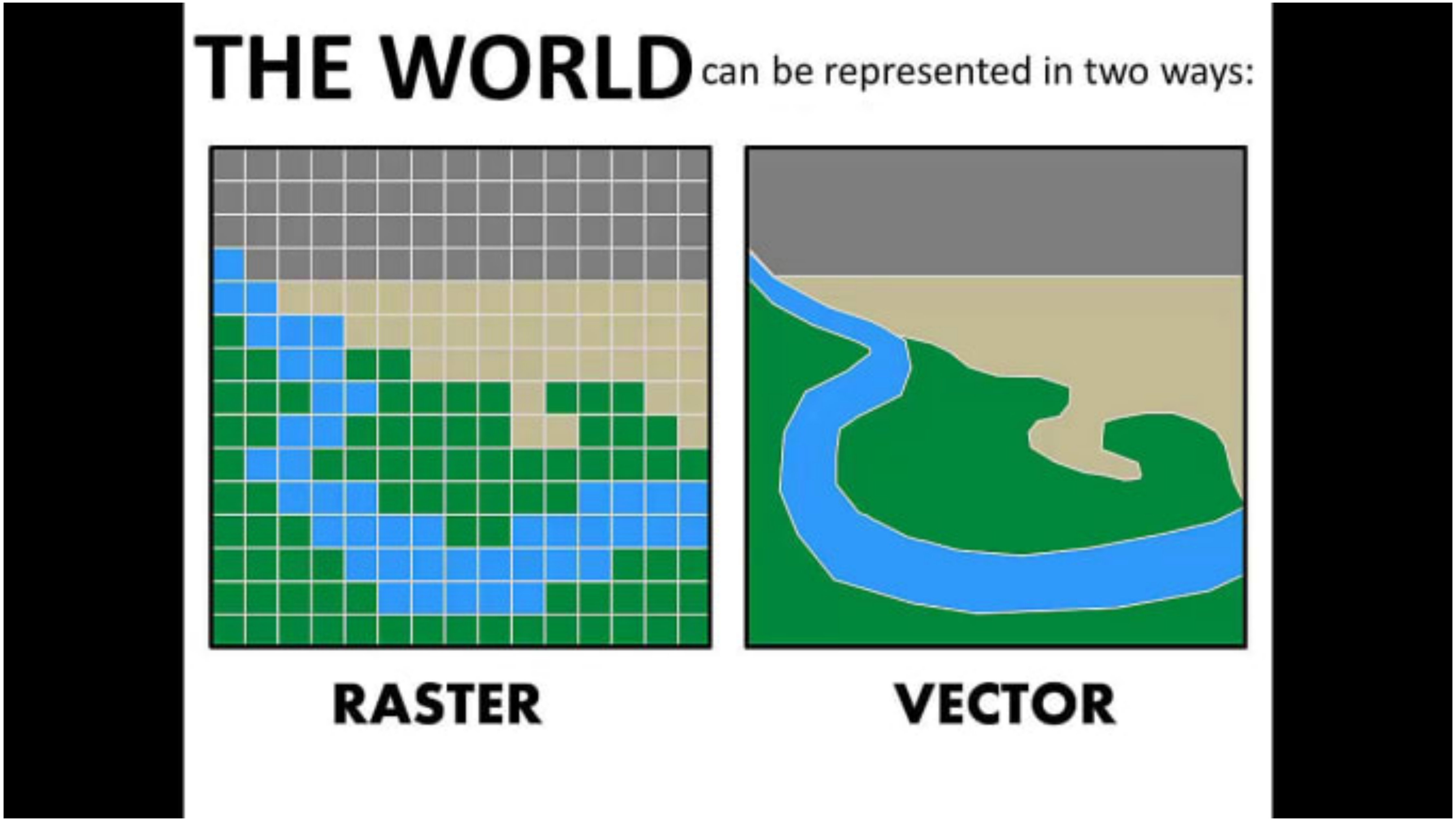
Next class:

April 28nd, 8.30, Teams (?)

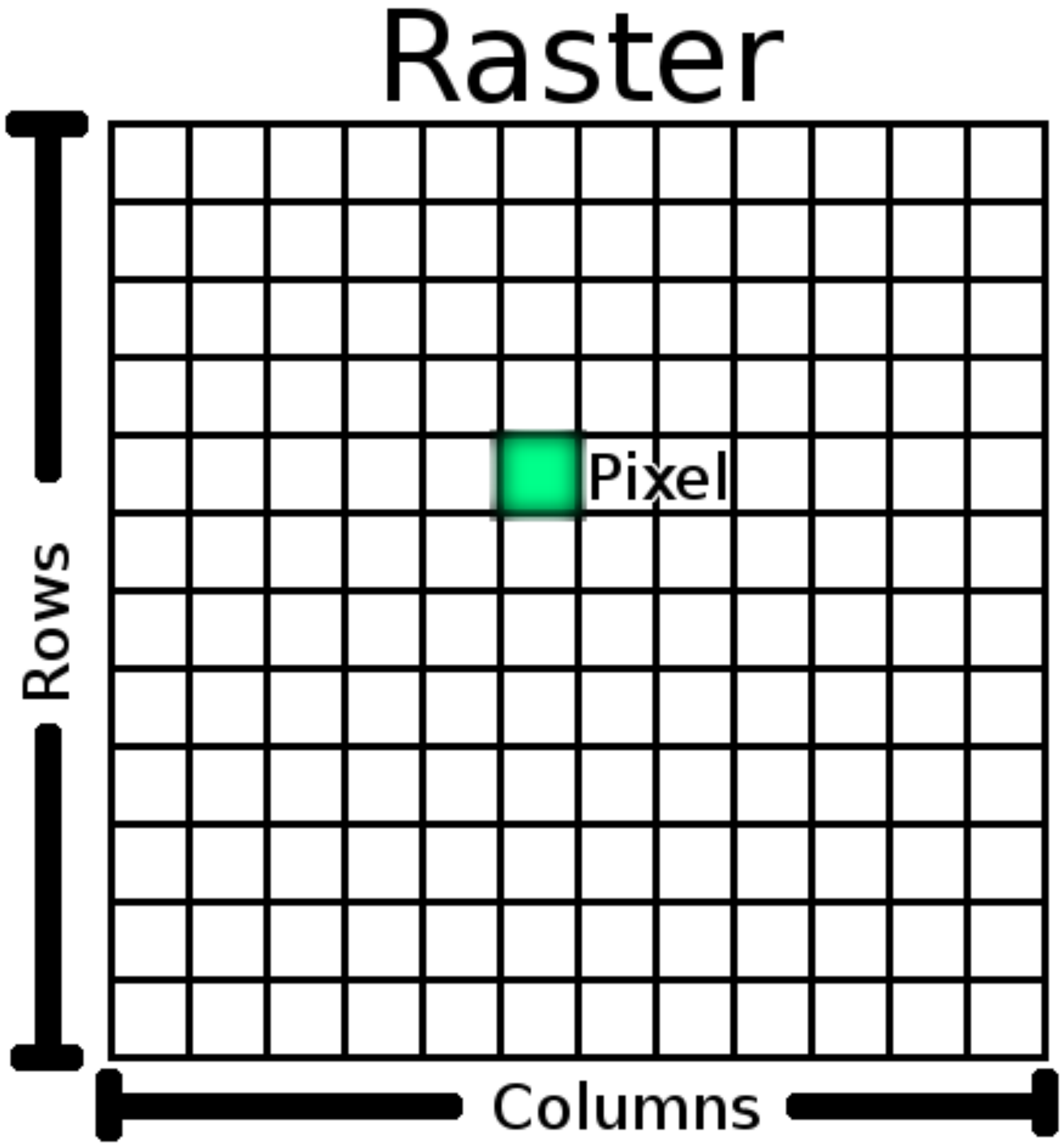
Raster, and vector layers

Raster, and vector layers



Raster

GIS data is commonly stored in a raster format, to encode geographic data as the pixel values. Each pixel in a layer contains an attribute for the variable.



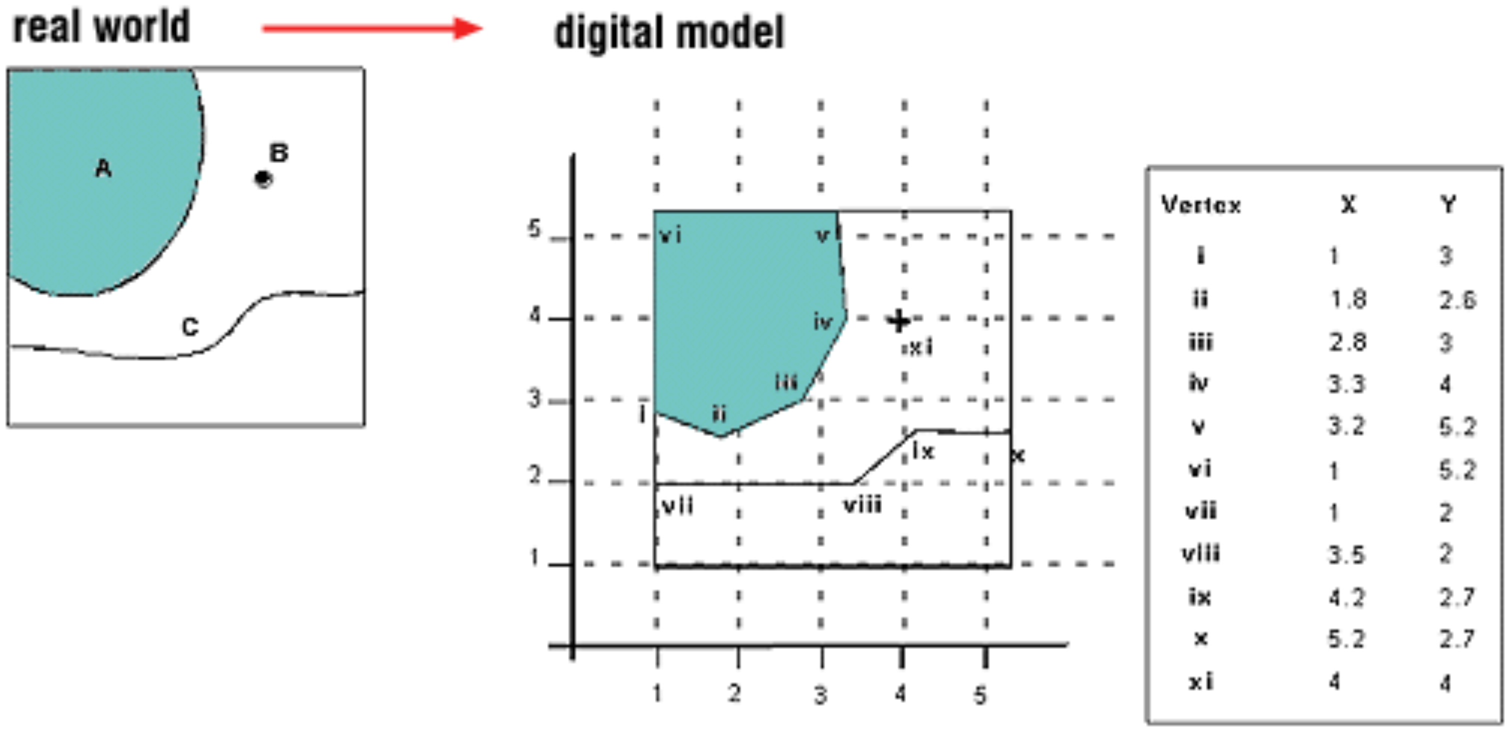
Resolution is clearly dependent on the size of each pixel. The smaller the pixels, the higher the resolution. Plus, raster data cannot scale up to an arbitrary resolution without loss of quality.

