

Species distribution modelling with Bayesian statistics in R

A. Márcia Barbosa

<https://modtools.wordpress.com/barbosa>

Course outline

Course outline

Day 1:

Quick refresher on SDM / ENM

Species occurrence data

Predictor variables

[PRACTICAL]

Defining the modeling region (extent and resolution)

Background vs. (pseudo)absence data

[PRACTICAL]

Course outline

Day 2:

Basics of decision trees and tree ensembles

BART model building

BART prediction, uncertainty and credible intervals

[PRACTICAL]

Model evaluation

Block cross-validation

[PRACTICAL]

Course outline

Day 3:

Variable importance

Selection of relevant predictors

[PRACTICAL]

Partial dependence plots (one- and two-dimensional)

Spatial partial dependence plots

Machine-learning maps

[PRACTICAL]

Course outline

- **not** a course on **GIS & mapping**
- **not** a course on **data management**
- **not** a course on **R & programming**

we'll use a bit of each, and I'll **help with anything I can**, but we **can't stray away** from the core subject: using BART for SDM

Course outline

- also **not** a **general SDM** course!

- sampling bias
- data cleaning
- uncertainty in predictor variables

... methodological issues that apply to all SDMs

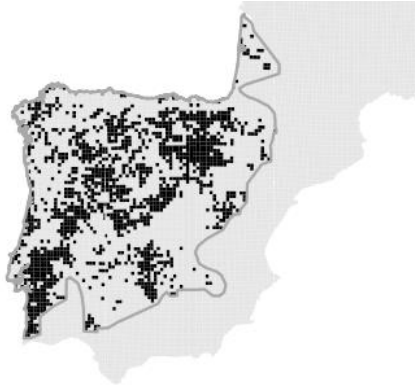
we'll use minimally adequate procedures, but we can't address these issues thoroughly in a short course focused on **using BART for SDM**

SDM / ENM refresher

Species Distribution / Ecological Niche Models

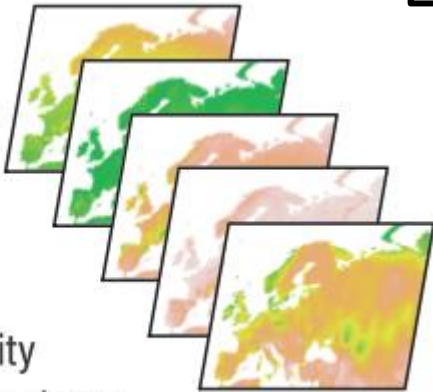
recorded
distribution

+

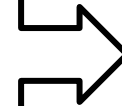


predictor
variables

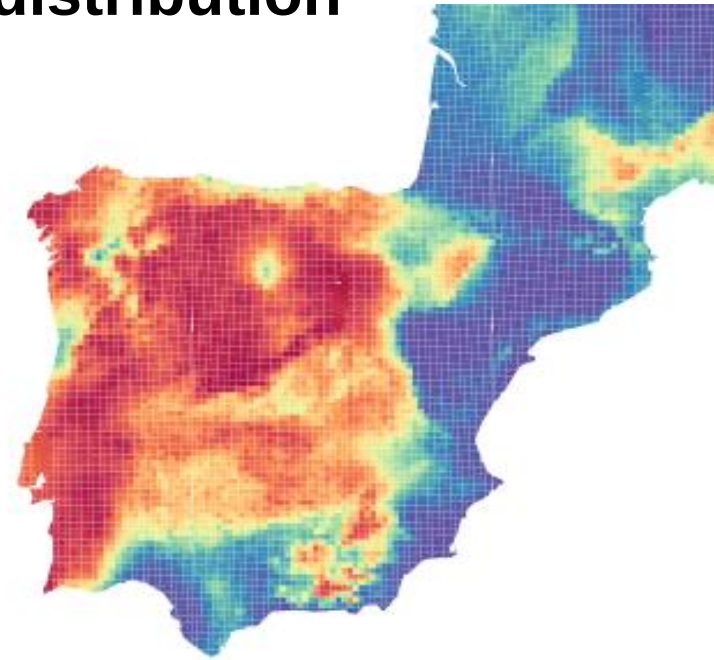
Elevation
Slope
Aspect
Soil properties
Wind speed
Cloud cover
Relative humidity
Max/min temperatures



