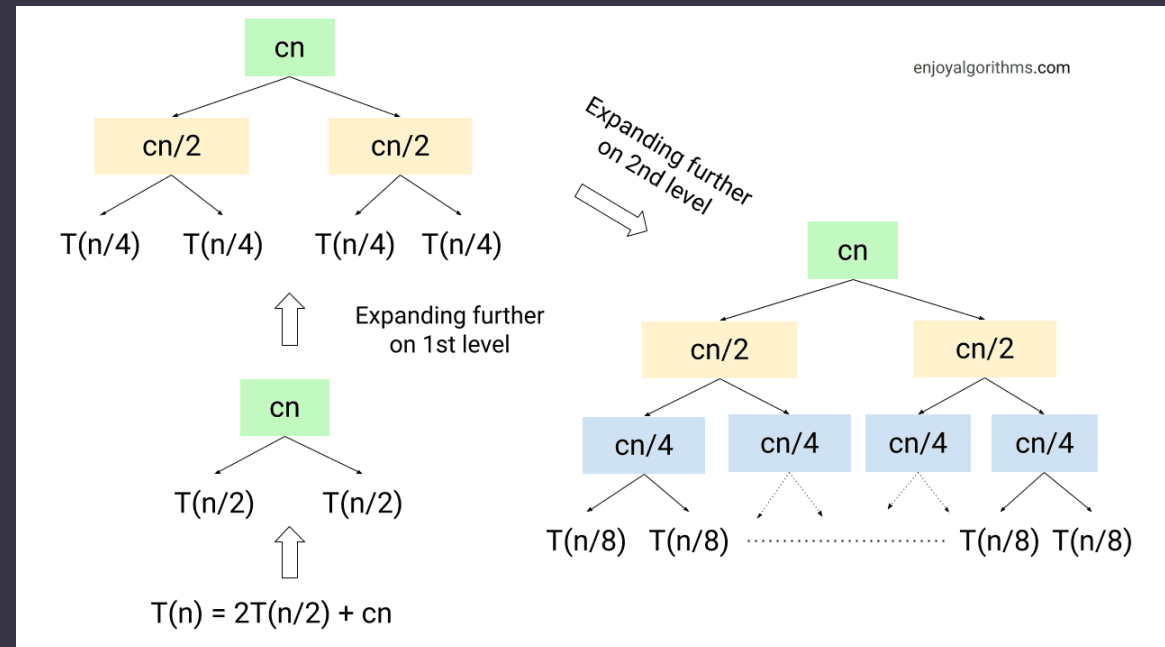


Tree Complexity

Tree Complexity (TC): The number of splits in a tree

By tweaking this, for absolutely any data set would could construct a tree that would perfectly fit the data

So, why don't we just do this all the time?