However, an high number of pixels requires an high computational effort, and often the grane of the layer is more dependent on the available computational facilities, than on the requirements of the study.

Normally continuous environmental variables are available in the form of raster geotiff files.

Vector

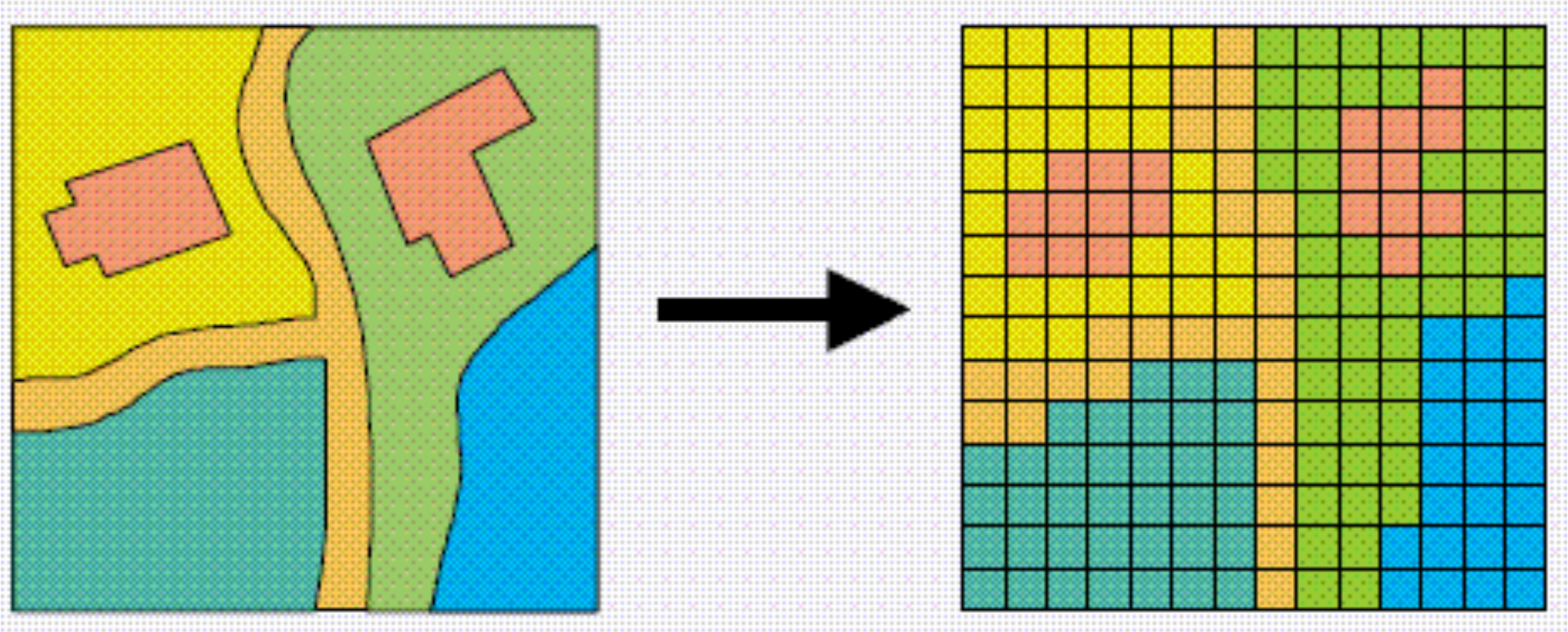
Vector data provide a way to represent real world features within the GIS environment. A feature is anything you can see on the landscape. Each things is a feature when it is represented in a GIS Application. Vector features have attributes, which consist of text or numerical information that describe the features. A vector feature has its shape represented using geometry. The geometry is made up of one or more interconnected vertices. Vector layers can be scaled up without loss of quality.



Vertex	X	Y
i	1	3
ii	1.8	2.6
iii	2.8	3
iv	3.3	4
v	3.2	5.2
vi	1	5.2
vii	1	2
viii	3.5	2
ix	4.2	2.7
x	5.2	2.7
xi	4	4

Raster, and vector layers

Continuous data, such as elevation, or climate data, are not effectively represented in vector form. Plus, spatial analysis and filtering within polygons is impossible. Thus, before being used in models, vector layers are “rasterized”, i.e. converted into raster layers. This process however could lead to decreased data quality, and loss of information. Since a pixel can contain 2 or more vector features, but it can have one attribute only, normally the most “relevant” attribute is chosen. This is however an approximation, which could influence (and bias) the resulting model.



Raster, and vector layers

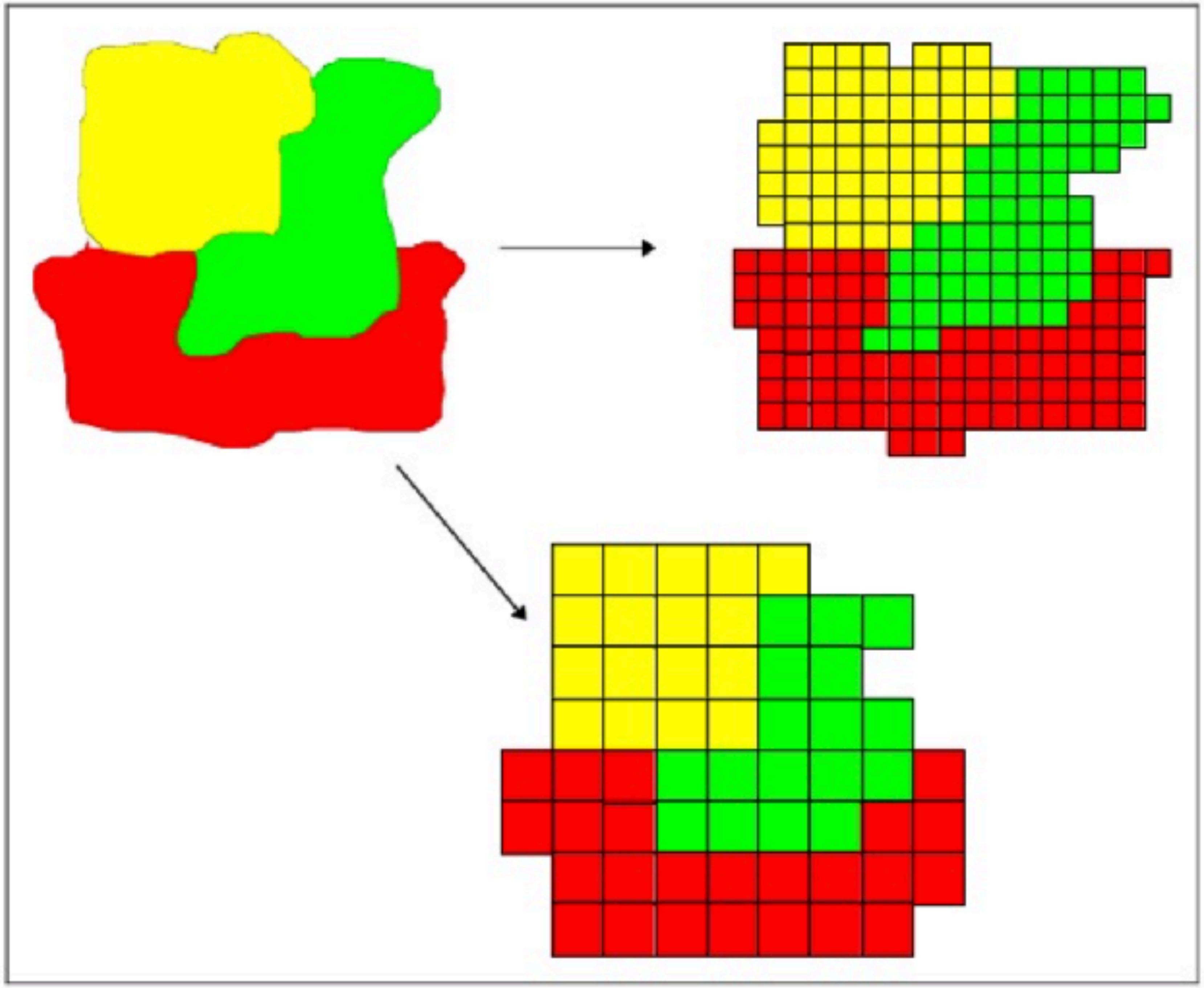


Figure 1. Rasterization of vector-based GIS maps for stand subdivision at different cell resolutions.

Types of environmental predictors

Climate Data

There are several large-scale datasets available, and depending on the nature of the analysis, researchers can choose the one dataset that is appropriate for their analysis.

Worldclim (https://www.worldclim.org) is probably the most widely used global climate dataset for ecological analyses.

It is available in geographic projection at a spatial resolution of 30 arc seconds (~1 km), but also at coarser resolutions (2.5, 5, and 10 arc minutes).

Worldclim maps are based on a large number of climate stations using long-term (1950–2000 in general, but 1961–1990 for most stations) monthly mean climate information for precipitation (47,554 stations), maximum (24,542 stations), and minimum (14,930 stations) temperature.

Due to the globally uneven distribution of climate stations, the mapping uncertainty varies substantially in space.

Climate Data

In addition to basic climate parameters such as monthly mean, minimum, and maximum temperature and precipitation, this dataset provides also a set of other so-called bioclimatic variables (bioclim), which are supposed to be more biological relevant than the original monthly climate layers, from which they are derived.

BIO1 = Annual Mean Temperature
BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3 = Isothermality (BIO2/BIO7) (×100)
BIO4 = Temperature Seasonality (standard deviation ×100)
BIO5 = Max Temperature of Warmest Month
BIO6 = Min Temperature of Coldest Month
BIO7 = Temperature Annual Range (BIO5-BIO6)
BIO8 = Mean Temperature of Wettest Quarter
BIO9 = Mean Temperature of Driest Quarter
BIO10 = Mean Temperature of Warmest Quarter
BIO11 = Mean Temperature of Coldest Quarter
BIO12 = Annual Precipitation
BIO13 = Precipitation of Wettest Month
BIO14 = Precipitation of Driest Month
BIO15 = Precipitation Seasonality (Coefficient of Variation)
BIO16 = Precipitation of Wettest Quarter
BIO17 = Precipitation of Driest Quarter
BIO18 = Precipitation of Warmest Quarter
BIO19 = Precipitation of Coldest Quarter

Climate Data

The datasets are primarily made available for current climate.