compute
model



predicted
(potential)
distribution



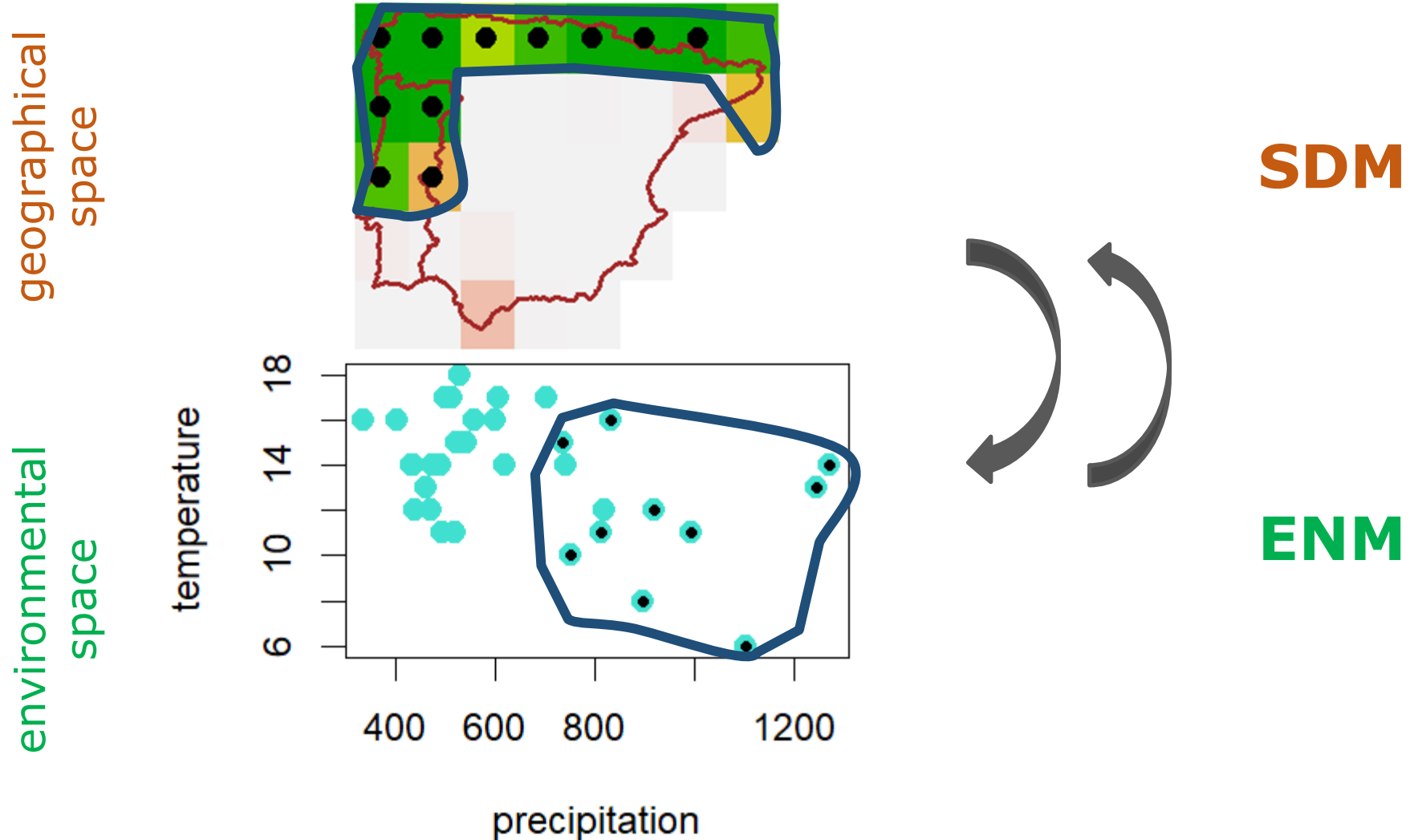
Species Distribution / Ecological Niche Models

- to **assess** the conditions under which a species has been observed
- to **predict** the locations where a species can live

a SDM identifies areas with similar conditions to where the species is (more often) observed

Species Distribution / Ecological Niche Models

from geographical to environmental space and back



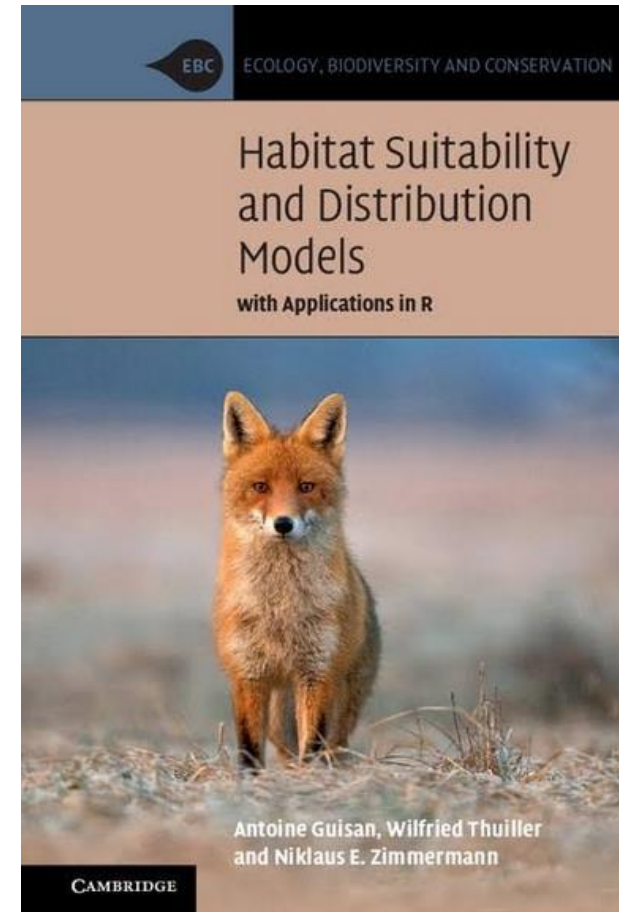
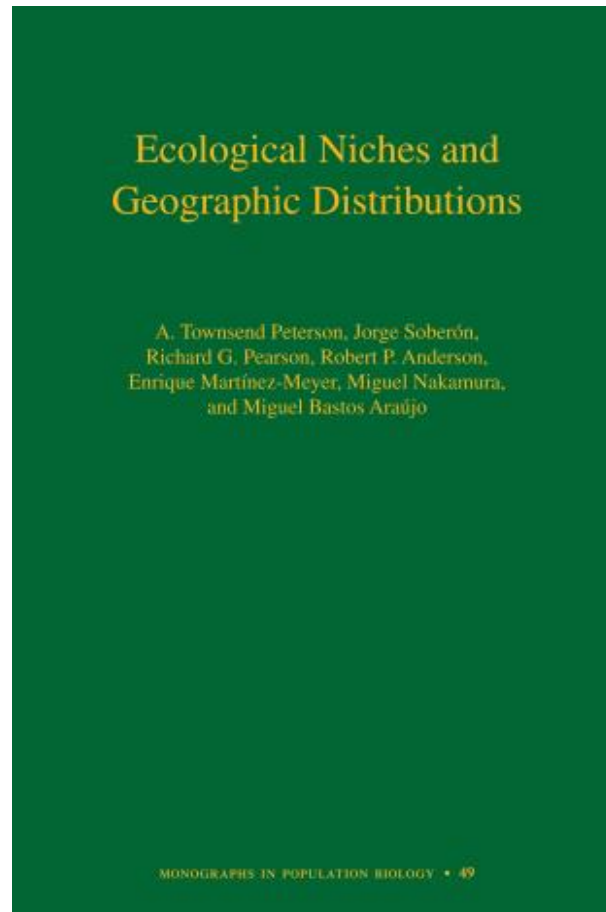
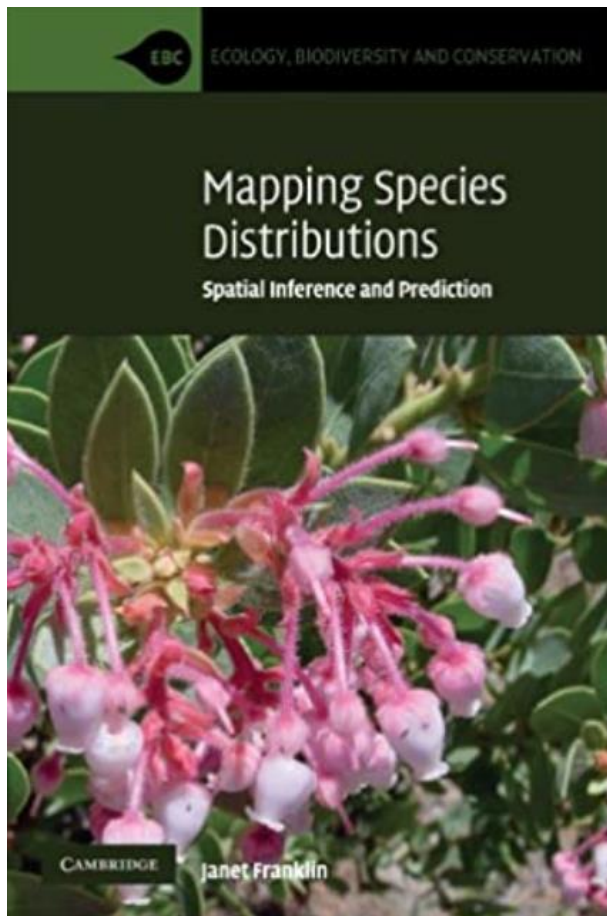
Species Distribution / Ecological Niche Models

