Introduction to species distribution models

Maria Grazia Pennino

Statistical Modeling Ecology Group

Instituto Español de Oceanografía (IEO, CSIC)

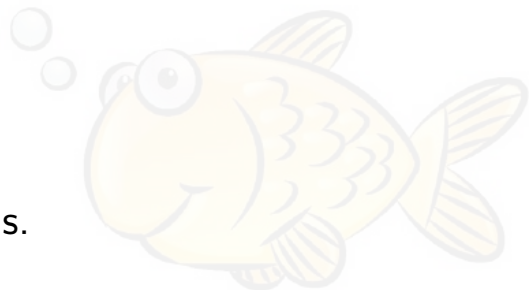
The Statistical Modeling Ecology Group

What is SMEG?

The SMEG is an inter-disciplinary research group, linking statisticians, biologist and ecologists (https://smeg-bayes.org). Our remit is to develop and apply advanced statistical methods to analyzing ecological/biological data, thereby improving the understanding and management of natural resources and their environment.

Active areas of research include:

- Spatial-temporal distribution of biological resources;
- Economic evaluation of natural resources;
- Modelling population dynamics;
- Modelling ecological point process data;
- Integrated biological and socioeconomic models.



1 Spatial data

Spatial data

Depending on the type of data and the purpose of the spatial analysis itself, we classify spatial data sets into one of three categories:

- lattice data (or areal data),
- point pattern data,
- geostatistical data.
- The main dissimilarity between these resides in the set of observations.

Lattice data

- Observations associated with spatial regions.
- It is partitioned into a finite number of geographical areas with well-defined boundaries.
- This type of data have a clear neighborhood structure.
- *Examples*:
  - observed locations from agricultural field trials,
  - cancer rates for Washington counties,
  - catch data for state, etc.

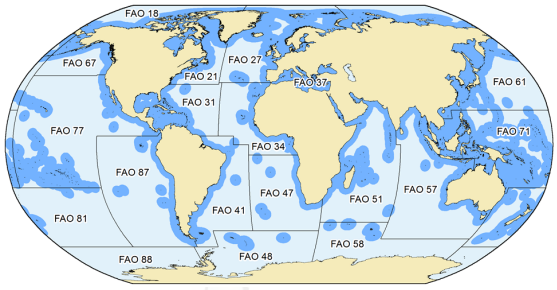


Figura: Catch allocation in the Exclusive Economic Zones (EEZs).

Point pattern data

- Locations are the variable of interest.
- The own locations determined phenomena that occurred randomly in one place.
- There is no attribute involved.
- *Examples*:
  - locations of earthquakes,
  - locations where fishers go to fish, etc.

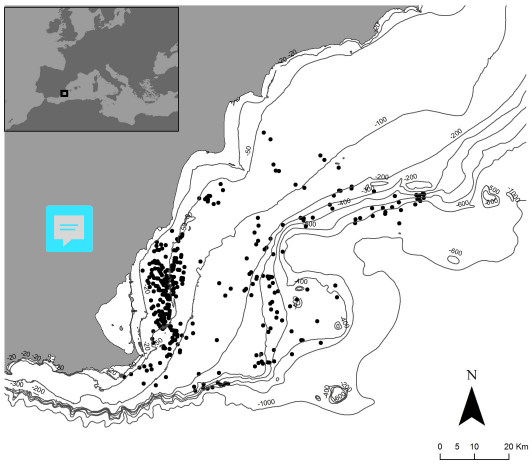


Figura: Fishery hauls in the south-eastern Mediterranean Sea.

Geostatistics

- Observations describe a process where the locations vary continuously in space.
- These data are characterized by spatial dependence between the locations
- *Examples*:
  - occurrence or abundance of species in a region,
  - annual acid rain deposition in a town, etc.

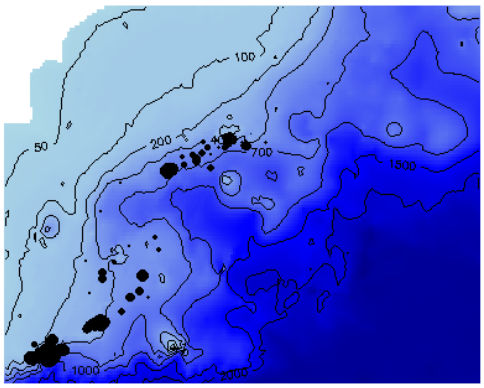


Figura: Black dots rapresent the abundance of the European hake in the south-eastern Mediterranean Sea.

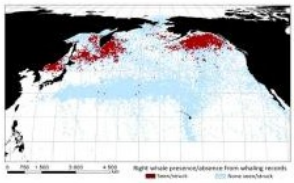
2 SDMs

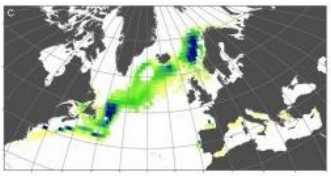
SDMs

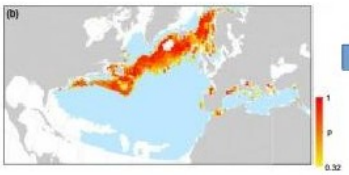
- SDMs is also known under other names including climate envelope-modeling, habitat modeling, and niche-modeling.
- Typical examples:
  - identifying the suitable habitat of a species;
  - predicting the response of species to climate changes;
  - identifying are with more biodiversity of richness;
  - identifying managing conservation areas;
  - assessing reintroducion areas for species;
  - predicting where alien species could extend their distributions;
  - predicting the possible impact of antropic effects on species distribution (Marine Spatial Planning, Cumulative Effect Assessment).