However, there are also datasets for historical data (three time slices for the last interglacial, last glacial maximum, and mid-Holocene), as well as for projected future climates for large numbers of global circulation models (GCM) and scenarios originating from CMIP (Coupled Model Inter-comparison Project), and the Assessment Reports of the Intergovernmental Panel on Climate Change(IPCC).

Currently, projections in the future are available for 20 year periods (2021-2040, 2041-2060, 2061-2080, 2081-2100). Their resolution is coarse, at the moment (2.5, 5, and 10 arc minutes). However, a finer resolution (30 arc seconds) will be available soon.
The climate variables available are: tn - monthly average minimum temperature (°C), tx - monthly average maximum temperature (°C), pr - monthly total precipitation (mm), and bc - bioclimatic variables. They were processed for nine global climate models (GCMs): BCC-CSM2-MR, CNRM-CM6-1, CNRM-ESM2-1, CanESM5, GFDL-ESM4, IPSL-CM6A-LR, MIROC-ES2L, MIROC6, MRI-ESM2-0, and for four Shared Socio-economic Pathways (SSPs): 126, 245, 370 and 585.

Land Cover/Land-Use Data

Land cover and land-use (which is often reclassified from land cover datasets) are important and useful for a range of large to regional scale applications.

Both land cover and land-use change are primary threats to biodiversity, thus it can be relevant to include them when assessing, and modeling species patterns in space and time.

One of the early datasets was the International Geosphere Biosphere Program (IGBP) classification of land cover containing 17 classes. It is a classification of data collected on a daily basis between April 1992 and March 1993 by the Advance Very High Resolution Radiometer (AVHRR) scanner, a satellite operated by the National Oceanic and Space Administration (NOAA). The first version of this dataset was released in 1997. This basic 17-item legend is now continued with a product from the MODIS sensor on board the TERRA satellite, available yearly since 2000 at 500 m and 0.05° spatial resolution.

A similar land cover product was generated using AVHRR data and a classification scheme of 12 classes globally, available at 1 and 8 km, as well as at 1° spatial resolution.

Land Cover/Land-Use Data

The European Space Agency (ESA) also provides a global land cover product. This Climate Change Initiative (CCI, http://cci.esa.int) land cover product was generated from time-series of ENVISAT MERIS images (the spatial resolution is 300m for the Full Resolution (FR) data and 1000m for the Reduced Resolution (RR) data).

The dataset covers the majority of the globe (75°N to 56°S, excluding Antarctica). The classification follows the 22-class levels of the Land Cover Classification System (LCCS).

Other land cover products are based on much higher resolution sensor data such as Landsat. One example is the National Land Cover Database (NLCD), a Landsat-based, 30 m resolution land cover database for the USA. NLCD provides land surface information such as thematic class, percent impervious surface, or percent tree canopy cover. NLCD data are free for download from the MRLC website.

Another product often used in Europe is the Coordinated Information on the European Environment (CORINE) land cover (also known as CLC). This combined land cover/land-use dataset is built on a hierarchical legend, which distinguishes >50 classes at finest scales. It comes in two spatial resolutions, 100 m and 250 m. The dataset covers more or less the whole European Union.

Land Cover/Land-Use Data

Often though, we are interested in land-use rather than in land cover.

This is more difficult to map, since it involves interpreting human use of what can objectively be seen from the above, e.g. mapped grassland might be a meadow or a pasture. It might be hard to distinguish between the two, since the difference is the use and not the cover. Furthermore, grasslands can originate from agricultural use, and would naturally revert to forest by means of succession if this use is stopped, while other grasslands are permanent because the conditions are not suitable for forests.

However, when projecting any model into the future, we should be aware that climate is not changing independently of changes in human land-use. Human activities - including land-use change - are among the causes of global change, and therefore future projections of species or biodiversity may suggest that we include the effects of associated land-use change.

Some aspects of land-use are included to a greater or lesser extent in land cover raster. However, at a global scale, land-use data are most often available at very coarse spatial resolutions, which are generally too coarse to be used effectively in SDMs.

Digital Elevation Data

Digital elevation data are a three-dimensional representation of a terrain's surface, very useful for deriving altitude, slope, and aspect.

There are several datasets used extensively at global scale.

Probably one of the most important is maintained by the United States Geological Service (USGS, https://earthexplorer.usgs.gov). It is accessible online, and the world is divided into several tiles in geo tiff format. It is available in geographic projection at a resolution of 30 arc seconds (~1 km resolution at the equator).

It provides a full global coverage, but is not very precise at the finest scales. Nonetheless, it is perfectly sufficient for most global modeling approaches.

For finer scale analyses, most researcher use their own data, yet there exist, at least for some portions of the globe, data at very fine scale, even 30 m.

Global data are usually available in the WGS84 geographic coordinate reference system, while regional applications often have the spatial data transformed to projections that correct for angles or area.

Borders, Political Units, and other vector data

There are several data sources available from which users can access and download free spatial vector datasets.

One source is the Global Administrative Areas database (GADM, https://gadm.org), which contains the spatial data of the world's administrative areas. The data are available in several formats, and can be downloaded either by country or for the whole world.

Another source is Natural Earth (https://www.naturalearthdata.com), which also includes countries, disputed areas, first-order admin (e.g. departments, states).

A source for European-oriented (but also global) datasets is the ESPON database (https://www.espon.eu/espon-database). It provides regional, local, urban, neighborhood, and historical data, etc.

Historical data can be obtained from the CShapes database (http://nils.weidmann.ws/projects/cshapes.html). It contains historical maps of country boundaries and capitals in the post-World War II.

Finally, a database for marine information is Marine Regions (https://www.marineregions.org). It includes marine boundaries, fishing zones, ecological classifications of marine ecosystems, etc.

Let’s switch to R for some examples

Modelling approaches

Modelling approaches

There are different statistical modeling approaches that can be used to predict habitat suitability for species, or other biological entities.

We will give a look to modeling techniques that

(i) are most commonly used, and

(ii) are implemented in R or can be easily called from R.

Selecting the appropriate modeling approaches is ultimately based upon the ecological questions the researcher would like to address, and the availability and accuracy of data to fit the models.

With the development of new powerful statistical techniques, the use of SDMs has increased rapidly. These models are static and probabilistic in nature, since they statistically relate the distributions of populations, species, communities or biodiversity to their environment.

Most of the best-known algorithms have pros and cons, and there is no ultimate algorithm to answer every possible question in ecological modeling. Rather, each algorithm has its own strengths and weaknesses.

Since the emergence of R, most of the algorithms available for analyzing and predicting species distributions can be run jointly and comparatively with the same data and on the same platform. In addition, several packages have been developed to make the best of the different algorithms implemented in R (e.g. packages *biomod2*, or *dismo*).

Recent literature recommends using combinations of data, algorithms, models and predictions, taking advantage of this possibility to use different algorithms on the same platform. Combining models is by no means a new idea. Model selection and multi-model inference have long been discussed in ecology, and it has been proposed that predictions from the same algorithm should be averaged and run over some competitor models or over all subsets of the available variables. The advantage of multi-model inference (**ensemble approach**) is that the inference can be made from more than one single “best” model, by extending the concept of likelihood of the parameters given the model and data, to a concept of the likelihood of the model given the data. Thus, modelers have started to average outputs from different algorithms to get the best out of them, and analyse the uncertainty around the mean.

Modelling approaches

Method(s)	Model/software name	Species data type
Climatic envelope	BIOCLIM	Presence-only
Gower Metric	DOMAIN	Presence-only
Ecological Niche Factor Analysis (ENFA)	BIOMAPPER	Presence/background
Maximum Entropy	MAXENT	Presence/background
Genetic algorithm	GARP	Presence/(pseudo-)absence
Regression: Generalized linear model (GLM) and Generalized additive model (GAM)	GRASP	Presence/(pseudo-)absence
Artificial Neural Network (ANN)	SPECIES	Presence/(pseudo-)absence
Classification and regression trees (CART), GLM, GAM and ANN	BIOMOD	Presence/(pseudo-)absence
Boosted regression trees	BRT (implemented in R)	Presence/(pseudo-)absence
Multivariate adaptive regression splines	MARS (implemented in R)	Presence/(pseudo-)absence

Presence-only approaches

Presence-only approaches (those which use presence data alone) are the simplest and oldest methods available.

They are particular in that they deal with presence-only data with no need to create any background or pseudo-absence data.

They can roughly be separated into two categories:

A) envelopes (e.g. BIOCLIM), and

B) distance-based approaches (e.g. ENFA)

There are two types of envelope approaches, geographic and environmental.

Geographic envelopes are models that focus on the geographic distribution of a species or population. They usually define the “extent of occurrence” of a species as the area contained within the shortest continuous geographic boundary, and are typically the approach used by the IUCN for monitoring changes in species ranges and deriving threat status. Different refinements have been made to remove potential outlier populations, and to provide more conservative estimates of species’ ranges.

Environmental envelopes, on the contrary, are rather more elaborate as they are based on the potential environmental drivers of species distributions. The pioneering approach, **BIOCLIM**, defines the ecological niche of a species as the *n*-dimensional bounding box (i.e. minimal rectilinear envelope) that encloses all the records of the species in the environmental space defined by *n* pre-selected variables. This is similar to Hutchinson’s view of the realized niche, except that it only considers presence data, and does not provide an estimate of habitat suitability. The BIOCLIM-type approach has the advantage of having been the first model to predict the geographic distribution of a species more than 25 years ago, when the use of computer technology in ecology was still in its infancy.

Modelling approaches

The rectilinear envelope is defined in the environmental space by means of the most extreme (minimum and maximum) records of the species along each selected environmental variable.

