Marine Ecological Modelling
Global Climate Change

The diversity of algorithms of ENM and ensembles

Jorge Assis, PhD // jmassis@ualg.pt // jorgemfa.medium.com
2020, Centre of Marine Sciences, University of Algarve

Algorithms to fit ecological niche models

Geographic algorithms

Use the location of occurrences, and **do not rely on the values of environmental variables** at these locations (not a niche model *per se*).

e.g.,

Circles;

Geographic Distance.

Major limitations

Does not use environmental variables to predict species occurrence;

Does not make quantitative predictions (e.g., probability of occurrence).

Algorithms to fit ecological niche models

Profile algorithms

Profile methods **only consider presence data, not absence or background data**. Only rely on the values of predictor variables at presence locations to find similar environmental regions.

e.g.,

Bioclim;

Surface Range Envelope.

Major limitations

Susceptible to overprediction of potential distributions;

Does not make quantitative predictions (e.g., probability of occurrence).

Algorithms to fit ecological niche models

Regression model algorithms**

Build a function that estimates the effect of different environmental predictors on the distribution of a species.

e.g.,

Generalized Linear Model;

Generalized Additive Model;

Multivariate Adaptive Regression Splines.

Major limitations

Some are susceptible to overfitting** (GAM);

Need relatively large datasets. Also, the more predictor variables (e.g. environmental layers), the larger the sample size required;

Sensitive to outliers.

Generalized Linear Models (GLM) are an **extension of ‘simple’ linear regression models.**

Able to deal with non-normal distributed data. Linear models imply a straight line describing the relationship between the response and the predictor variables (an assumption often violated in ecological data).

In GLMs, the relationship between the response and the predictors is not linear, and a **link function provides a transformation of the response** so that the transformed response is linearly related to the predictors.

A GLM with binomial data, such as the presence/absence of a species, is commonly called “logistic regression”. In this case, the link function is a logit function, which is the log of the odds ratio (probability of presence/probability of absence).

Advantages of GLM:

Able to deal with categorical predictors (e.g., subtract type);

Relatively easy to interpret providing a clear understanding of how each of the predictors are influencing the outcome;

Less susceptible to overfitting** than other algorithms.

Algorithms to fit ecological niche models

Machine learning algorithms**

Construct a function that estimates the effect of different environmental variables on the distribution of a species.

e.g.,

MaxEnt;

Boosted Regression Trees;

Artificial Neural Network.