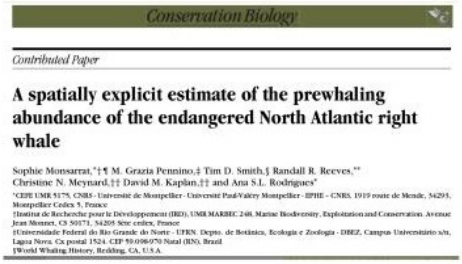
SDMs: some examples

(1) Understanding the historical distribution and abundance of marine mammals using SDMs, historical and environmental data.

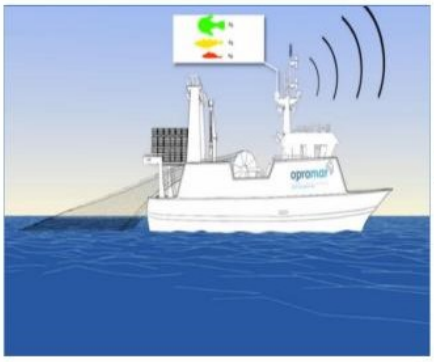






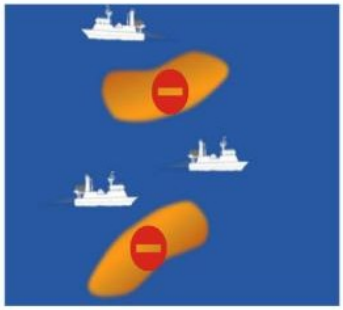


SDMs: some examples



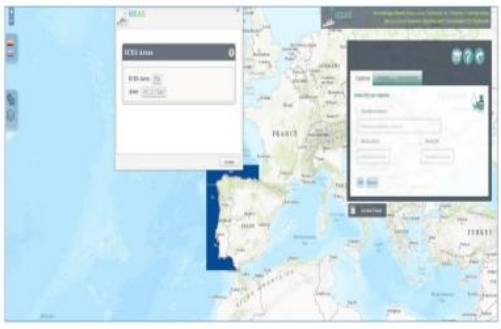
iSEAS project

Real-time transmission of the collected data trough the red-box system



High-density discards areas

An automatize and quickly SDM that analyze data and identify high-density discard areas in a on-line web-server.

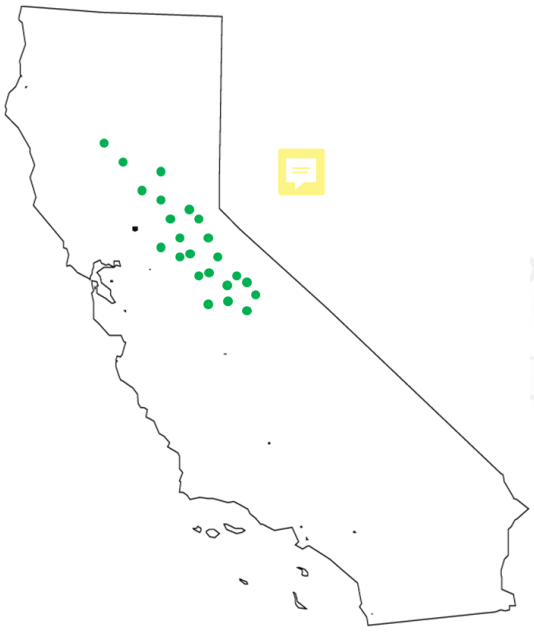


SDMs

- The core problem in all these cases: Predict the spatial distribution of a phenomenon (species occurrence or abundance, etc.).
- HOW???: links spatially referenced records of the studied phenomenon with predictor variables for two main purposes:
  - (1) understand the relationships between the species and its environment;
  - (2) predict the found relationship outside the sampled locations.

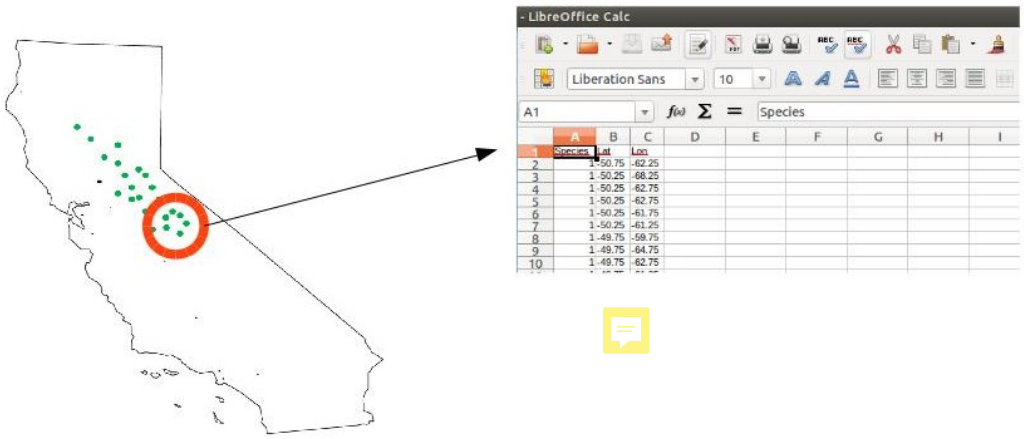
SDMs

In the most general sense, here's how SDM works.
(1) Locations of occurrence of a species (or other phenomenon) are compiled:



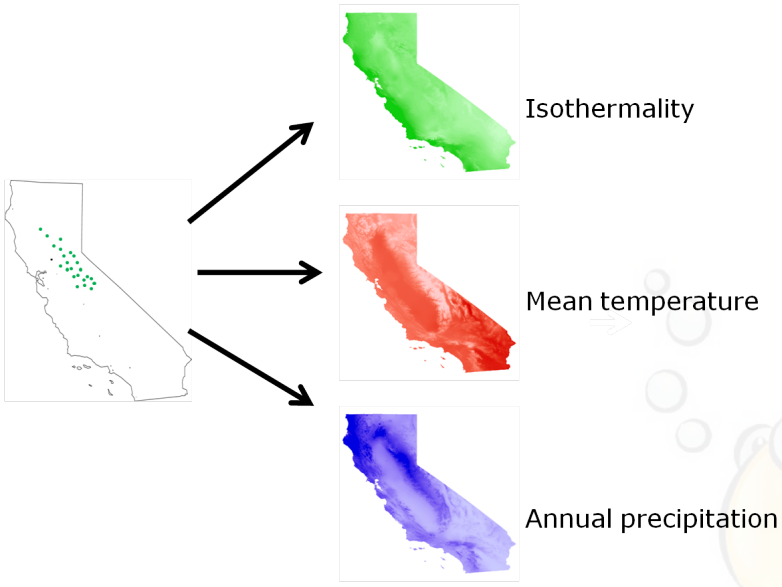
SDMs

In the most general sense, here's how SDM works.
(1) Locations of occurrence of a species (or other phenomenon) are compiled:



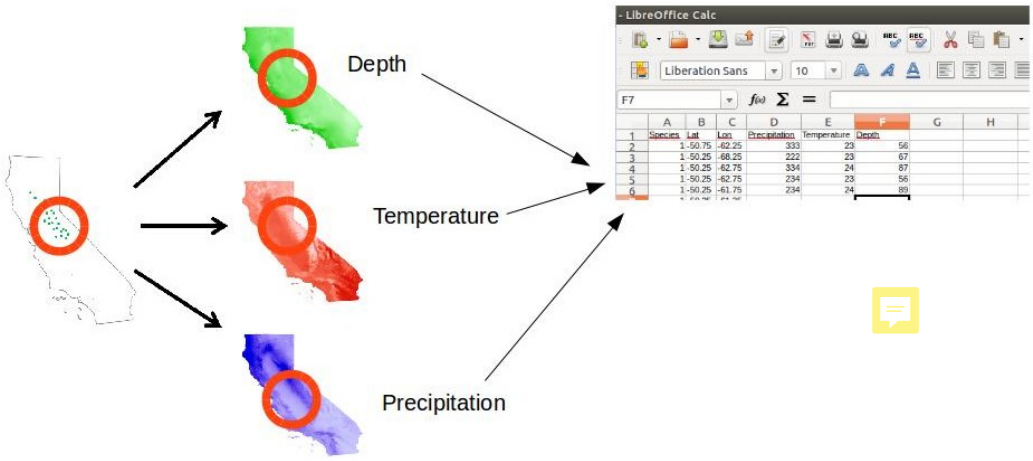
SDMs

(2) Values of environmental predictor variables (such as climate) at these locations are extracted from spatial databases:



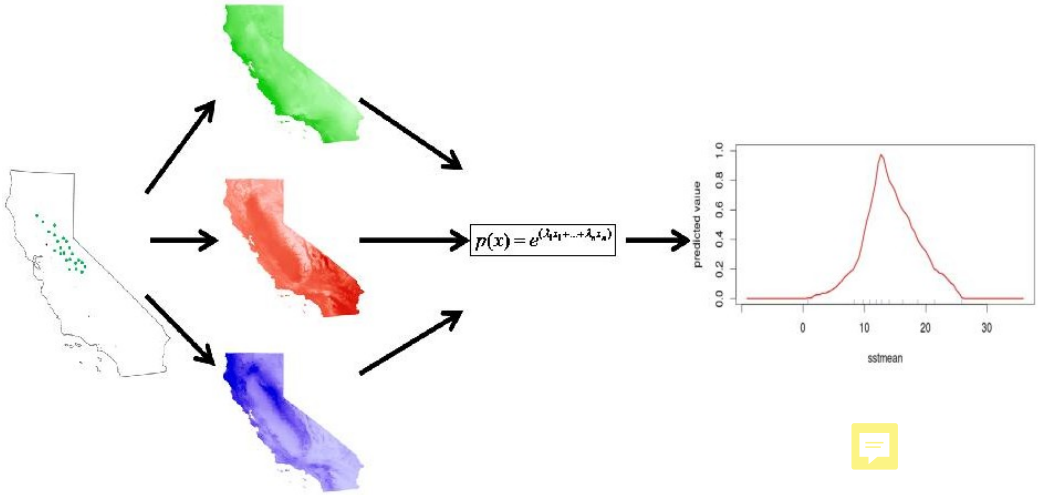
SDMs

(2) Values of environmental predictor variables (such as climate) at these locations are extracted from spatial databases:

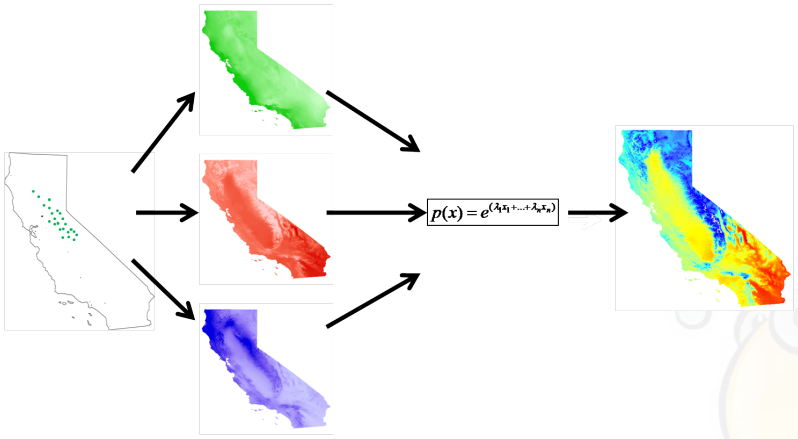


SDMs

(3) The environmental values are used to fit a model predicting likelihood of presence, or another measure such as abundance of the species:



(4) Once we know the relationship among species and the environmental variables we can extrapolate it and predict the probability of occurrence of the studied species in the entire area on interest (and/or for future or past climate).



3 Data types

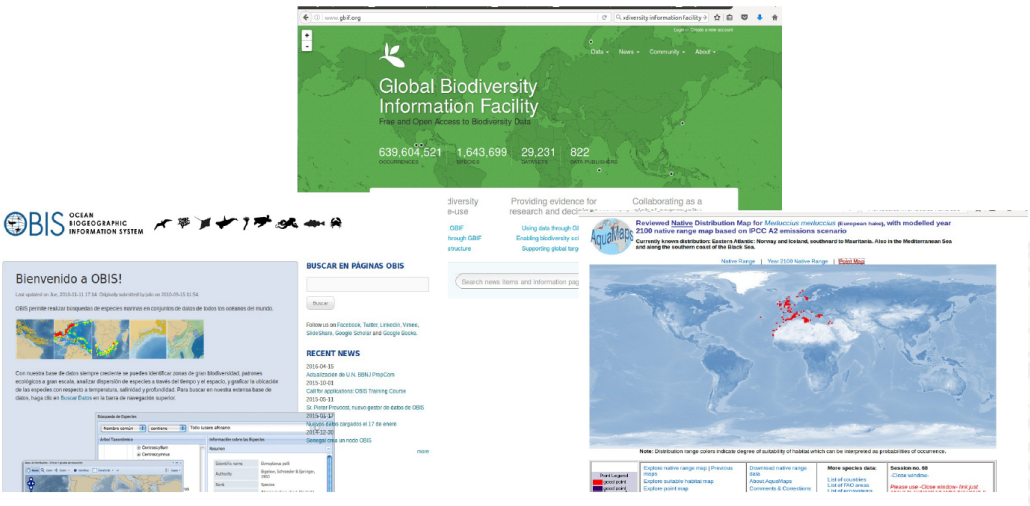
Biological data: response variable

Data describing the known distribution of a species may be obtained in a variety of ways:

- Personal collection: occurrence localities can be obtained during field surveys.
- Large surveys: distribution information may be available from surveys undertaken by a large number of people. *i.e.* Araújo et al. (2005a) built distribution models using data from *The new atlas of breeding birds in Britain and Ireland: 1988-1991* (Gibbons et al. 1993), which represents the sampling effort of hundreds of volunteers.
- Museum collections: occurrence localities can be obtained from collections in natural history museums.
- Opportunistic surveys: data can be obtained from correlated activities (fishery, cetacean sightings from whale watching, etc.)

Biological data: response variable

- On-line resources: distributional data from a variety of sources are increasing being made available over the internet. For example, Global Biodiversity Information Facility, AquaMaps and Ocean Biogeographic Information System.



Biological data: response variable

Species distribution data may be either:

- Presence-only: records of localities where the species has been observed.
- Presence/absence: records of presence and absence of the species at sampled localities.
- Abundance: records of the number of individuals (count data) or the weight of the species.
- What is Pseudo-absences? Generating random absence in the study area.

Biological data: response variable

Types of pseudo-absence selection methods:

- Background absence: involves taking randomly pseudo-absence points from the background data (environmental variables).
- Pseudo absence points with limited geographical extent: involves selection of pseudo-absence points within a certain geographic distance from presence points.
- Pseudo-absence points based on environmental variables: involves selection of pseudo-absence points in areas that are environmentally dissimilar from presence points.

Biological data: possible issues

There are a number of potential sources of bias and error that should be carefully considered when analysing species distribution data.

- Identification of species;
- Inaccurate spatial referencing of samples;
- Preferential sampling: collectors tend to sample in easily accessible locations;
- False absences: can occur when a species could not be detected although it was present.

Environmental data: explicative variables

A wide range of environmental input variables have been employed in SDMs. Most common are variables relating to:

- Climate: *e.g.* temperature, precipitation;
- Topography: *e.g.* elevation, aspect;
- Soil type and land cover type.