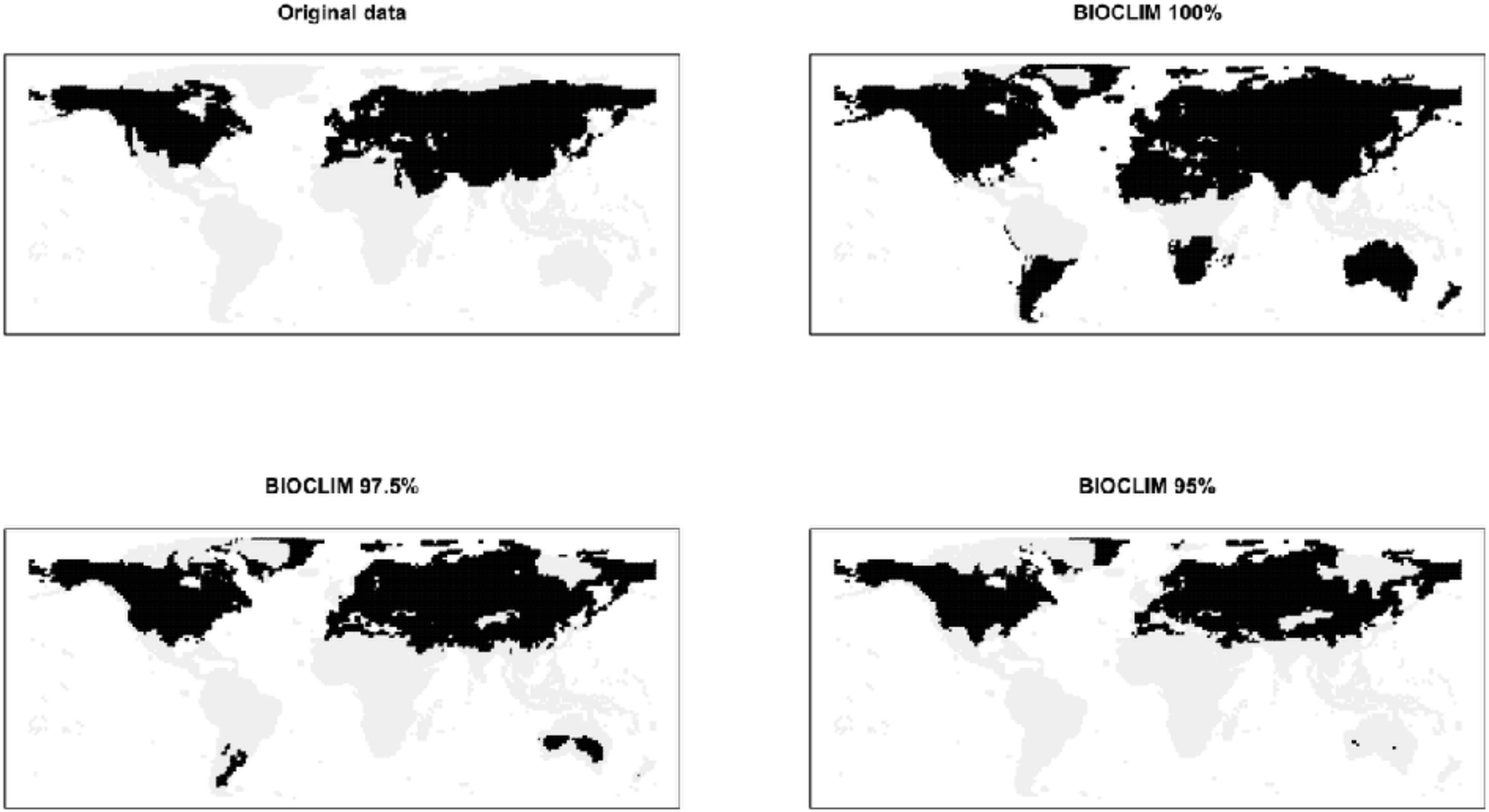
In order to reduce the sensitivity of model predictions to outliers, species records can be sorted along each variable and only the records that lie within a certain percentile range of these environmental gradients (e.g. 5–95%) can be used for model construction. In this way, the model is less sensitive to outliers (i.e. sink populations).

The *biomod2* package proposes a flexible function – species range envelope (SRE) – which essentially reproduces the original BIOCLIM, with the possibility of applying different percentiles.

The *dismo* package also provides more refinement to produce continuous probability maps.

Let’s switch to R,
and make an example

Modelling approaches



Original data

BIOCLIM 100%

BIOCLIM 97.5%

BIOCLIM 95%

We can see that predictions from SRE using 100 percent of the data erroneously predict the southern hemisphere as being suitable for the red fox.

Using the core 95 percent quantile allows for more accurate prediction of the southern hemisphere, but at the cost of underestimating the distribution in Russia.

Generally speaking, such over- and under-predictions highlight the relatively low predictive accuracy of SRE. Indeed, it assumes independent rectilinear bounds, and that all variables are known, and it will cause over-prediction when not enough variables are included and under-prediction with too many variables.

This approach, although quite simple, should thus be used with parsimony and care. However, it does give a quick rough estimate of the habitat suitability of a given species without much effort.

The first major shortcoming of rectilinear envelopes is that they assume the relationship between the presence of a given species and any given variable is binary. In other words, a single presence record under an extreme climatic condition at the edge of a species' range, for example, has the same weight as thousands of presences recorded in the core of the range.

This can be dealt with by adjusting the percentile of the data included. Nonetheless, as we have seen in the example, strongly reducing the percentile can also lead to the exclusion of relevant range information.

The second major problem is that every explanatory variable modeled is apportioned the same weight when constructing the complete species model, and that explanatory variables are treated as independent. This highlights the importance of carefully selecting the variables. In BIOCLIM, even if 100 variables were selected, they would all be used with the same weight, and thus all contribute equally to defining the multidimensional envelope for the given species.

Such a highly constrained model might prove highly accurate in defining the current extent of a given species, but it would expectedly perform relatively poorly when used to project the distribution of the species in space and time.

Distance-based approaches are refined alternatives to simple envelope approaches. Instead of building on rectilinear discrimination, they are usually built on the distance between the environmental centroid of the study area and that defined for the species. This approach is meant to overcome some of the limitations previously discussed such as variable selection and variable importance, which can be used to calculate the axes of the environmental space.

ENFA calculates a measure of habitat suitability based on the analysis of marginality (to what extent a species' mean of the environmental space differs from the global environmental mean across the whole study area, known as background in ENFA), and environmental tolerance, or specialization (to what extent a species' variance in environmental space differs from the global environmental variance).

A threshold of suitability value can then be applied to determine the boundaries of the ecological niche. In this way, ENFA measures the ecological niche that is actually occupied by a given species by comparing its distribution in the ecological space (i.e. a species' distribution) with the distribution of the environment across the whole study area (i.e. the global distribution). As ENFA takes into account background it is not a presence-only method in the strict sense of the word, but rather a presence-background data approach.

With respect to the definition of Hutchinson's niche, a species' marginality indicates the species' niche position (i.e. niche optimum), while the environmental breadth negatively correlates with a species' specialization.

A generalist species, a species that tolerates a large range of environmental conditions, will have a large estimated niche breadth, and *vice versa*. Environmental niche breadth usually strongly correlates to other ecological niche dimensions such as functional traits, or sensitivity to environmental changes.

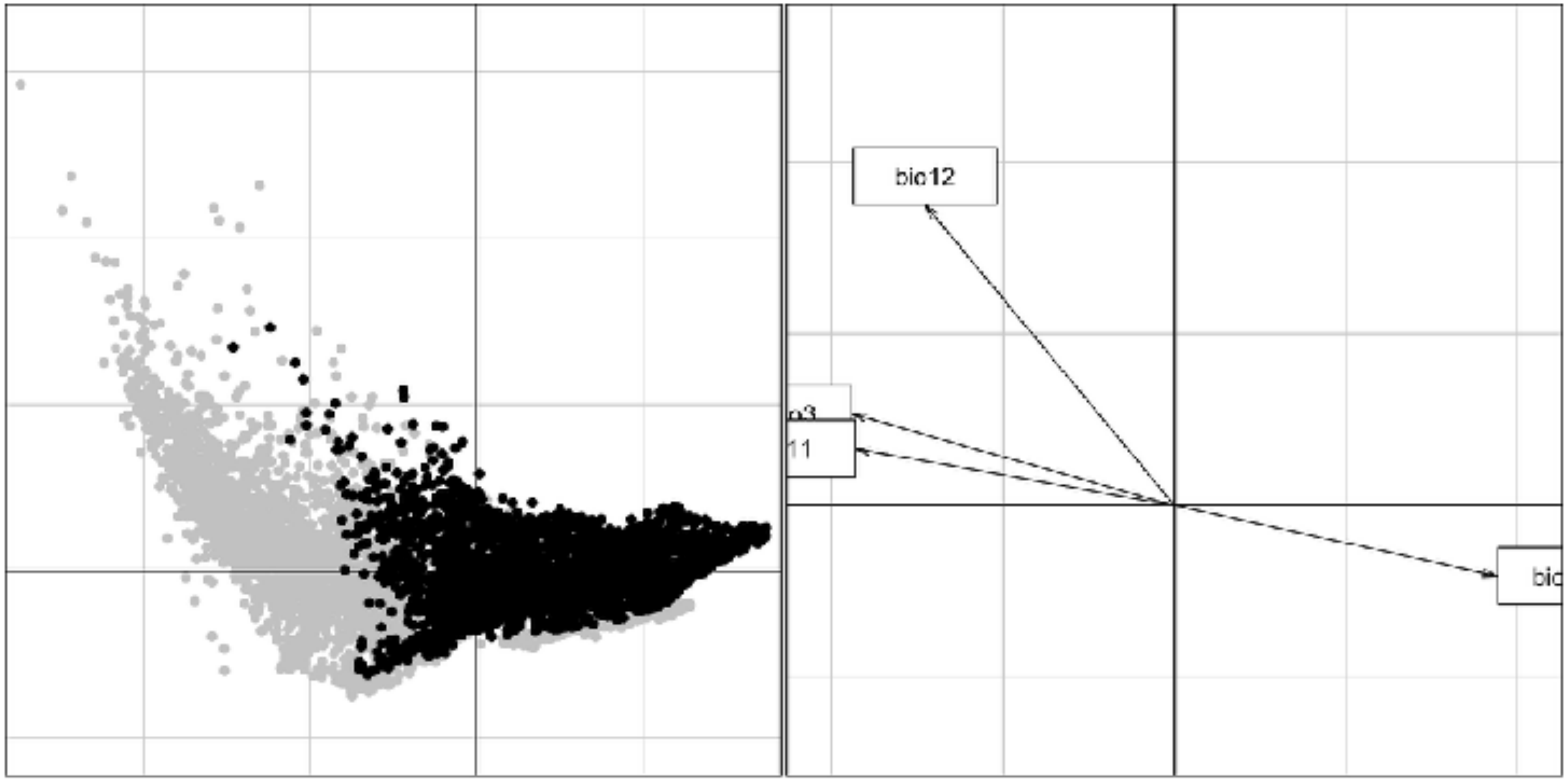
More specifically, ENFA:

(i) transform the predictor variables into a set of uncorrelated factors (as in principal component analysis), and

(ii) construct the axes in a way that accounts for all the marginality of the species on the first axis, and then minimizes species' ecological tolerances along all following axes.