fundamental aspects

- quality of species occurrence data
- quality of predictor variables
- extent and resolution of study area

**GARBAGE IN,
GARBAGE OUT!**

Find out more



Find out more

Ecological Modelling 456 (2021) 109671

Contents lists available at [ScienceDirect](#)

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel



Ecological Modelling

Volume 476, February 2023, 110242



A curated list of R packages for ecological niche modelling

Neftalí Sillero ^{a,*}, João Carlos Campos ^a, Salvador Arenas-Castro ^b, A.Márcia Barbosa ^a



Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling

Neftalí Sillero ^{a,*}, Salvador Arenas-Castro ^{a,b}, Urtzi Enriquez-Urzelai ^{c,d}, Cândida Gomes Vale ^e, Diana Sousa-Guedes ^a, Fernando Martínez-Freiría ^c, Raimundo Real ^f, A.Márcia Barbosa ^a

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2021, VOL. 35, NO. 2, 213–226
<https://doi.org/10.1080/13658816.2020.1798968>



Taylor & Francis
Taylor & Francis Group



FORESIGHT PAPER

Common mistakes in ecological niche models

Neftalí Sillero ^{id} and A. Márcia Barbosa

ECOGRAPHY

Open Access

Review | Open Access |

Top ten hazards to avoid when modeling species distributions: a didactic guide of assumptions, problems, and recommendations

Mariano Soley-Guardia, Diego F. Alvarado-Serrano, Robert P. Anderson

First published: 31 January 2024 | <https://doi.org/10.1111/ecog.06852>

Software note | Open Access |

Achieving higher standards in species distribution modeling by leveraging the diversity of available software

Jamie M. Kass , Adam B. Smith, Dan L. Warren, Sergio Vignali, Sylvain Schmitt, Matthew E. Aiello-Lammens, Eduardo Arlé, Ana Márcia Barbosa, Olivier Broennimann, Marlon E. Cobos, Maya Guéguen, Antoine Guisan, Cory Merow, Babak Naimi, Michael P. Nobis, Ian Ondo, Luis Osorio-Olvera, Hannah L. Owens, Gonzalo E. Pinilla-Buitrago, Andrea Sánchez-Tapia, Wilfried Thuiller, Roozbeh Valavi, Santiago José Elías Velazco, Alexander Zizka, Damaris Zurell

First published: 19 November 2024 | <https://doi.org/10.1111/ecog.07346>

Species occurrence: data types and sources

Data setup

example of a data frame that can be used for computing a model

presence	x	y	cells	bio_1	bio_2	bio_3	bio_4
0	26.125	-25.125	5954	19.64711	16.02806	55.41713	482.3851
0	26.375	-25.125	5955	20.27168	16.56423	55.74219	487.0897
0	26.625	-25.125	5956	20.01786	16.05557	55.92709	470.6123
1	26.875	-25.125	5957	19.65874	15.92198	56.15314	465.0204
1	27.125	-25.125	5958	19.52998	15.71729	55.63602	465.7927
0	27.375	-25.125	5959	19.54671	15.34325	54.76509	470.6145
0	27.625	-25.125	5960	19.77601	15.25366	54.40098	475.5373