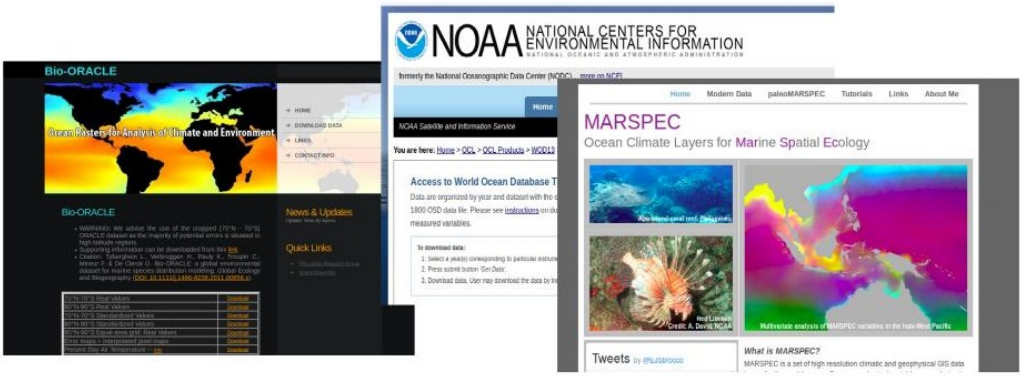
Environmental variables may comprise either:

- Continuous data: that can take any value within a certain range, such as temperature or precipitation;
- Categorical data: that are split into discrete categories, such as land cover type or soil type.

Environmental data: explicative variables

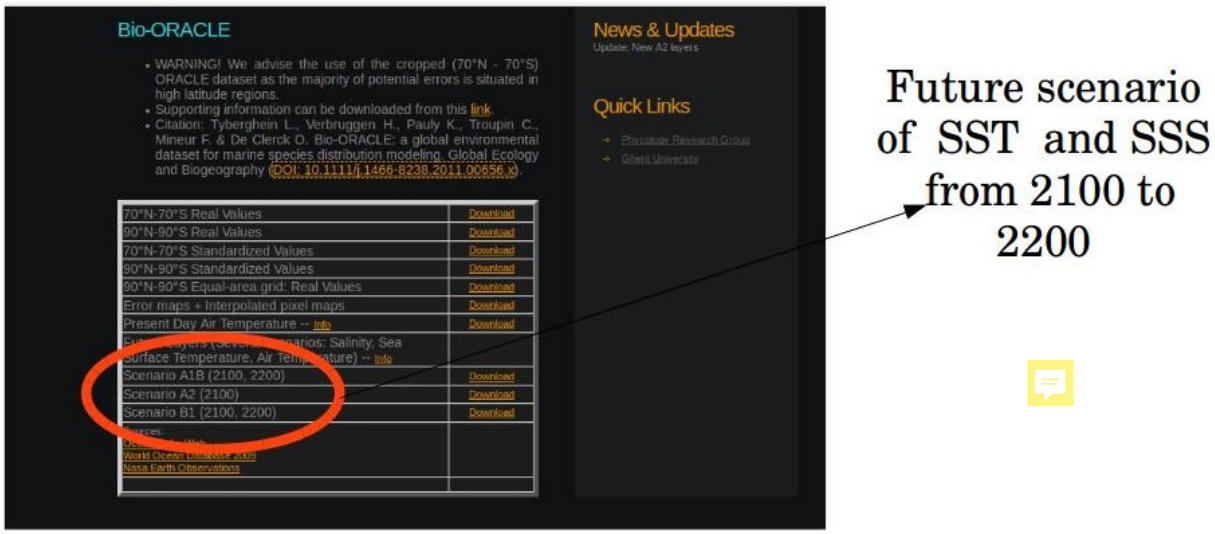
How get environmental data?

- Local measures: sites where a species has been observed, or locations of weather stations;
- Remote sensing (satellite): various atmospheric and land products (will see them later);
- Ocean models: scenarios of future climate change for the globe, reconstructed palaeoclimates, etc;



Environmental data: explicative variables

- Ocean models: scenarios of future climate change for the globe, reconstructed palaeoclimates, etc;



Environmental data: explicative variables

Key point

- Spatial scale: the appropriate data resolution for studying ants is likely to be very different from that for studying elephants.

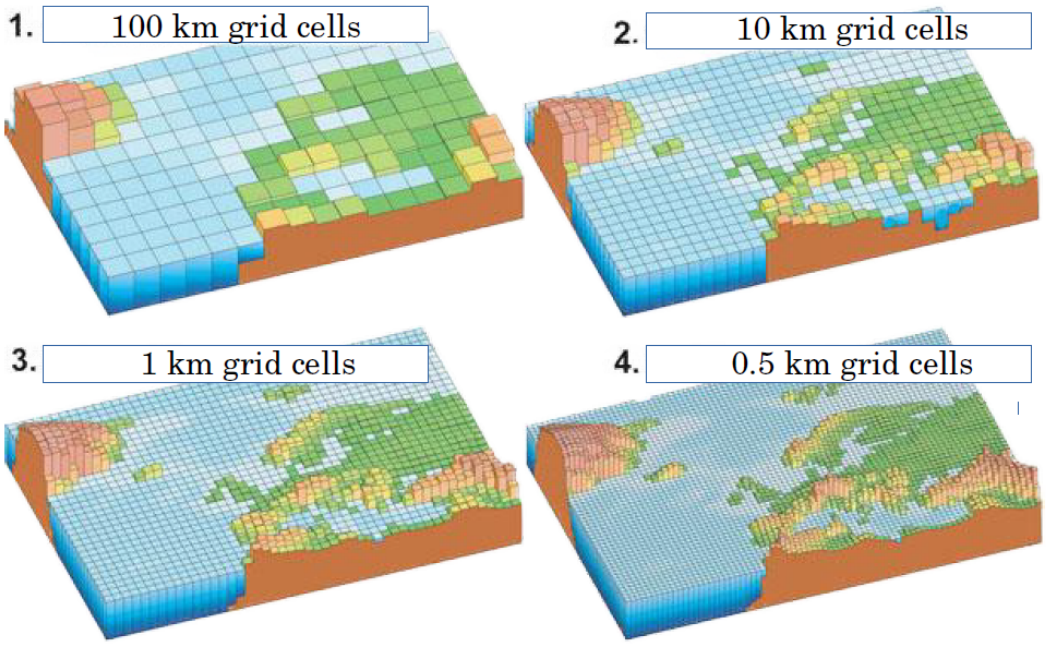
Spatial scale has two components:

- Extent: refers to the size of the region over which the model is run (e.g. New York state or the whole of North America);
- Resolution: refers to the size of grid cells (*e.g.* $1km$ or $10km$).

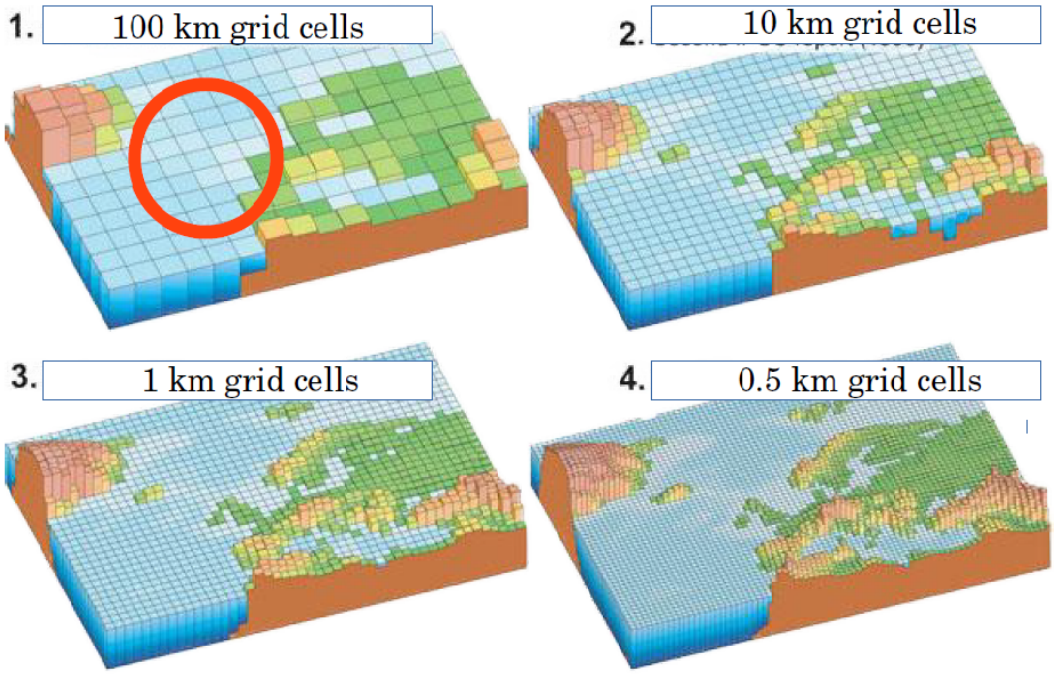
NOTE

- Note that it is common for datasets with large extent to have coarse resolution (e.g. data for North America at $10km$) and datasets with small extent to have fine resolution (e.g. New York state at $1km$). Spatial scale can play an important role in the application of a species' distribution model.

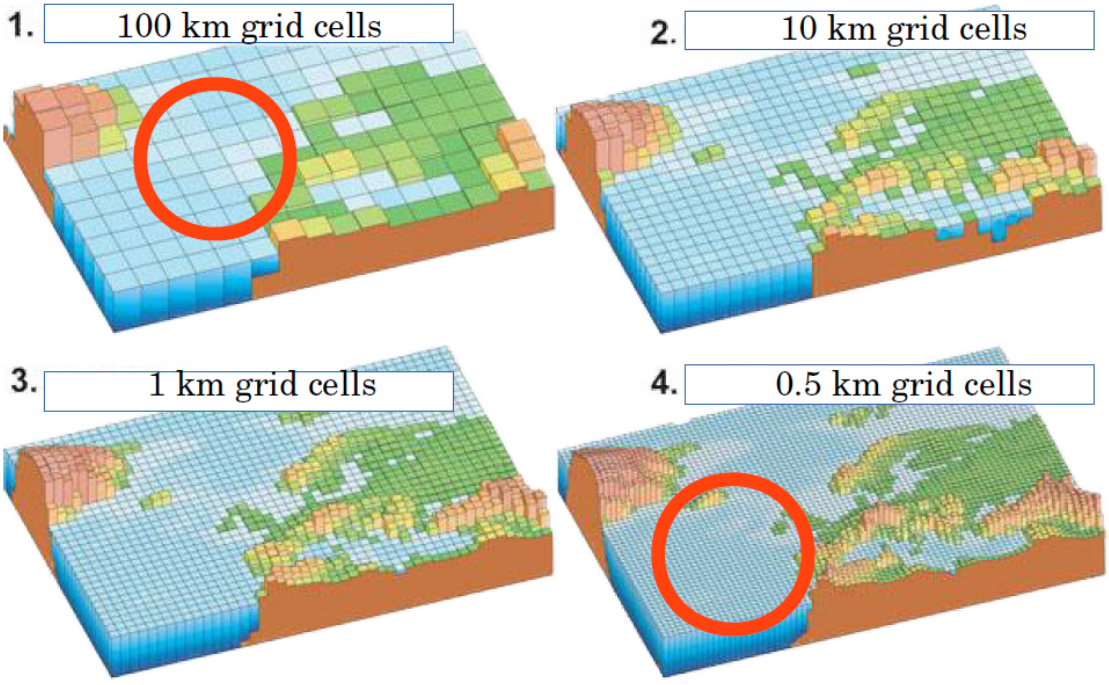
Spatial scale



Spatial scale



Spatial scale



Environmental data: possible issues

There are a number of potential sources of bias and error that should be carefully considered when analysing environmental data.