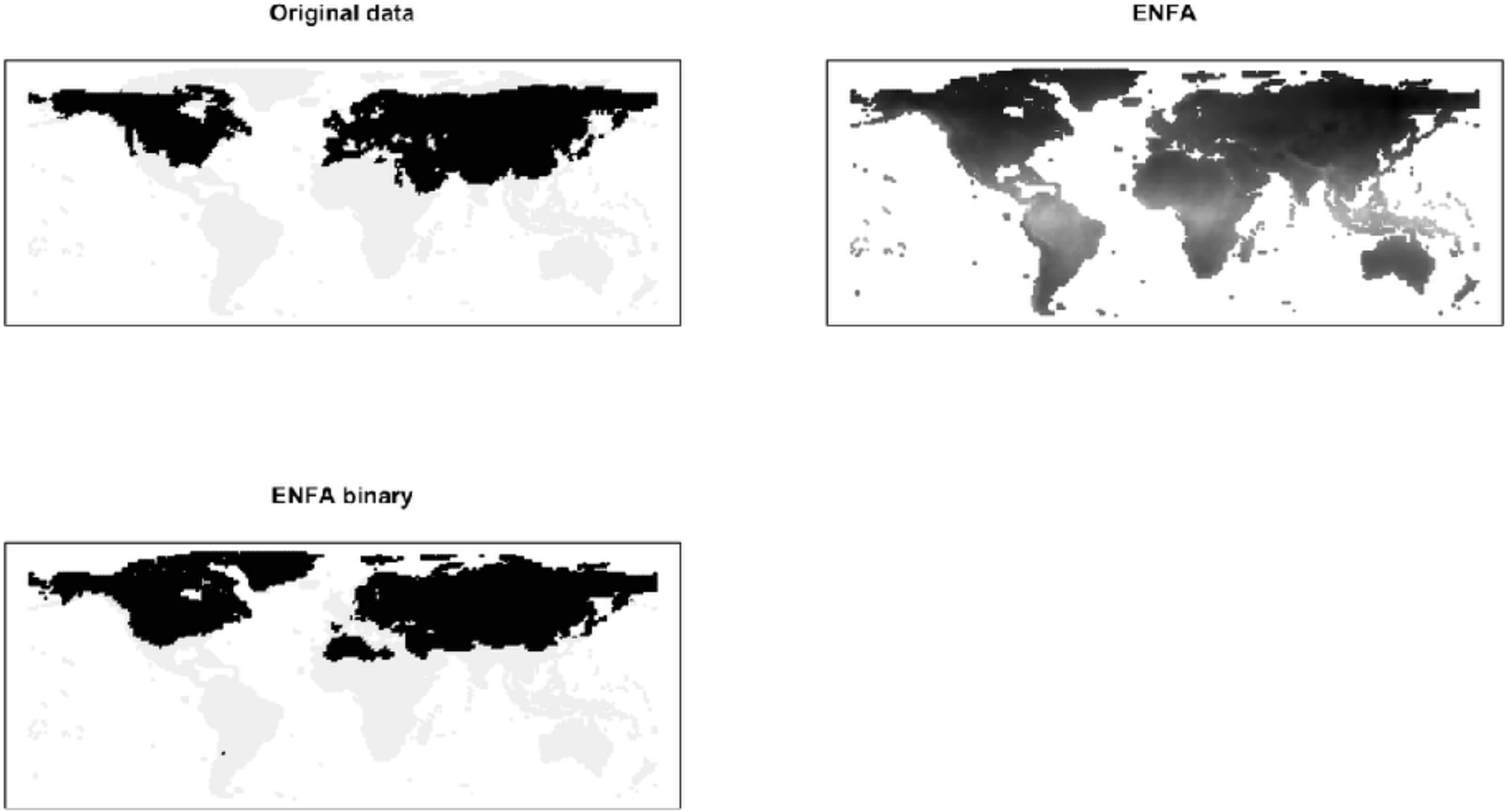
ENFA has been fully implemented in a standalone package called *BIOMAPPER*, but can also be found in the *adehabitatHS* package in R.

Let’s switch to R,
and make an example



The scatterniche plot represents the environment used by the species of interest against the global environment (environmental conditions for the whole world). The major axis of variation is mainly determined by bio3, bio7, and bio11, whereas the second is mostly influenced by bio12.

Modelling approaches



Original data

ENFA

ENFA binary

ENFA generates the environmental suitability of a species by using the Mahalanobis Distance (a measure of the distance between a point and a distribution) between any presence (occurrence), and the centroid of the environmental niche of the species.

ENFA thus produces habitat suitability values as distances, and not as values between 0 and 1.

In order to compare ENFA results to those from BIOCLIM, for instance, we need to transform the ENFA values into binary presence–absence information. Here we used a function from the pRoc package called roc(), which balances the percentage of presence and background data (here assuming they represent non-suitable areas).

BIOCLIM and ENFA enable us to make predictions of potentially suitable habitats based on relatively limited assumptions, and using fairly simple algorithms.

These methods have been extensively compared in isolation, or against methods using either presence and absence or pseudo-absence data. ENFA and BIOCLIM generally have lower predictive accuracy than standard methods using presence and absence data. Anyway, among the two, ENFA generally performs best.

Presence-absence approaches

All the following approaches we will discuss do make use of presence AND ABSENCE data.

Primary biodiversity data are instances of the distribution of an organism. Models use these data to infer the actual distribution.

However, while we know where an organism do occur, we normally do not know where it does **not** occur.

An organism could have not been sampled in an OGU because:

- it is actually absent

- it was not detected

The second case, however, can be split in two subcases:

- the areas was underexplored

- the area was fully explored.

The approaches used to estimate the ecological niche can make use of *pseudo-absences* instead of actual absence data, which are normally difficult to obtain. While actual absences can be used, it is in fact possible to generate *pseudo-assences* from an assessment of the actual distribution of the taxon, or on its fundamental niche.

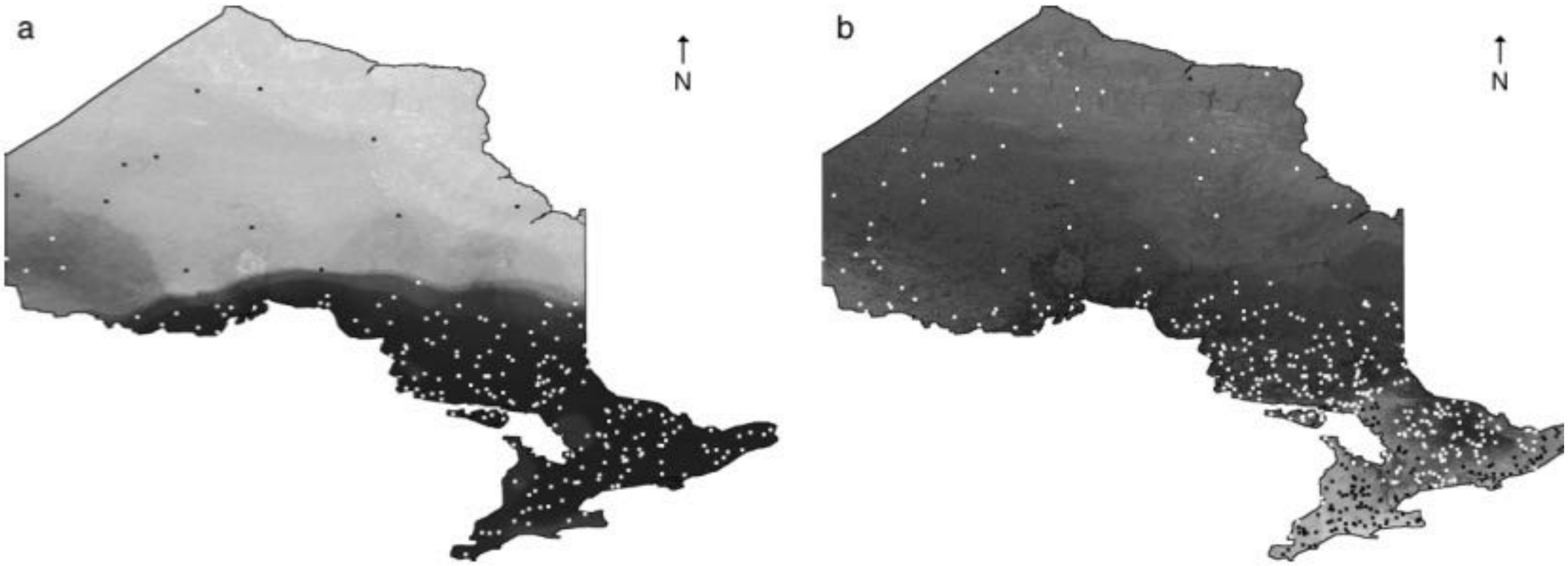
Normally, algorithms using absence data do require a number of absences higher than presence data, better if ten times higher, or more.

The use of pseudo-absences can overcome, or smooth biases due to a bad, or absent sampling strategy. In fact, often sampling is not systematic, hence data do not cover all of a survey area. Furthermore, especially as far as historical data are concerned, systematic sampling was never used as standard approach, and collectors do operate only in "interesting" ecotopes.

So, primary data are often heavily biased, and do not cover several ares. ENMs can overcome such bias by predicting the presence of a taxon in under-sampled (or unexplored) areas.

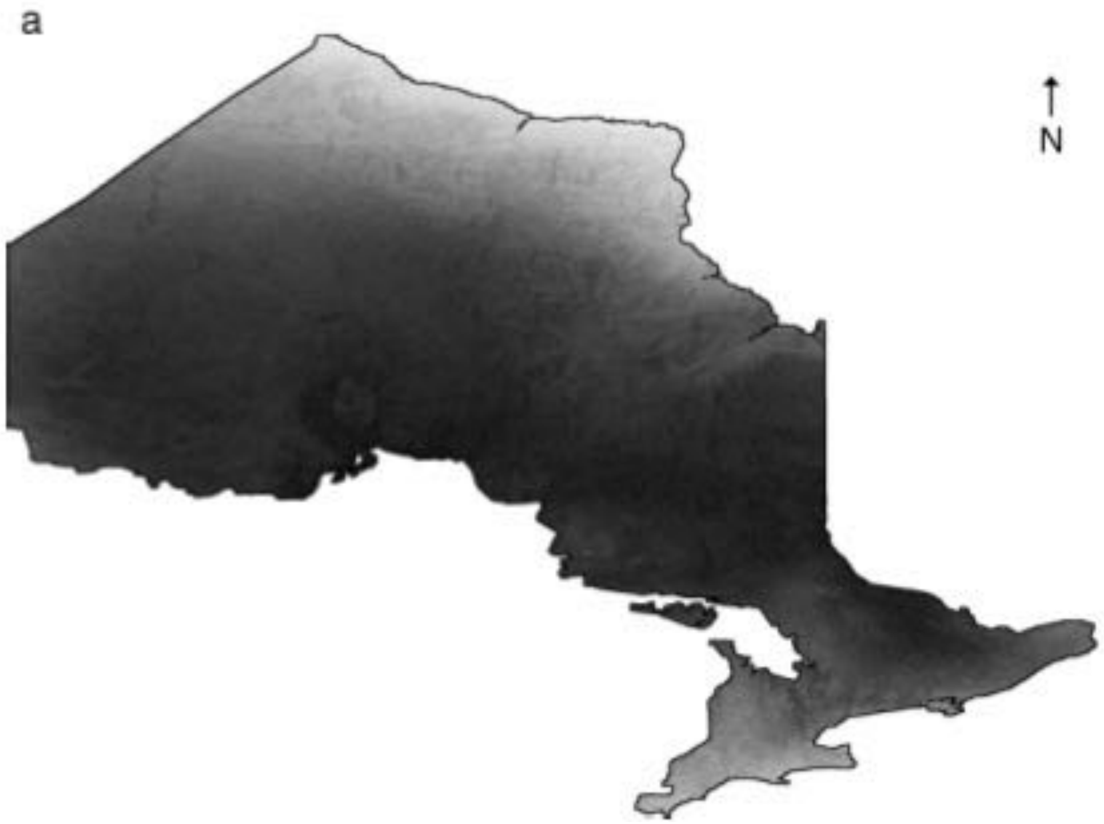
Modelling approaches

Algorithms which use presences only normally overestimate the presence of a taxon in oversampled areas, hence underestimating it in unexplored areas, and lowering the predictive value of the model (a). Model using absences as well, on the contrary, can have far higher predictive value (b).