*species occurrence data:
1 = presence,
0 = no recorded presence*

*values of the
predictor variables*

Occurrence data types

- point occurrences

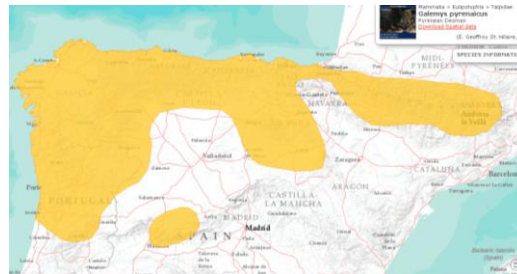


most
common

- gridded data

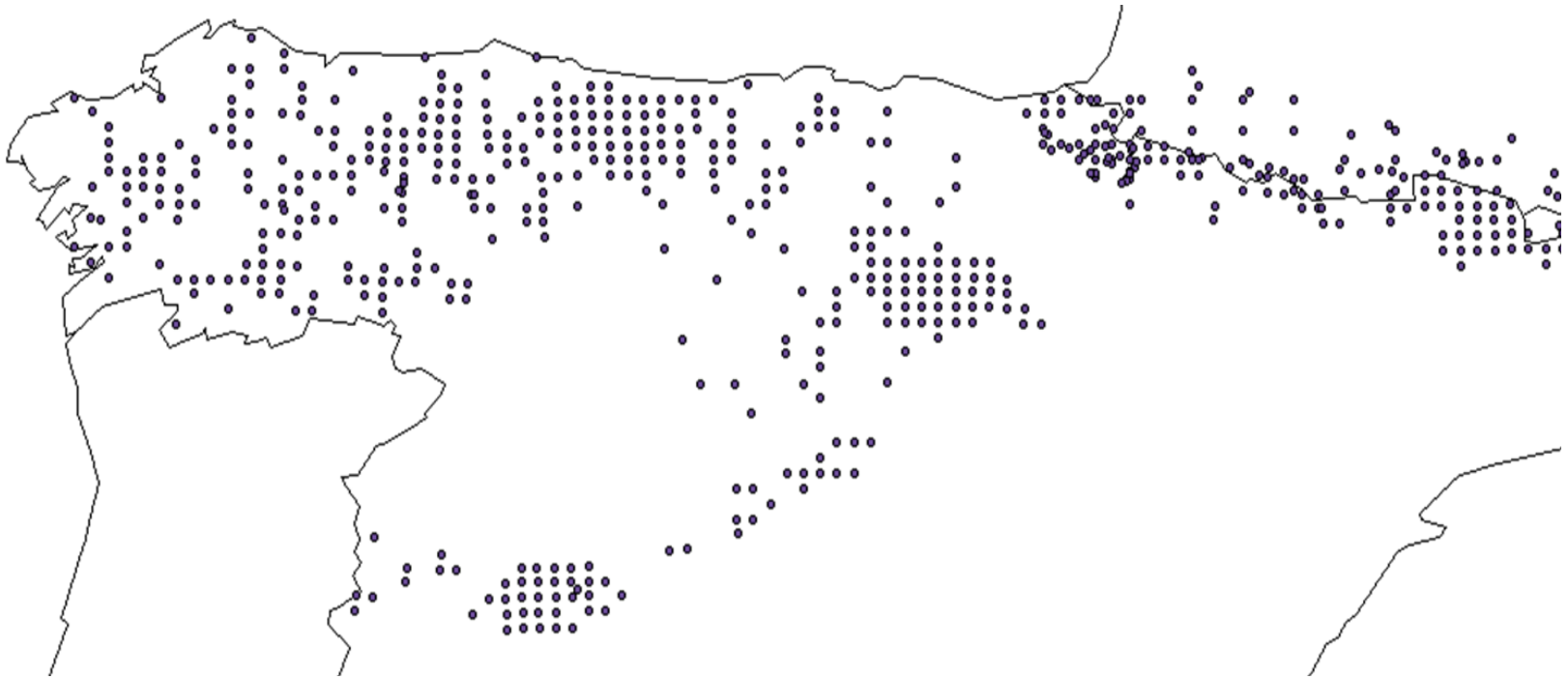


- range map



- others? what's your data type?