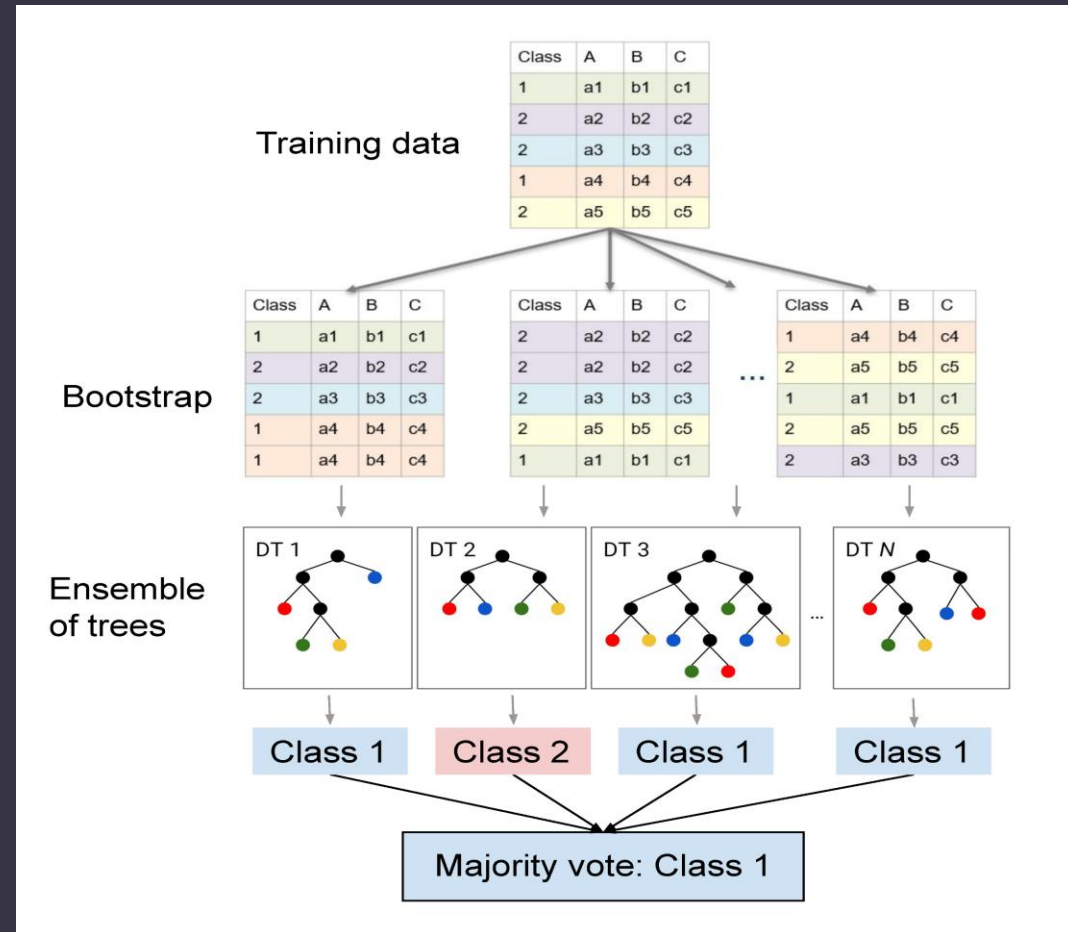


Random Forest

Random Forest (RF) is one of the most prevalent SDM algorithms

- 1) Takes *bootstrapped* samples from the training data
- 2) Constructs individual tree for each of these bootstrap samples
- 3) *Aggregates* prediction via majority vote (51 yes: 49 no == yes)

Bootstrap + Aggregate = Bagging

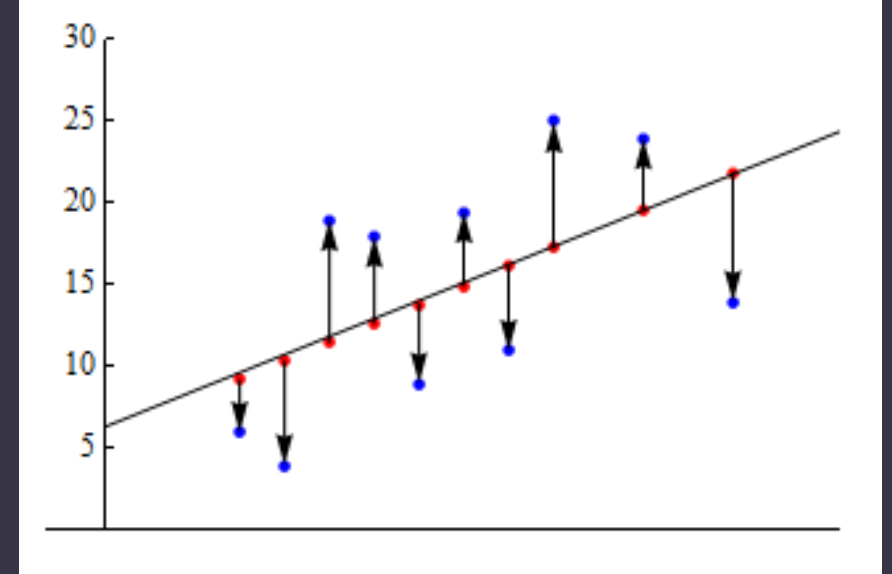


Gradient Boosting

Gradient boosting makes sequential models, each of which explains variation not explained by the previous model

General Approach

1. Tune model 1 to best explain some response variable (here, occurrence)
2. For each point estimate residual deviance
3. Continue until deviance no longer decreases
4. Each model contributes less than previous
=Regularization/Shrinkage