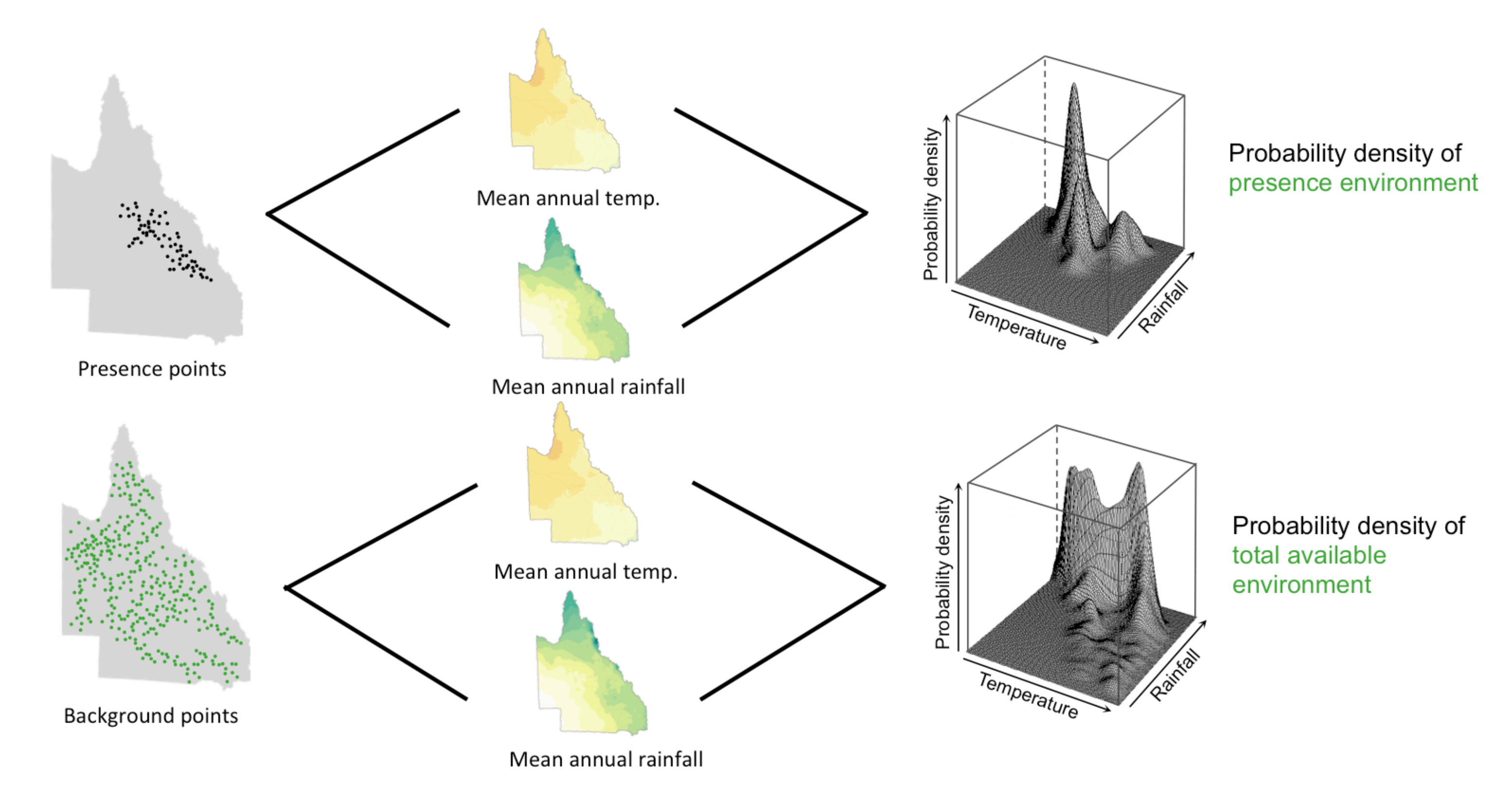
Major limitations

Susceptible to overfitting if not properly parameterized**;

Some need at least 2 predictor variables to run (BRT).

Maximum Entropy (in short MaxEnt) is the most widely used algorithm in ENM. Predicts occurrences by **comparing the density in the environmental conditions at the locations where the species has been found, to the environmental conditions across the study region** - samples a large number of background points.