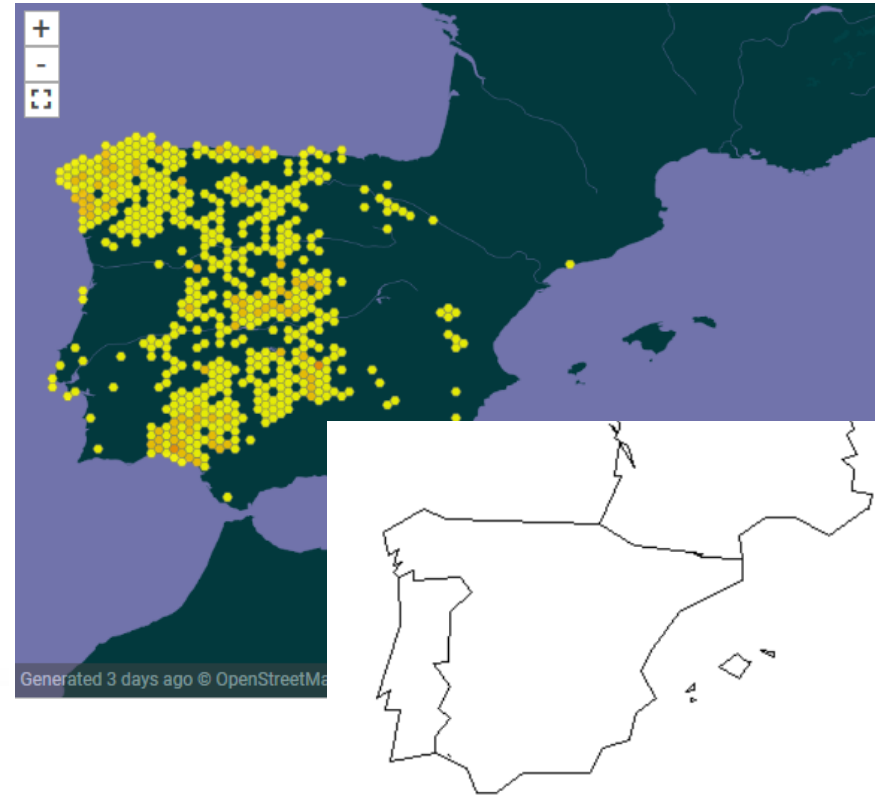
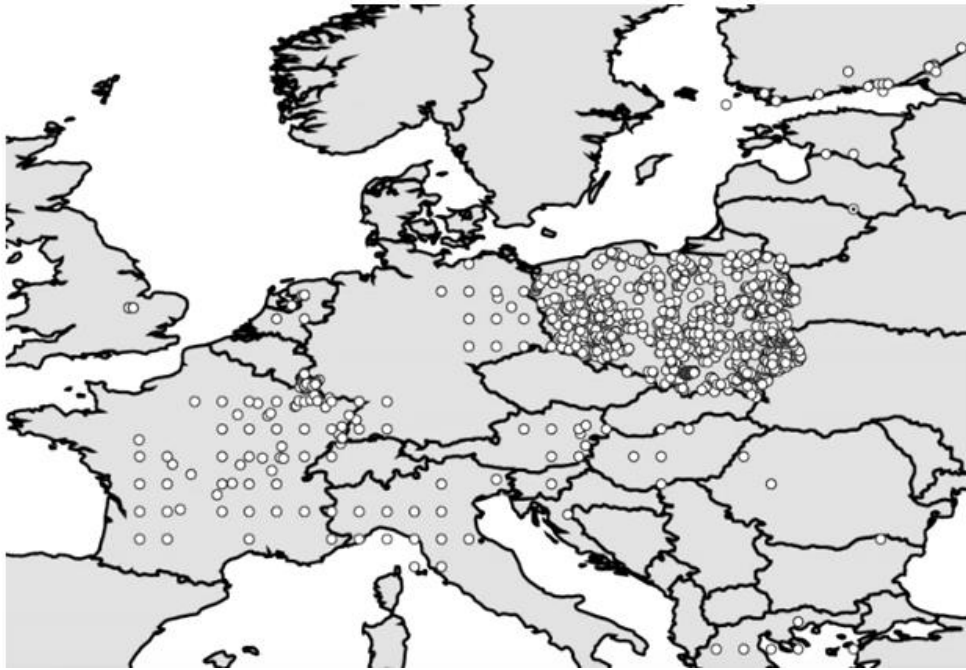
Occurrence data: precision vs. accuracy



- many **points** are actually grid **centroids**!
- check spatial resolution / spatial accuracy / location error / coordinate uncertainty
- don't use a finer resolution in the predictor variables!

(more on this later)

Occurrence data: spatial bias



- exclude under-surveyed regions from modelling
- spatial thinning / gridding to homogeneous resolution
- target-group background/absences: sites where other species of the group were recorded

Occurrence data bases

<https://www.gbif.org>



<https://www.inaturalist.org/observations>



<https://observation.org>



Observation.org

Add ▾ Observations ▾

<https://www.marinespecies.org>



<https://obis.org>



OCEAN BIODIVERSITY
INFORMATION SYSTEM

<https://www.idigbio.org>



<https://www.neotomadb.org>



just some
examples!

Occurrence data papers

Global Ecology and Biogeography

A Journal of
Macroecology

DATA PAPER |  Open Access |  

BioTIME: A database of biodiversity time series for the Anthropocene

Maria Dornelas , Laura H. Antão, Faye Moyes, Amanda E. Bates, Anne E. Magurran, Dušan Adam, Asem A. Akhmetzhanova, Ward Appeltans ... [See all authors](#) ▾

PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals

Ecological Archives E090-184

Kate E. Jones , Jon Bielby, Marcel Cardillo, Susanne A. Fritz, Safi, Wes Sechrest, Elizabeth H. Boakes, Chris Carbone, C

Updated distribution and biogeography of amphibians and reptiles of Europe

Neftali Sillero^{1,*}, João Campos¹, Anna Bonardi², Claudia Corti³, Rayan Pierre-Andre Crochet⁵, Jelka Crnobrnja Isailović^{6,7}, Mathieu Denoël⁸, Genti João Gonçalves⁹, Sergei Kuzmin¹⁰, Petros Lymberakis¹¹, Philip de Pous¹², Roberto Sindaco¹⁵, Jeroen Speybroeck¹⁶, Bert Toxopeus¹⁷, David R. Vieites

just some
examples!

