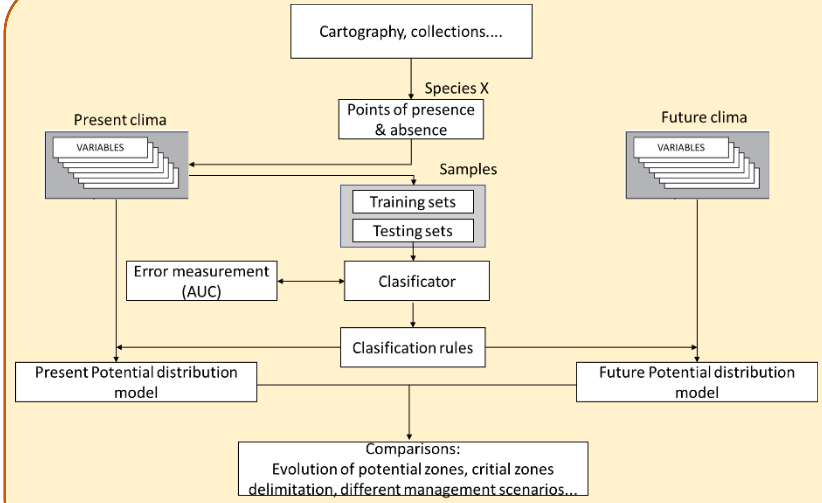
Species Distribution Models

Simplifying a complex reality

Mathematical thinking allows us to capture part of the complex world around us into formulas and relational structures that can be more easily analyzed than the real world. In ecology, Species Distribution Models (SDMs) try to capture the adequacy of a particular space for the presence of an organism (i.e. a species), taking into account a set of environmental variables of the area where the species live.

SDMs integrate specific values of environmental and human variables with the presence or absence of a particular species, capturing the ecological niche of the species under study. Essentially, they represent the fundamental niche of a species (also called potential niche) taking information of the realized niche of the species. Most of the time, models do not include biotic interactions and refer only to the environmental part of the ecological niche, without interaction between the points, which is called the “Grinellian niche” in classical Ecology.



Cartography, collections....
Species X
Points of presence & absence
Present clima
VARIABLES
Future clima
VARIABLES
Samples
Training sets
Testing sets
Error measurement (AUC)
Clasificator
Clasification rules
Present Potential distribution model
Future Potential distribution model
Comparisons:
Evolution of potential zones, critial zones delimitation, different management scenarios...

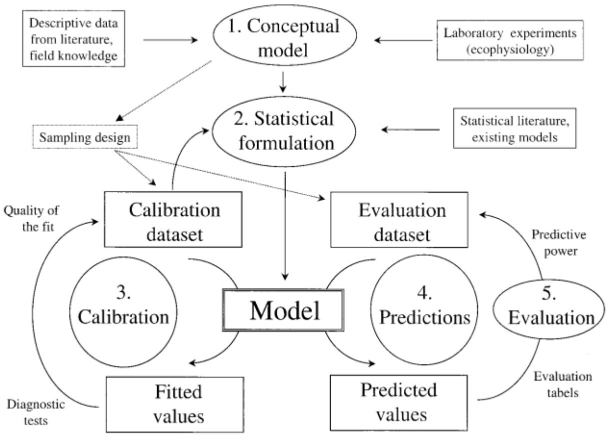
Figure 1. *Modified from Mateo et al. 2011*

- Presence and/or absence data is nowadays available in many repositories such as **museum collections**, **online databases** such as GBIF, and **private records** from NGOs, **governments** and **scientific institutions.**
- Present climatic and environmental **parameters** can be obtained locally through **scientific equipment** such as meteorological stations or CTDs (in the marine environment), and globally through **satellite imagery** and/or interpolation of local measurements.
- In recent years, the development of **climatic models** allowed for **future predictions** of some **climatic variables** such as Temperature and Precipitation, which can be used in the models to predict future potential distribution of species.

Following Guisan & Zimmermann 2000, the model building process is composed of five steps:

1. Conceptual model refers to the process of formulation of an ecological model which should contain the underlying conceptual framework. Here, the user should remind which environmental variables will be used, which niche is seeked for, the organism selected, etc.

2. Statistical formulation addresses the choice of an appropriate algorithm for predicting a particular response variable. Often more than one technique may be appropriately applied.



Descriptive data from literature, field knowledge
1. Conceptual model
Laboratory experiments (ecophysiology)
Sampling design
2. Statistical formulation
Statistical literature, existing models
Quality of the fit
Calibration dataset
Evaluation dataset
Predictive power
3. Calibration
Model
4. Predictions
5. Evaluation
Diagnostic tests
Fitted values
Predicted values
Evaluation tabels

Figure 2. *Modified from Guisan & Zimmermann 2000.*

3. Calibration refers to the adjustment of the mathematical approach previously selected (i.e. adjusting knots in GAMs, or number of trees in BRT).