MaxEnt gives the relative environmental suitability of a species for each point in the study area (i.e., the potential distribution).



Adapted from Elith et al. (2011) *A statistical explanation of MaxEnt for ecologists*. Diversity and Distributions, 17, 43-57.

Advantages of Maximum Entropy:

Can use both continuous and categorical predictor variables;

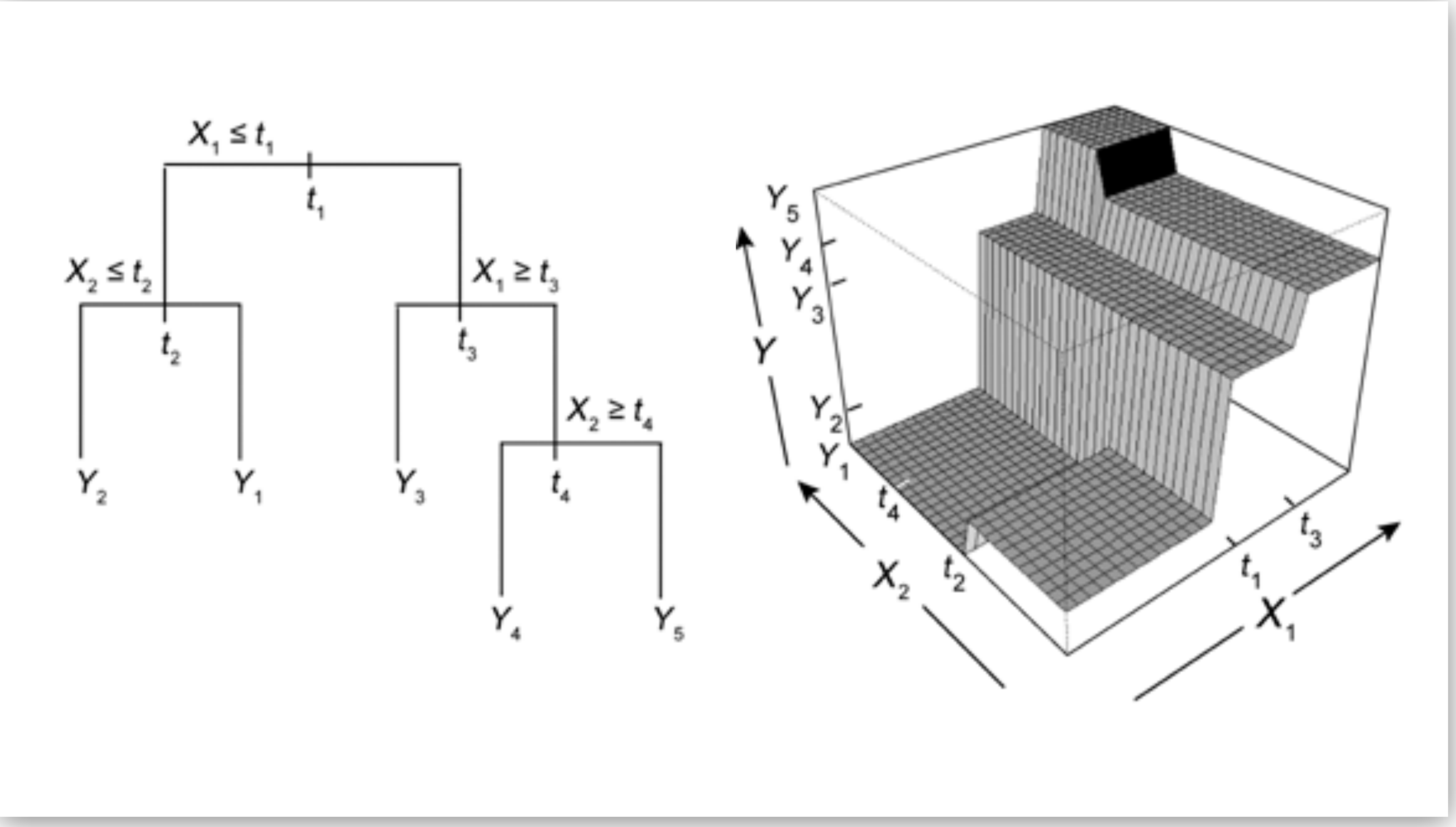
Includes a proper set of parameters to protect against overfitting;