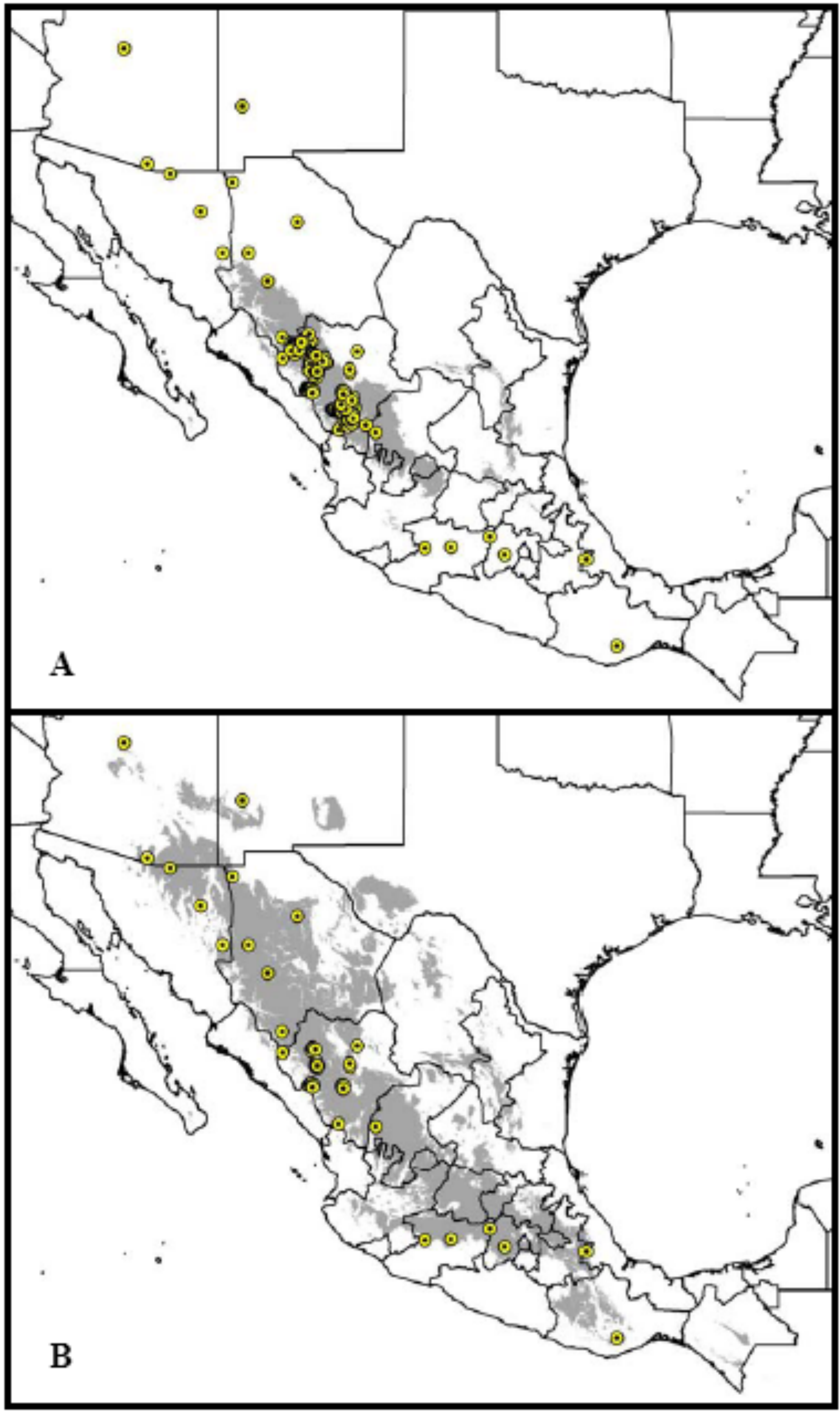


Pseudo-absences

If we compare the two results with the actual distribution of the taxon (see below), we can see how much reliable the prediction of the second model is. Thus, algorithms using absence data can overcome sampling biases far better than others.



Modelling approaches

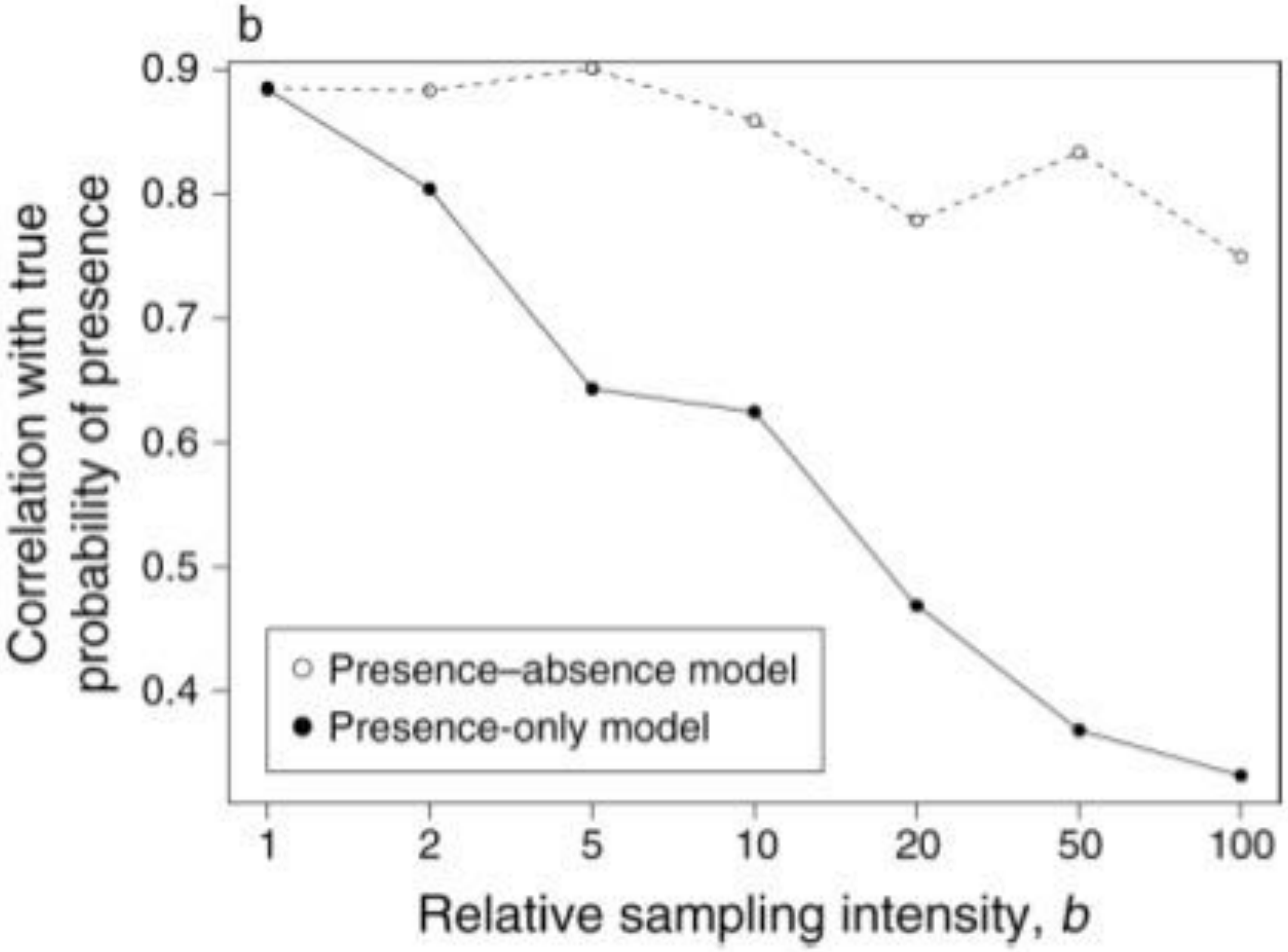


To overcome sampling biases, the approach of removing selectively presence data can be also applied.

In this case, a grid with cells size fitting the ecological diversity of the area of investigation is overlapped to the map of presence data, and multiple presences in grid cells are removed.

Sampling biases are far more important when presence data from natural history specimens are used. Historical collections, until 30-40 years ago, were mostly made by taxonomists, whom collected only "interesting" specimens, and without a sampling strategy, following a preferential sampling approach.

Algorithms using absence data are also useful in overcoming this bias, as shown in the graph.



Pseudo-absences can be produced by using several methods. The simplest is a random generation in all the survey area. In this case, clearly, pseudo-absences could be generates also in OGUs where presence data do occur.

For this reason, several author suggest to remove the pseudo-absences when the fall in OGUs with presences.

To avoid the generation of pseudo-absence in cells where presence data do exist, a two-step modelling approach can be adopted, i.e. an approach which produces a first model by using a presences-only algorithm, and then generate the pseudo-absences outside the predicted distribution.

Pseudo-absences can also be generated by expert assessment, i.e. By using the expertise of a scientist to define OGUs in which the taxon is absent.

Another method uses strictly ecologically related taxa to delimitate areas where a taxon has low probabilities of occurring.

Anyway, number and distribution of the pseudo-absences are major issues for the overall quality of the models.

Whichever way they are generated, as well as their number in comparison with the occurrences, do obviously influence the outcome of the model.

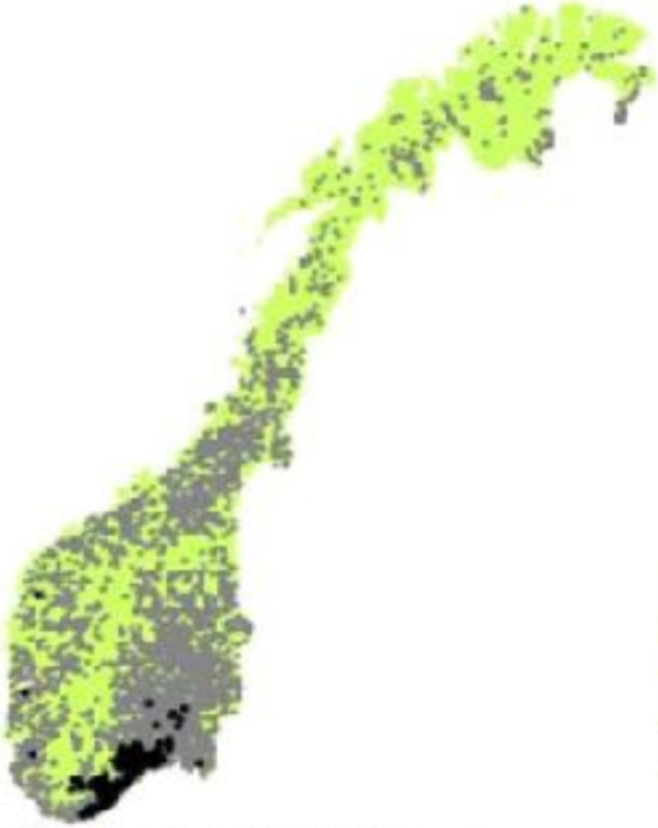
To check the magnitude of the bias, several studies were carried on.