Data Paper

NeoBiota 64: 1-21

<https://doi.org/10.3897/neobiota.64.55597> (11 Jan 2021)

Distribution of alien tetrapods in the Iberian Peninsula

▼ Fernando Ascensão, Marcello D'Amico, Ricardo C. Martins, Rui Rebelo, A. Márcia Barbosa, Joana Bencatel, Rafael Barrientos, Pedro Abellán, José L. Tella, Laura Cardador, José D. Anadón, Martina Carrete, Enrique Murru, Pedro Fernandes, Sara M. Santos, António Mira, Maria da Luz Mathias, Vítor S. Paulo, Henrique M. Pereira, César Capinha

ECOLOGY
ECOLOGICAL SOCIETY OF AMERICA

DATA PAPER |  Free Access

MAMMALS IN PORTUGAL: A data set of terrestrial, volant, and marine mammal occurrences in Portugal

Clara Grilo , Beatriz C. Afonso, Filipe Afonso, Marta Alexandre, Sara Aliácar, Ana Almeida, Ivan Prego Alonso, Francisco Álvares, Paulo Alves, Paulo Célio Alves, Pedro Alves, Anabela Amado, Vitor Amendoeira, Francisco Amorim, Guilherme da Silva Aparício, Ricardo Araújo, Fernando Ascensão, Margarida Augusto, Victor Bandeira, A. Márcia Barbosa, Soraia Barbosa, Sérgio Barbosa, Silvia Barreiro, Paulo Barros, Tânia Barros, Filomena Barros, Mafalda Basto, Joana Bernardino, Sara Bicho, Luis Eduardo Biedma, Marta Borges, Luis Braz, José Carlos Brito, Tiago Brito, João Alexandre Cabral, Javier Calzada, Cláudia Camarinha, Mafalda

Review of the systematics, distribution, biogeography and natural history of Moroccan amphibians

WOUTER BEUKEMA^{1,*}, PHILIP DE POUS^{2,3,4}, DAVID DONAIRE-BARROSO⁵, SERGÉ BOGAERTS⁶, JOAN GARCIA-PORTA³, DANIEL ESCORIZA⁷, OSCAR J. ARRIBAS⁸, EL HASSAN EL MOUDEN⁹ & SALVADOR CARRANZA³

packages for obtaining species occurrence data

- **geodata**



this is what we'll
use in class

- **rgbif**

- **rinat**

- **rvertnet**

- **rebird**

just some examples!

- **ecoengine**

- **rbison**

- **ridigbio**

- **spocc** (interface to several data sources)

Citing species occurrence data

- GBIF is not a data source/provider, but a repository
- we must **cite the actual sources** within it
- this video (~ minute 19-20) explains how to do that:



<https://www.youtube.com/watch?v=cO7b477Fo6w>

packages for cleaning species occurrence data

- `scrubr` (archived)
- `fuzzySim::cleanCoords`
- `biogeo`
- `bdclean`
- `CoordinateCleaner`

← this is what we'll
use in class

just some examples!

Import species occurrence data

```
# download GBIF occurrence data for a particular species:  
my_species <- "Loxodonta africana"  
download_window <- ext(10, 50, -40, -10)
```

```
library(geodata)
```

```
gbif_data <- sp_occurrence(genus = "",  
species = my_species, ext = download_window)
```

(optional)

if you only want to
download records from a
limited part of the world
(useful if your species is
too widespread)

NOTE: analyses will take long to run
if you work on a large area! Choose
a relatively small area (not much
larger than the example) to analyse
during the course

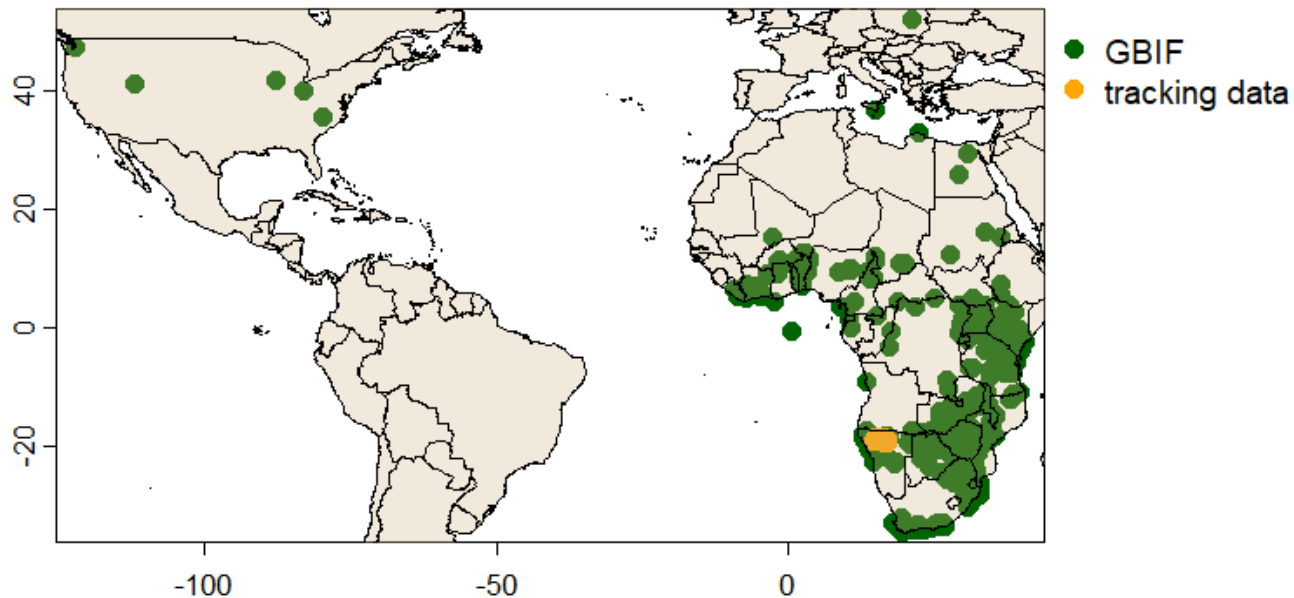
Import species occurrence data

You can (additionally or instead) use your own occurrence data from e.g. field sampling or telemetry. Then:

```
presences <- appendData(gbif_data, tracking_data)
```

```
plot(presences, "source")
```

```
plot(countries, add = TRUE)
```