4. Predictions are extrapolation of the model based on particular points to the rest of the area. From the mathematical formula and the predictors, it is possible to predict the species presence and/or abundance.

5. Evaluation refers to validation procedures such as checking the correlation between predicted and sampled measurements, to verify the quality of the model.

SDMs have strongly evolved in recent years and include a wide range **of individual modelling techniques,** which consist of different mathematical and conceptual approaches to design the models, and that can be classified into three different groups, depending on the method through which they determine the adequacy of a specific location for a species habitat:

Discriminant techniques: These methods require presence and absence data to build a criterion to classify the space into adequate, or not adequate, for the species to live. Here we find most of the techniques used in the present:

- **Classification and Regression Trees** → Classification trees partition the ecological space in rectangles using a series of rules to identify regions with the most homogeneous responses to predictors. Classification and Regression Trees differ in the way through which they assign a constant to each region. Trees are insensitive to outliers but are usually not as accurate as other methods such as linear modelling.

- **Boosted Regression Trees**→ This method consists in the boosting of many regression trees. Boosting is a method for improving accuracy based on the **sequential and stagewise construction of the final model**. This procedure constructs a Regression Tree and then uses methods to increase emphasis on poorly modelled observations. Different boosting algorithms differ in the way they measure the lack of fit for each observation

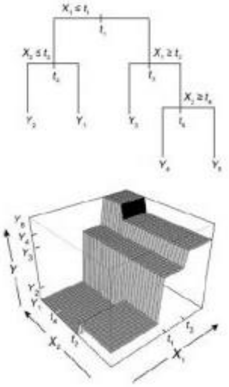


***Figure 3**. A single decision tree (upper panel), with a response Y, two predictor variables X_1 and X_2 and split points t_1, t_2, etc. The bottom panel shows its prediction surface.*

- **Generalized Linear Models (GLMs)→** These models were developed in 1972 to relate responses following any of the **exponential family distributions** (Normal, Binomial, Poisson, gamma or negative binomial), to predictor variables. Instead of using the classical Ordinary Least Squares method (OLS), this technique uses iterative weighted linear regression to obtain maximum likelihood estimates of the parameters for the predictors.

- **Generalized Additive Models (GAMs)→** GLMs require a linear relationship between each predictor variable and the response variable, but this is usually not the case between species abundance and environmental variables. GAMs allow for a nonlinear relationship between response and predictors by using **smooth functions** of predictor variables. This allows for more flexibility in the inclusion of non-linearly related predictors to the models while maintaining the possibility to deal with those which do it linearly.

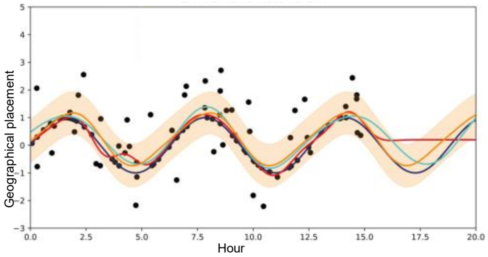


Figure 4. GAM model fit in a non linear way. This example uses polynomial fit s a smooth function to relate the geographical movement of an organism. This model represents a circadian cycle of about 5h.

- **Bayesian methods→** Bayesian thinking offers a probabilistic instead of a deterministic approach to problems. In ecological modelling, this translates into **probability distributions for parameters** of interest in the models. As an example, Bayesian inference can be beneficial to model temporal or spatial autocorrelation, individual level random effects, and hidden states.

- **Maximum entropy approximation (MaxEnt)→** This is the only Discriminant technique that does not require absence data, since it creates a background pseudo-absence layer with which it is able to construct a model. The idea of Maxent is to estimate the probability distribution of **maximum entropy** (i.e. closest to uniform) to assess the probability distribution of a species.

Descriptive techniques: Used to describe the available data: only require presence data. These techniques were the first mechanisms developed in the field and include:

- **Environmental space definition techniques (i.e. BIOCLIM, HABITAT)→** These procedures relate, using different approaches, the bioclimatic envelope of organisms to a number of bioclimatic variables.

- **Mathematical distance methods (i.e. DOMAIN) →** These procedures use different Environmental-Distance measurements to model habitat suitability for Species distributions. Some of them are based on density of observations and enable a good fit with complex distributions such as nongaussian.

Mixed techniques: Different individual **models can be combined** to obtain a **consensus model** using packages such as BIOMOD in R.

References:

- Guisan, A. & Zimmermann, N.E. Predictive habitat distribution models in ecology. Ecological Modelling **135,** 147-186 (2000)
- Mateo, R.G., Felicisimo, A.M. & Muñoz, J. 2011. Modelos de distribución de especies: Una revisión sintética. Revista Chilena de História Natural. **84 ,** 217-240 (2011).
- Faith, D.P., Minchin, P.R. & Belbin, L. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69, 57–68 (1987)
- Elith, J., Leathwick, J.R. & Hastie,T. A working guide to boosted regression trees. Journal of Animal Ecology **77,** 802-813 (2008)