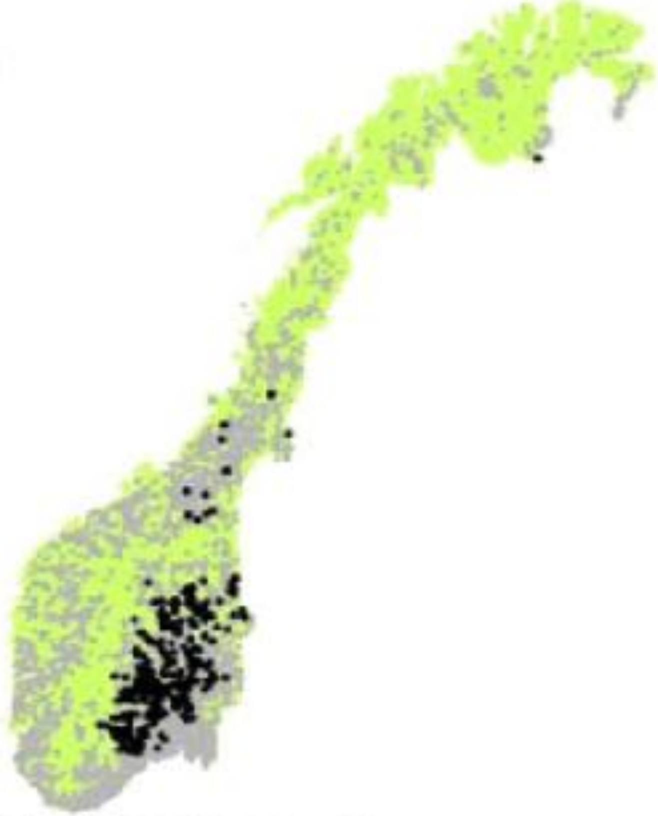
One in particular is of great interest. It was carried out in Norway, and analyzed the distribution of four taxa, two fungi of the group Polyporales (*Fomitopsis rosea* and *Xylobolus frustulatus*), and two insects (*Leptura maculata* and *Anoplodera sexguttata*).

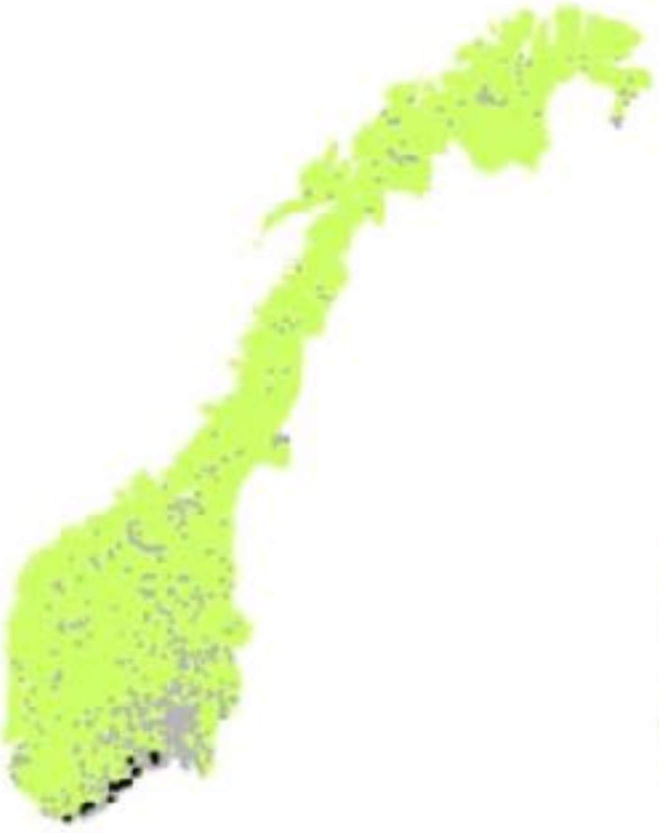
Modelling approaches

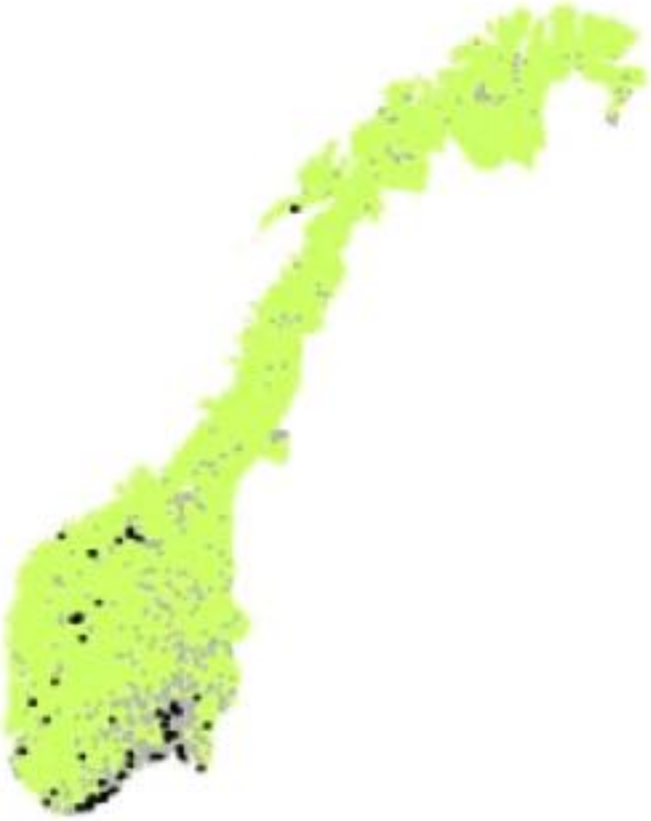
To check the magnitude of the bias, several studies were carried on.

One in particular is of great interest. It was carried out in Norway, and analysed the distribution of four taxa, two fungi of the group Polyporales (*Fomitopsis rosea* and *Xylobolus frustulatus*), and two insects (*Leptura maculata* and *Anoplodera sexguttata*).


Xylobolus frustulatus


Fomitopsis rosea


Anoplodera sexguttata


Leptura maculata

Pseudo-absences

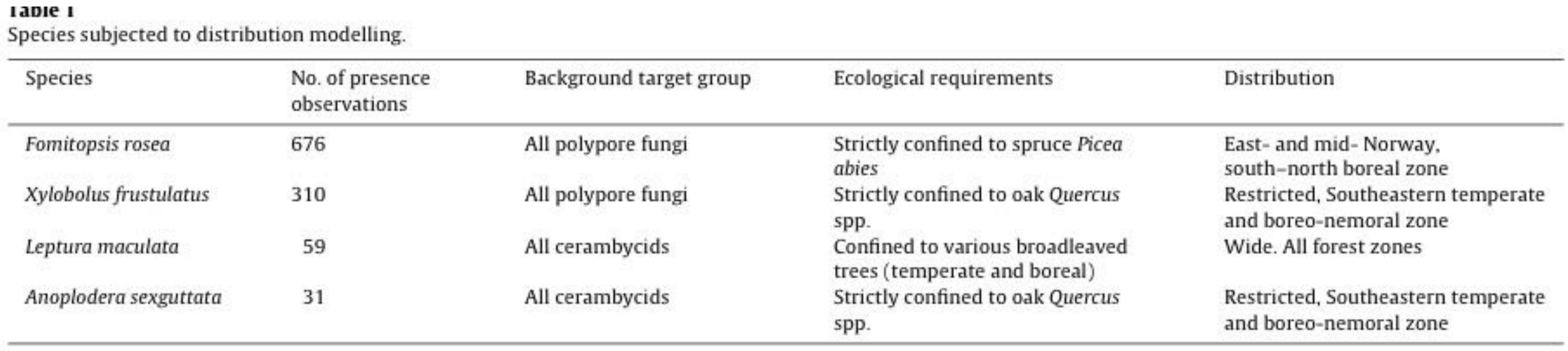


Table 1
Species subjected to distribution modelling.

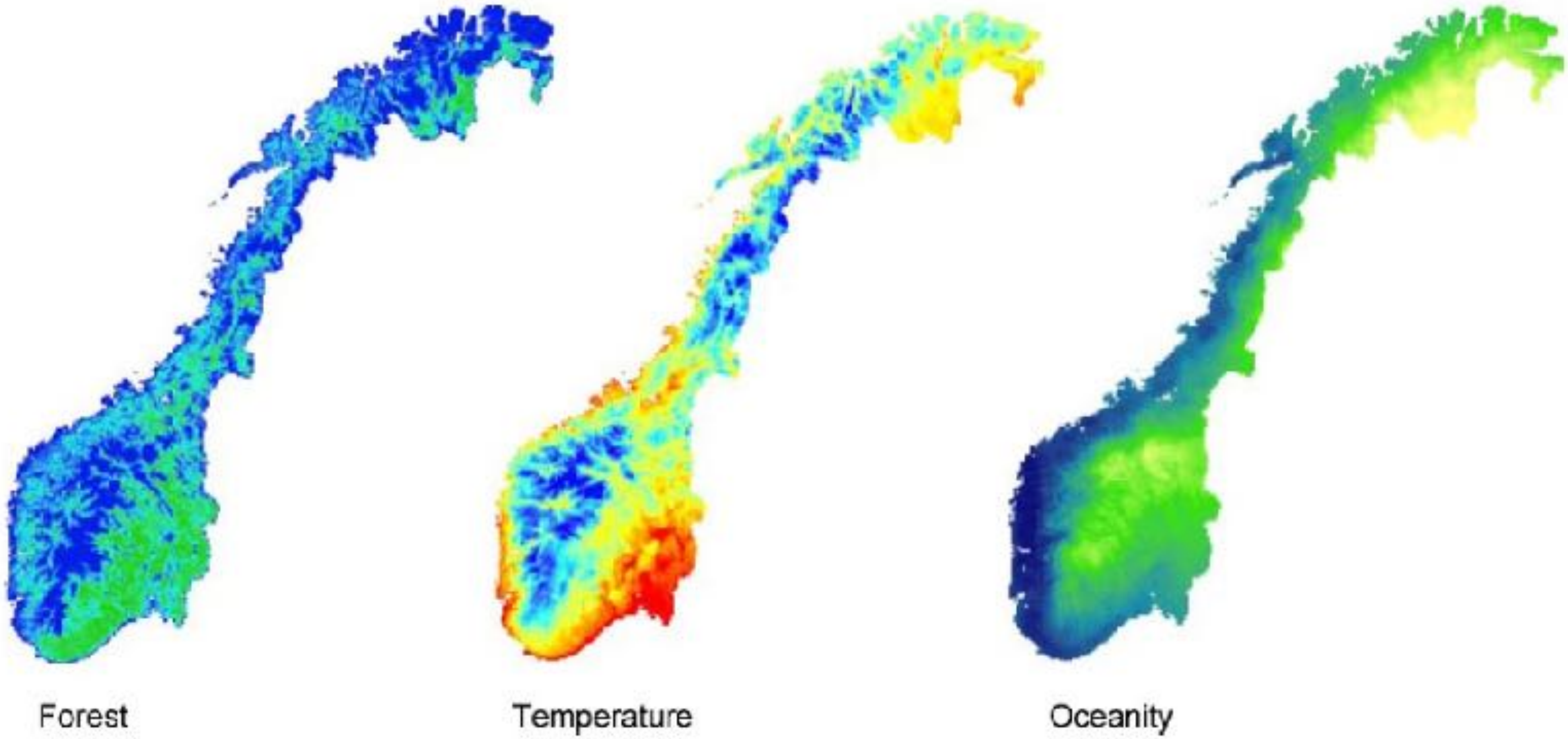
Species	No. of presence observations	Background target group	Ecological requirements	Distribution
Fomitopsis rosea	676	All polypore fungi	Strictly confined to spruce *Picea abies*	East- and mid- Norway, south–north boreal zone
Xylobolus frustulatus	310	All polypore fungi	Strictly confined to oak *Quercus* spp.	Restricted, Southeastern temperate and boreo-nemoral zone
Leptura maculata	59	All cerambycids	Confined to various broadleaved trees (temperate and boreal)	Wide. All forest zones
Anoplodera sexguttata	31	All cerambycids	Strictly confined to oak *Quercus* spp.	Restricted, Southeastern temperate and boreo-nemoral zone