Clean species occurrence data

```
# remove occurrences with duplicate, missing, unlikely,  
impossible, overly uncertain, or absence coordinates:
```

```
library(fuzzySim)  
  
presences <- cleanCoords(presences, coord.cols =  
c("decimalLongitude", "decimalLatitude"),  
  uncert.col = "coordinateUncertaintyInMeters",  
  uncert.limit = 10000,  
  abs.col = "occurrenceStatus")
```

NOTE: these are just some basic
automated procedures; for serious work,
carefully inspect and clean your data!

Predictor variables: data types and sources

Predictor variables: types

Abiotic

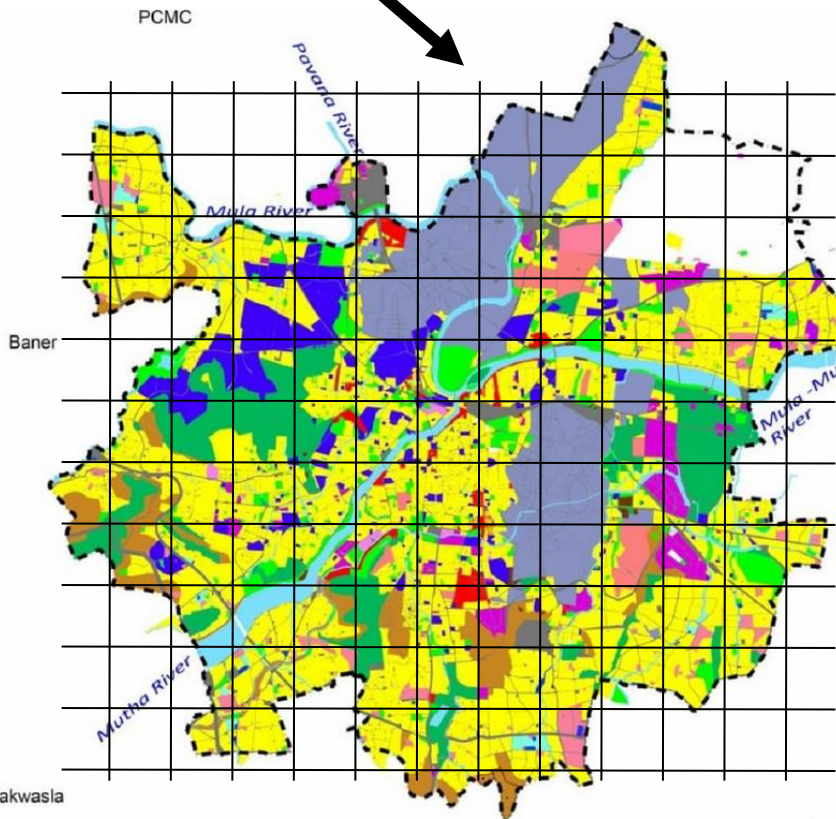
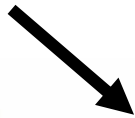
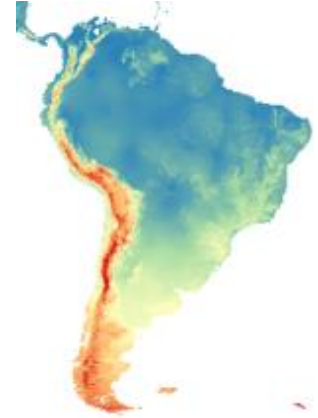
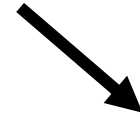
- climate: precipitation, temperature, seasonality, frost days, dry periods...
- topography: altitude, slope, aspect, bathymetry...
- physical/chemical properties of soil/water

Biotic

- (potential) presence or abundance of other species
- modifications of the environment caused by other species, incl. human influence (population density, urbanization, land use, traffic...)