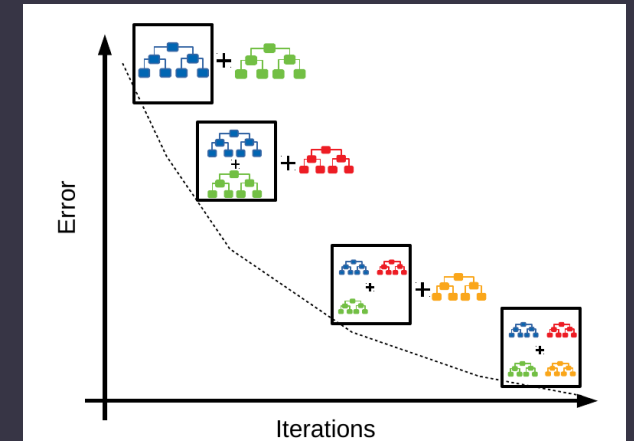


BRT == GBM

Gradient boosting is a technique applied in boosted regression trees (BRT; also known as gradient boosting machines [GBM])

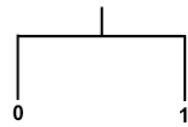
BRT Approach

1. Divide data into calibration and evaluation data sets
 - hold out data = bag fraction (fraction of data withheld)
2. Generate a random forest type decision tree
 - depth of tree is determined by tree complexity
3. Estimate deviance for calibration and bag-fraction
 - this is hold-out deviance
4. Return to step 1. Continue until residuals no longer decrease
5. Each successive model's contribution is scaled by a learning rate

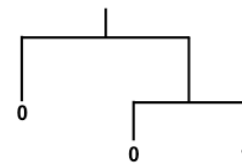


Learning Rate and Tree Complexity

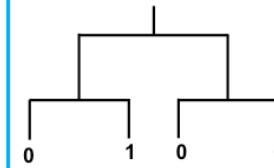
Tree complexity (tc)



$tc = 1$
no interactions



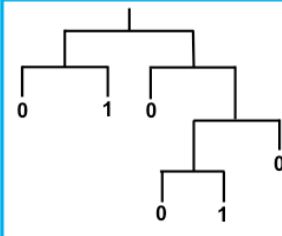
$tc = 2$



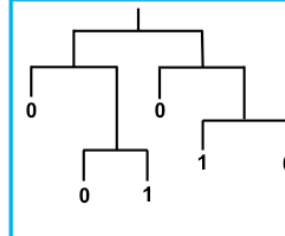
$tc = 3$

Learning rate (lr)