The species do differ in distribution, and in the density of occurrences, ranging from 674 to 31. In the maps, the distribution is depicted on the basis of occurrences only. Data are taken from the GBIF Norge national node.

For all species the sampling intensity is lower in the northern part of the country.

The whole survey area is divided in 14878 OGU (5x5 km each)

Modelling approaches



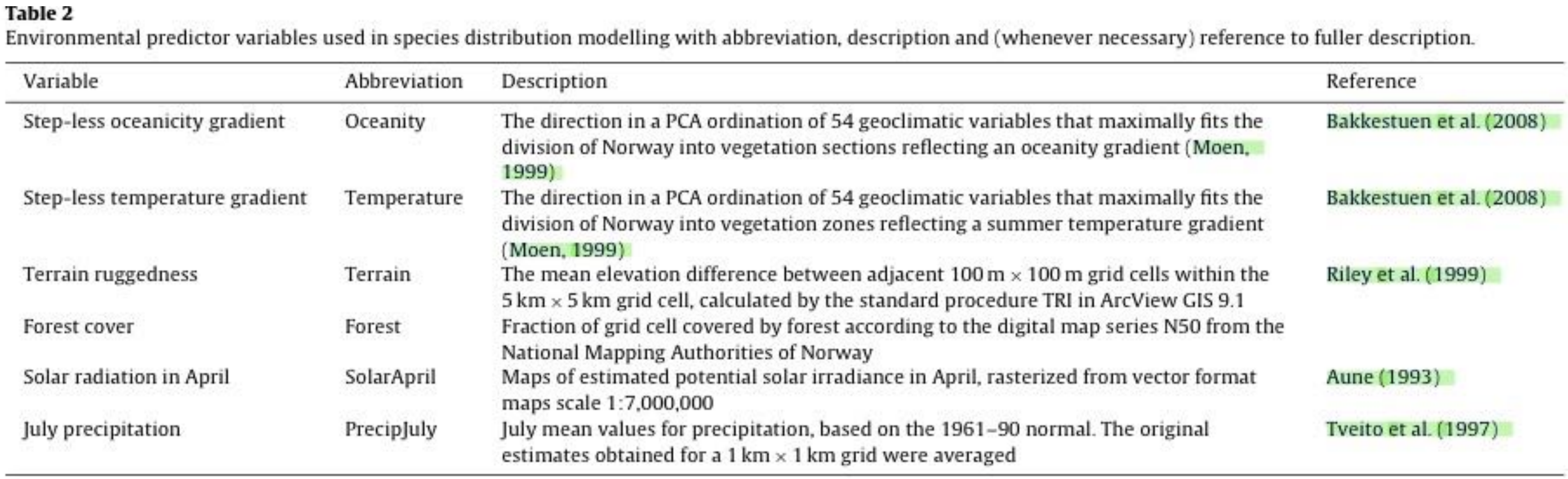


Table 2
Environmental predictor variables used in species distribution modelling with abbreviation, description and (whenever necessary) reference to fuller description.

Variable	Abbreviation	Description	Reference
Step-less oceanicity gradient	Oceanity	The direction in a PCA ordination of 54 geoclimatic variables that maximally fits the division of Norway into vegetation sections reflecting an oceanity gradient (Moen, 1999)	Bakkestuen et al. (2008)
Step-less temperature gradient	Temperature	The direction in a PCA ordination of 54 geoclimatic variables that maximally fits the division of Norway into vegetation zones reflecting a summer temperature gradient (Moen, 1999)	Bakkestuen et al. (2008)
Terrain ruggedness	Terrain	The mean elevation difference between adjacent 100 m × 100 m grid cells within the 5 km × 5 km grid cell, calculated by the standard procedure TRI in ArcView GIS 9.1	Riley et al. (1999)
Forest cover	Forest	Fraction of grid cell covered by forest according to the digital map series N50 from the National Mapping Authorities of Norway	
Solar radiation in April	SolarApril	Maps of estimated potential solar irradiance in April, rasterized from vector format maps scale 1:7,000,000	Aune (1993)
July precipitation	PrecipJuly	July mean values for precipitation, based on the 1961–90 normal. The original estimates obtained for a 1 km × 1 km grid were averaged	Tveito et al. (1997)

Pseudo-absences were generated by three different approaches:

- random

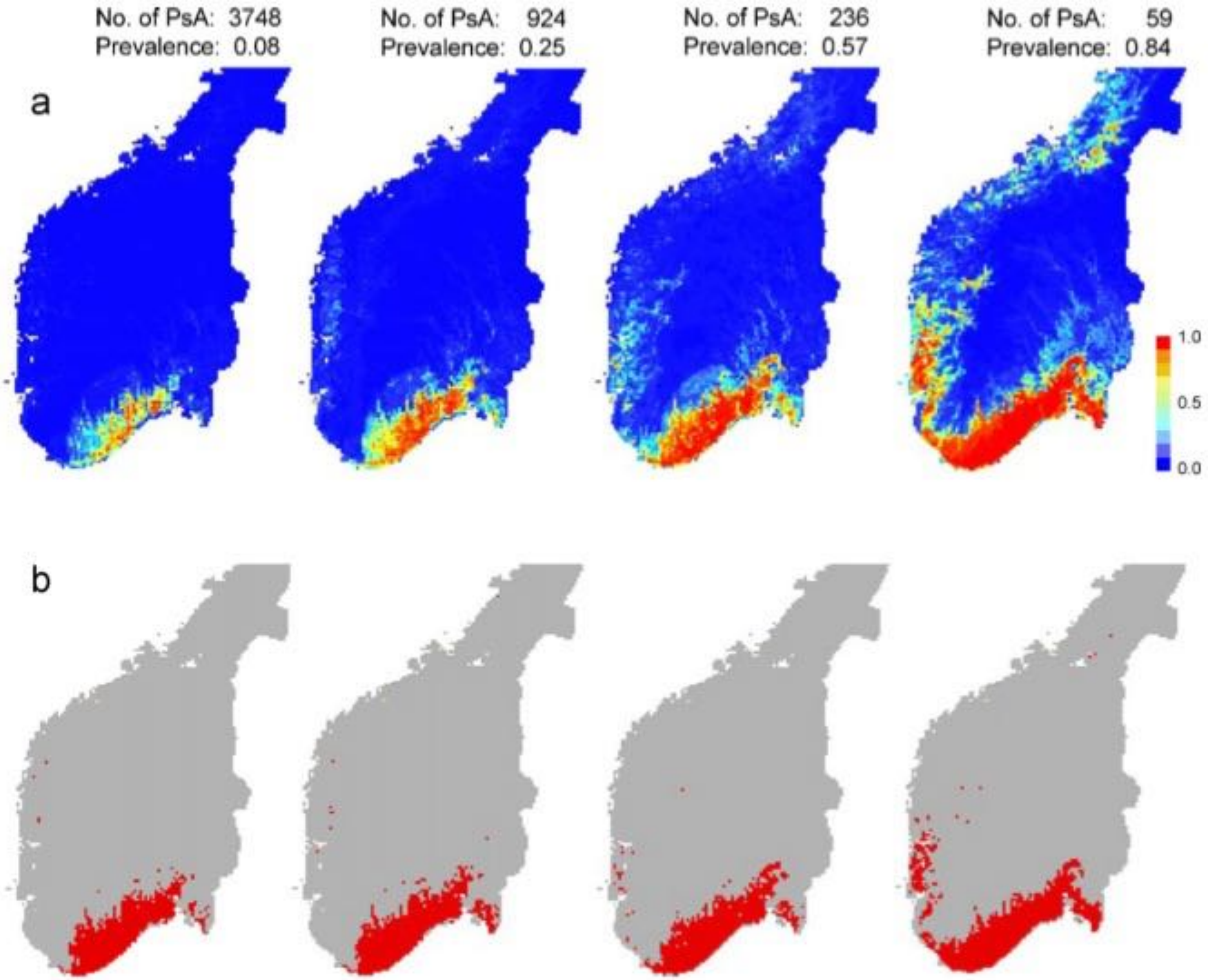
- fixed grid method, by overlapping to the survey area grids of different sizes, and taking the intersections as absence points

- target background method, i.e. by selecting sub-samples of occurrences of other species of the same group. This on the basis of the hypothesis that a species does occur only where other strictly related taxa do occur.

For each method, four numbers of pseudo-absences were generated (64, 256, 1024, 4096), in order to check whether this number influences the outcome.

The first result evidences that the PA generated with the third method produce models which are far from being effective. This because using PA taken from the areas actually occupied by related taxa, entire portions of the survey area are not taken into account. The third method excluded from the model 40% of the survey area, i.e. the whole mountain area, were ecological conditions were the most different from the ideal conditions for the selected taxa, and models do perform better when a certain ecological variability is present.

Modelling approaches



Models obtained for *Xylobolus frustulatus*

The number of PAs influences the outcome of the modelling algorithm, especially when probability maps are produces. In this case, the prevalence (P / PA) strongly influences the outcome. However, when suitability maps are produced, by cutting the presences at a certain threshold value, prevalence seems to be far less important, especially when PAs are ca. 10 times the occurrences.

An high prevalence, anyway, produces models in which (probably) sink populations are overestimated, hence lowering the predictive value of the model.

For this reason, normally it is preferable to produce a number of PAs which is ten to one hundred higher than the occurrences.