Can detect interactions* between predictor variables;

* An interaction occurs **when a predictor has a different effect on the outcome depending on the values of another independent predictor**.

Boosted Regression Trees (in short BRT) is one of the most promising newer statistical approaches for ENM, which combine the strengths of two algorithms **regression trees** (models that relate a response to their predictors by recursive binary splits) **and boosting** (combines simple models to give improved predictive performance).

BRT gives the probability of a species occurring at each point in the study area (i.e., also the potential distribution).

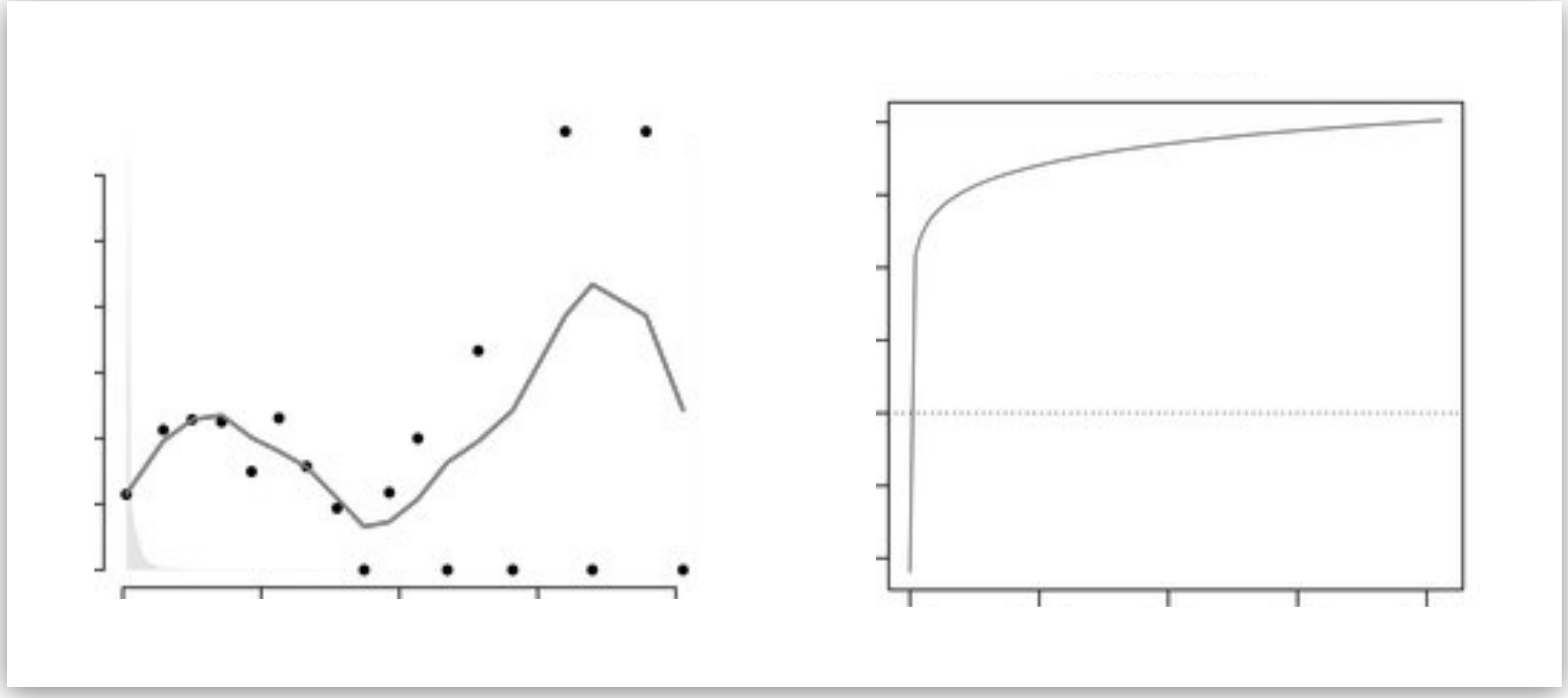


Advantages of BRT:

Boosted regression trees can **handles different types of predictor variables, accommodates missing data** and can fit **complex nonlinear relationships**, and automatically handle interaction effects between predictors.

It allows forcing monotonicity relationships between the output and predictor variables, which strongly reduces overfitting.

Overfitted response vs. monotonocity forcing eliminating overfitting



Performance of algorithms to fit ENM

Most algorithms have been tested and compared.

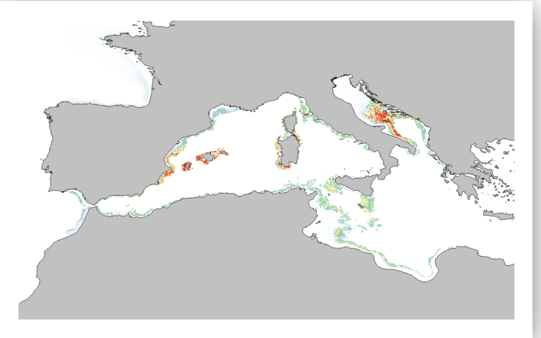
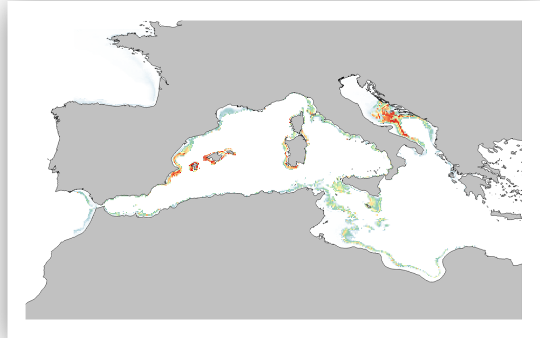
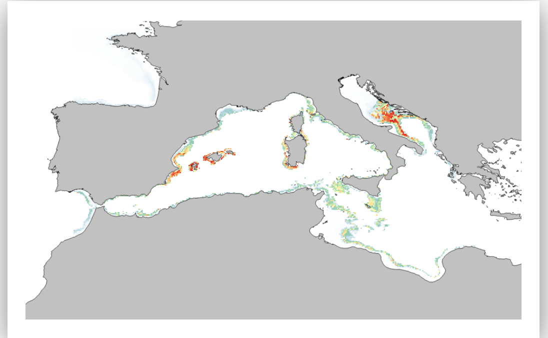
Differences in performance among different algorithm types tend to be smaller than differences among species. Some comparisons yield conflicting results about relative performance of algorithms, which is unsurprising as each study uses different data (i.e., species and environment).

Yet, **machine learning algorithms**, like MaxEnt and especially those that incorporate model averaging (e.g., BRT) **tend to have better performance** than more simple statistical methods such as GLMs.

(Segurado & Araújo, 2004; Elith & Graham, 2009)

Ensembling models

Instead of using a single-model to investigate species distributions, one can conduct an **ensemble experiment**. This **reduces the uncertainty of algorithms by combining their outputs** (e.g., habitat suitability maps) **with a general statistics** (mean, median, variance).

mean( , ) = 

For example, one can synthesise the results of two algorithms in a single map by averaging their outputs. Also this approach **allows to identify regions of agreement / disagreement between algorithms** (e.g., variance or standard deviation)