Random subset



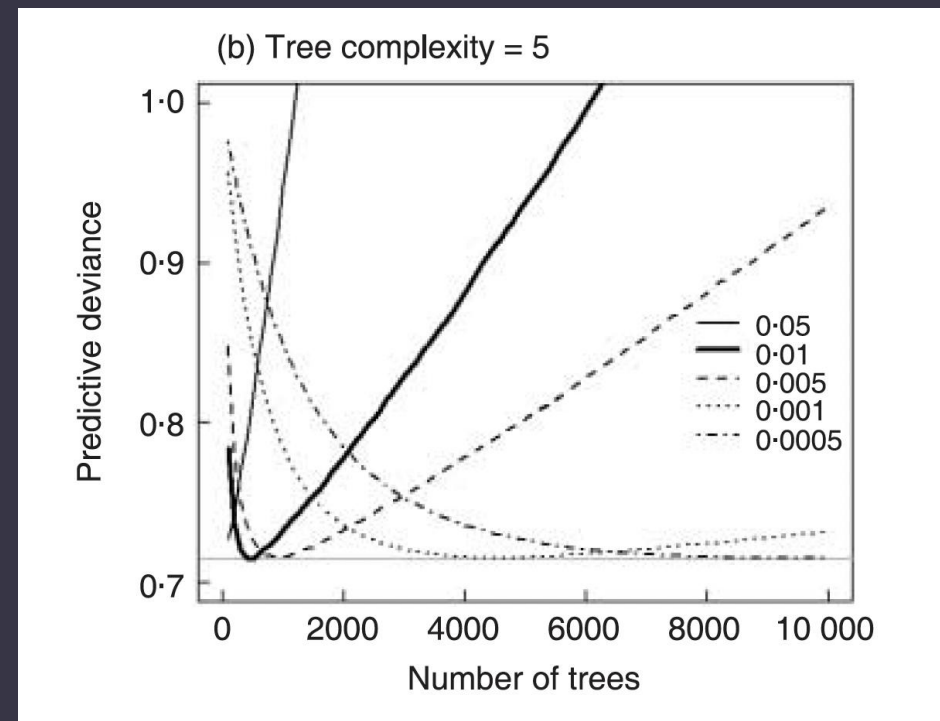
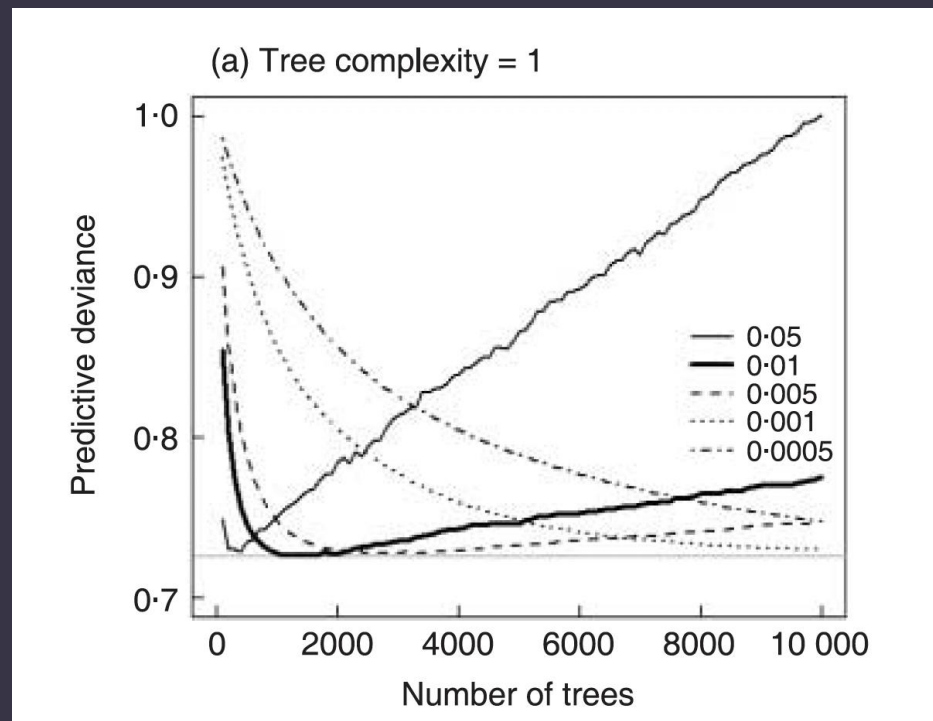
Random subset



lr : contribution to
growing model
small value = many trees

Hold-Out Deviance

Predictive deviance varies as a function of 1) learning rate, 2) tree complexity, and 3) how many data points you have



Decision Trees: In a Nutshell




Random forest uses **bootstrapped** samples to create many trees in parallel then **aggregates** the predictions via majority voting/consensus (this means RF uses **bagging** to make predictions)

BRT makes many sequential trees to minimize predictive deviance from an initial tree (via gradient boosting)

Modelling species presence-only data with random forests

Roozbeh Valavi , Jane Elith, José J. Lahoz-Monfort, Gurutzeta Guillera-Arroita

A working guide to boosted regression trees

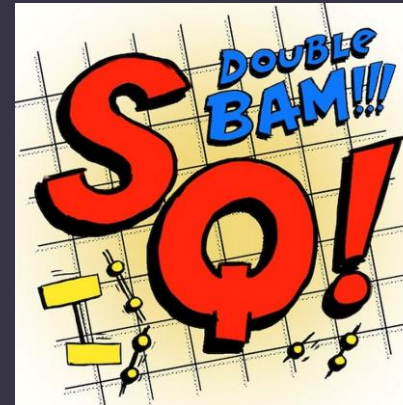
J. Elith , J. R. Leathwick, T. Hastie

Final Remarks

This is by no means comprehensive, but (hopefully) provides a working introduction

Thankfully, there are many, many good resources out there when it comes to learning the stats


If you apply even a bit of thought into the construction of your models, you are doing more than some published works



StatQuest YouTube Channel

Literature Guides

A statistical explanation of MaxEnt for ecologists

Jane Elith , Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, Colin J. Yates

A working guide to boosted regression trees

J. Elith , J. R. Leathwick, T. Hastie

Modelling species presence-only data with random forests

Roosbeh Valavi , Jane Elith, José J. Lahoz-Monfort, Gurutzeta Guillera-Arroita

Species distribution modeling

Robert J. Hijmans and Jane Elith

SDMs in R
Written by two of the GOATS