- Environmental data could be not available for the entire time series, especially for years before 1992.
- Environmental data could be not available with the needed temporal aggregation for the studied species (generally monthly menas but sometimes daily means are needed)
- Environmental data could be not available with the spatial resolution needed for the studied species.
- Missing data (inputation of data).

4 Models type

Modeling algorithms

- A number of alternative modeling algorithms have been applied to classify the probability of species presence (and absence or abundance) as a function of a set of environmental variables.
- All have the same objective: The task is to identify potentially complex linear and non-linear relationships in multi-dimensional environmental space and predict the distribution of a species in unsampled locations or in future (or past) period of time.
- They differ in how they deal with the mentioned issues and consequently in their estimates and predictions.

Modeling algorithms

Some published methods for species' distribution modeling:

- Envelope-model: Use the *BIOCLIM* software (also in R) and presence-only data;
- Gower Metric: Use the *DOMAIN* software and presence-only data;
- Maximum Entropy: Use the *MAXENT* model/software and presence-background data.
- ENFA (Ecological Niche Factor Analysis): Use the *BIOMAPPER* software and presence-background data.

NOTE

- These methods focus on how the environment where the species is known to occur relates to the environment across the rest of the study area (the *background*). An important point is that the occurrence localities are also included as part of the background.

Modeling algorithms

- Genetic algorithm: Use the *GARP* software and presence-pseudo-absence data.
- Artificial Neural Network (ANN): Use the *SPECIES* model/software and presence-absence (or pseudo-absence or abundance) data;
- Regression: generalized linear model (GLM), generalized additive model (GAM), boosted regression trees (BRT), Random Forest (RF), multivariate adaptive regression splines (MARS): presence-absence (or pseudo-absence or abundance): Implemented in R (*dismopackage*).

NOTE

- An important difference between the pseudo-absence approach and the background approach is that pseudo-absence models do not include occurrence localities within the set of pseudo-absences.

How should we choose which method to apply?

- A number of studies have demonstrated that different modelling approaches have the potential to yield substantially different predictions
  - Elith et al. (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography, 29, 129-151.
  - The authors compared 16 modeling methods using 226 species across six regions of the world.

Principles

- What biological data do you have access to?: whether the model uses presence-absence or presence-only data, etc.
- Categorical or continuous data?: *e.g.* Presence/absence or count data cannot be used in LM modeling. Also some models don't allow to use categorical environmental variables.
- What environmental data do you have access to?: *e.g.* linear relationship cannot be fitted in GLM models.
- What is the resolution and extent of this data?: the scale of the model is important. Thuiller et al. showed in 2003 that GAMs are better at performing consistently across scales due to their ability to model complex response curves. However, certain other models might work better at a specific scale.

Principles

- What is the question you want to answer?: only prediction or also estimation. For example, ANN have shown good predictive ability but identifying the relative contribution of each input variable to the prediction is difficult.
- Predictive power: some models may have excellent predictive power but do not enable us to easily understand how the algorithm is operating, black box.
- Uncertainty: whether the model is able to quantify uncertainty.
- Spatial correlation: whether the model is able to take into account spatial correlation.
- Temporal correlation: whether the model is able to take into account temporal correlation.
- Whether the model is able to take into account more complicated issues as detectability, preferential sampling, missing data,zero-inflated data, physical barriers, etc.

5 Model calibration

Model selection and calibration

- Once you have decided on a model type, then you need an methodology to select the best model from a suite of potential predictors.
- Which main effects do we include?
- Which interactions between predictors do we include?
- Model selection tries to simplify this task. This methodology is designed to test the accuracy of prediction of a model.

Model selection and calibration

This is an "unsolved" problem in statistics: there are no magic procedures to get you the best model.

- To "implement" the model selection you need:
  - Check multicollinearity between predictors (e.g. Draftsman's plots, Pearson correlation index, Variance inflation factor (VIF));
  - A criterion to compare models (e.g. Akaike's Information Criterion (AIC), R2, Deviance Information Criterion (DIC));
  - With a limited number of predictors, it is possible to search all possible models;
  - With a large number of predictors you can use multivariate analysis as Principal Component Analysis (PCA).
- Aim: find the best compromise between fit and complexity.

Are our data spatial correlated?

- Everything is related to everything else, but near things are more related than distant things (Tobler, 1970).
- This phenomenon is called spatial autocorrelation by statisticians.
- The property of random variables taking values, at pairs of locations a certain distance apart, that are more similar (positive autocorrelation) or less similar (negative autocorrelation) than expected for randomly associated pairs of random observations (Legendre, 1993).
- In the case of species distributions, spatial autocorrelation occurs mostly because of habitat heterogeneity, or because of biotic processes such as dispersal, conspecific attraction, competition with another species or other complex dynamics (e.g. source-sink).
- When modeling species distributions, the presence of spatial autocorrelation in the residuals is very often an indication that an important covariate was not included in the model (or that the model was misspecified in another way).

6 Predictions

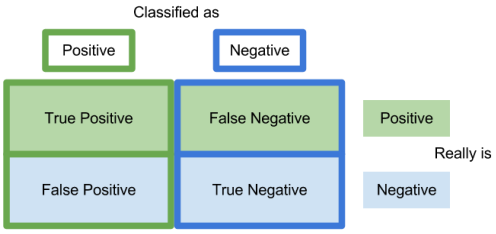
How good is the prediction?

Assessing the accuracy of a models predictions is commonly termed validation, and is a vital step in model development.

- In order to test predictive performance it is necessary to have data against which the model predictions can be compared:
- We can split the data-set in two subsets:
  - Training data: that are used to build the model.
  - Evaluation data: it is fairly common for studies to assess predictive performance by simply testing the ability of the model to predict the evaluation data.
- To repeat for k times (commonly 10)
- The relative proportions of data included in each data set are somewhat arbitrary, and dependent on the total number of locality points available (commonly 70 % for training data and 30 % for evaluation).

How good is the prediction?

- Predictive performance can be summarized in a confusion matrix, which records the frequencies of the four possible types of prediction:



- (a) true positive (TP) (the model predicts that the species is present and test data confirms this to be true);
- (b) false positive (FP) (the model predicts presence but test data show absence);
- (c) false negative (FN) (the model predicts absence but test data show presence);
- (d) true negative (TN) (the model predicts and the test data show absence).

How good is the prediction?

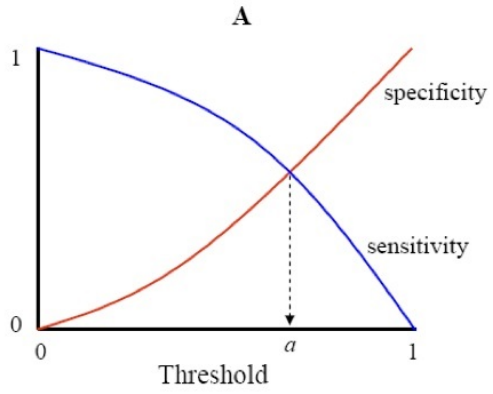
- Sensitivity: proportion of positives which are correctly identified (TP/(TP + FN));
- Specificity as TN/(FP + TN), that is the proportion of negatives which are correctly identified;
  True Skill Statistic (TSS) (Allouche et al. 2006):
  - The TSS is calculated as *sensitivity + specificity - 1*
  - Ranges from -1 to +1, where +1 indicates perfect agreement and values of zero or less indicate a performance no better than random.

NOTE

- These statistics can then be reported as the mean and range from the set of K samples.

Selecting thresholds of occurrence

- Binary predictions are necessary using statistics derived from the confusion matrix. Therefore it is necessary convert probability values in presence/absence by setting a threshold above which the species is predicted to be present.
- Liu et al. (2005) tested 12 methods for setting thresholds using presence/absence data for 2 European plant species. They conclude that the best methods for setting thresholds included maximizing the sum of sensitivity and specificity.



How good is the prediction?

How select an appropriate decision threshold? Depend of question that is being addressed.

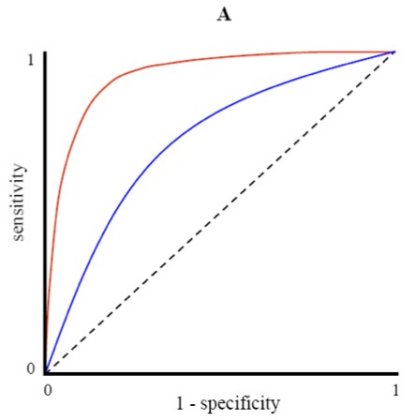
- If the purpose of modeling is to identify areas within which disturbance may impact a species negatively (e.g. as part of an environmental impact assessment), then the threshold may be set low to identify a larger area of potentially suitable habitat.
- In contrast, if the model was intended to identify potential introduction or reintroduction sites for an endangered species or species of recreational value, then it would be appropriate to choose a relatively high threshold.
- Choosing a high threshold reduces the risk of choosing unsuitable sites by identifying those areas with highest suitability.

How good is the prediction?

- When model output is continuous, assessment of predictive performance using statistics derived from the confusion matrix will be sensitive to the method used to select a threshold.
- It is often useful to derive a test statistic that provides a single measure of predictive performance across the full range of possible thresholds, the AUC: the Area Under the Receiver Operating Characteristic Curve.
- The AUC test is derived from the Receiver Operating Characteristic (ROC) Curve. The ROC curve is defined by plotting sensitivity against 1 – specificity across the range of possible thresholds.

How good is the prediction?

The ROC curve thus describes the relationship between the proportion of observed presences correctly predicted (sensitivity) and the proportion of observed absences incorrectly predicted (1- specificity).



Therefore, a model that predicts perfectly will generate an ROC curve that follows the upper curve (red), whilst a model with predictions that are no better than random will generate a ROC curve that follows the dashed line.

How good is the prediction?

- In order to summarize predictive performance across the full range of thresholds we can measure the area under the ROC curve, the AUC, expressed as a proportion of the total area of the square defined by the axes.
- AUC ranges from 0 to 1, with values below 0.6 indicating a performance no better than random, values between 0.7-0.9 considered as useful, and values > 0.9 as excellent.