Predictor variables: types

- **Continuous:** climate, topography
- **Categorical:** soil type, land use

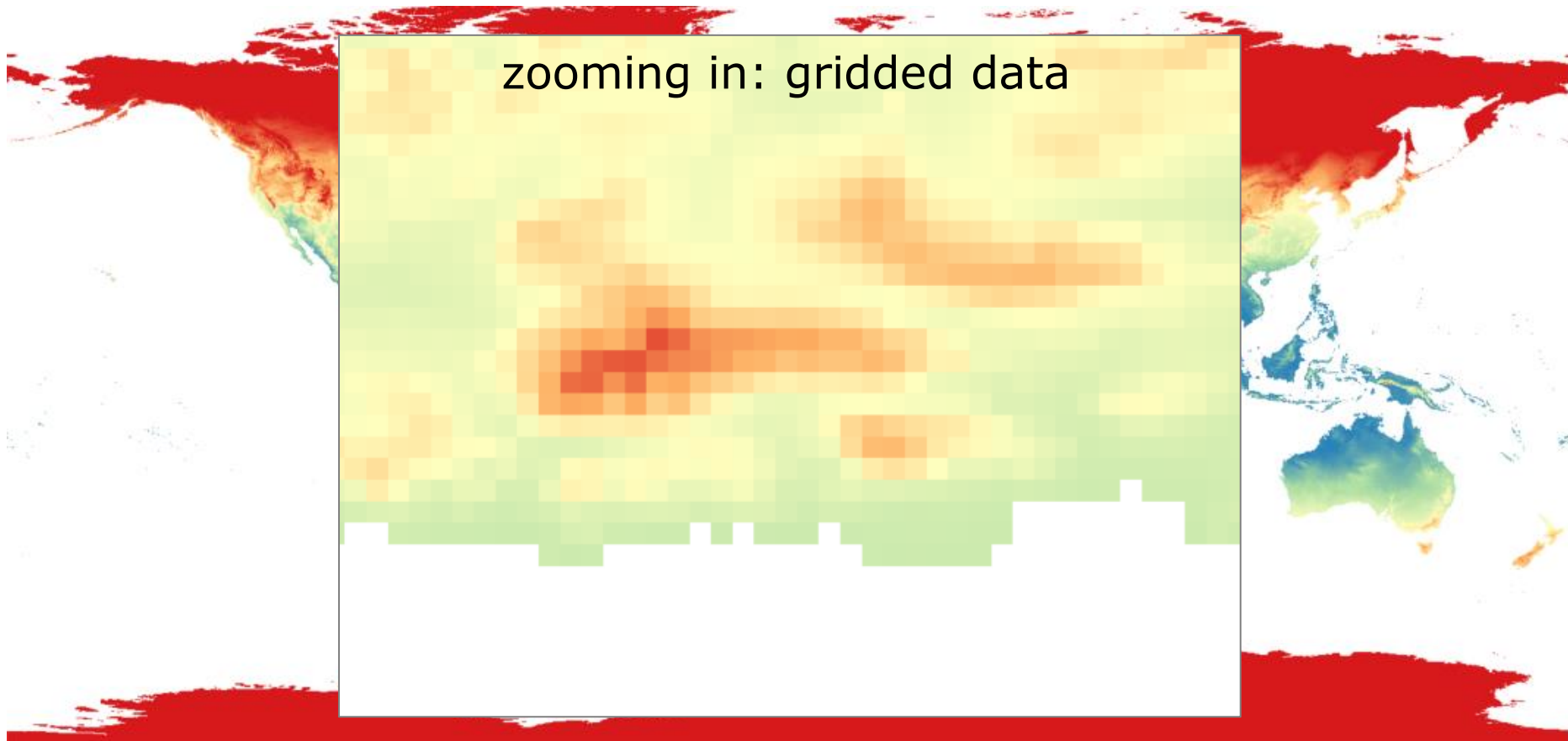


Can be converted to continuous:

- [target encoding](#)
- relevant classes
 - proportion
 - mean distance

Predictor variables

- obtained e.g. from interpolation, downscaling or remote sensing
- often in **grid** (raster) format: pixels = spatial units



[WorldClim interpolated temperature]

Predictor variables: data bases

<http://www.worldclim.org/>

WorldClim

<http://envirem.github.io/>

ENVIREM

<http://chelsa-climate.org/>

CHELSEA

<https://datadryad.org/resource/doi:10.5061/dryad.s2v81>

MERRAclim

<https://www.climatologylab.org/terraclimate.html>

TERRACLIMATE

<http://www.bio-oracle.org/>

Bio-ORACLE
Marine data layers for ecological modelling

<http://www.marspec.org/>

MARSPEC
OCEAN CLIMATE LAYERS FOR
MARINE SPATIAL ECOLOGY

<http://paleoclim.org/team-paleoclim/>

PaleoClim

just some
examples!

Predictor variables: data papers

International Journal of Climatology



Research Article | Full Access

WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas

Stephen E. Fick , Robert J. Hijmans

Methods in Ecology and Evolution



Free Access

CliMond: global high-resolution historical and future scenario climate surfaces for bioclimatic modelling

Darren J. Kriticos , Bruce L. Webber, Agathe Leriche, Noboru Ota, Ian Macadam, Janice Bathols, John K. Scott

SCIENTIFIC DATA

Data Descriptor | OPEN | Published: 05 September 2017

Climatologies at high resolution for the earth's land surface areas

Dirk Nikolaus Karger , Olaf Conrad, Jürgen Böhrer, Tobias Kawohl, Holger Kreft, Rodrigo W. Soria-Auza, Niklaus E. Zimmermann, H. Peter Linder & Michael Kessler

Global Ecology and Biogeography

A Journal of Macroecology

Full Access

Bio-ORACLE: a global environmental dataset for marine species distribution modelling

Lennert Tyberghein , Heroen Verbruggen, Klaas Pauly, Charles Troupin, Frederic Mineur, Olivier De Clerck

ECOGRAPHY

A JOURNAL OF SPACE AND TIME IN ECOLOGY

Research | Full Access

ENVIREM: an expanded set of bioclimatic and topographic variables increases flexibility and improves performance of ecological niche modeling

Pascal O. Title , Jordan B. Bemmels

SCIENTIFIC DATA

Data Descriptor | OPEN | Published: 20 June 2017

MERRAclim, a high-resolution global dataset of remotely sensed bioclimatic variables for ecological modelling

Greta C. Vega , Luis R. Pertierra & Miguel Ángel Olalla-Tárraga

Predictor variable selection

- ecological relevance
 - availability (at appropriate scale)
- [TODAY]
- collinearity (correlations with other variables)
 - statistical significance / informativeness /
contribution to the model
- [DAY 3]

Predictor variables: R packages

- `raster::getData`
- **geodata**  this is what we'll use in class
- `sdmpredictors`

not exhaustive lists!

for real work, do a thorough data search

Import predictor variables

```
# download/import variables with 'geodata' pkg:

# terrestrial:

layers <- worldclim_global(var = "bio", res = 5,
path = "../outputs")

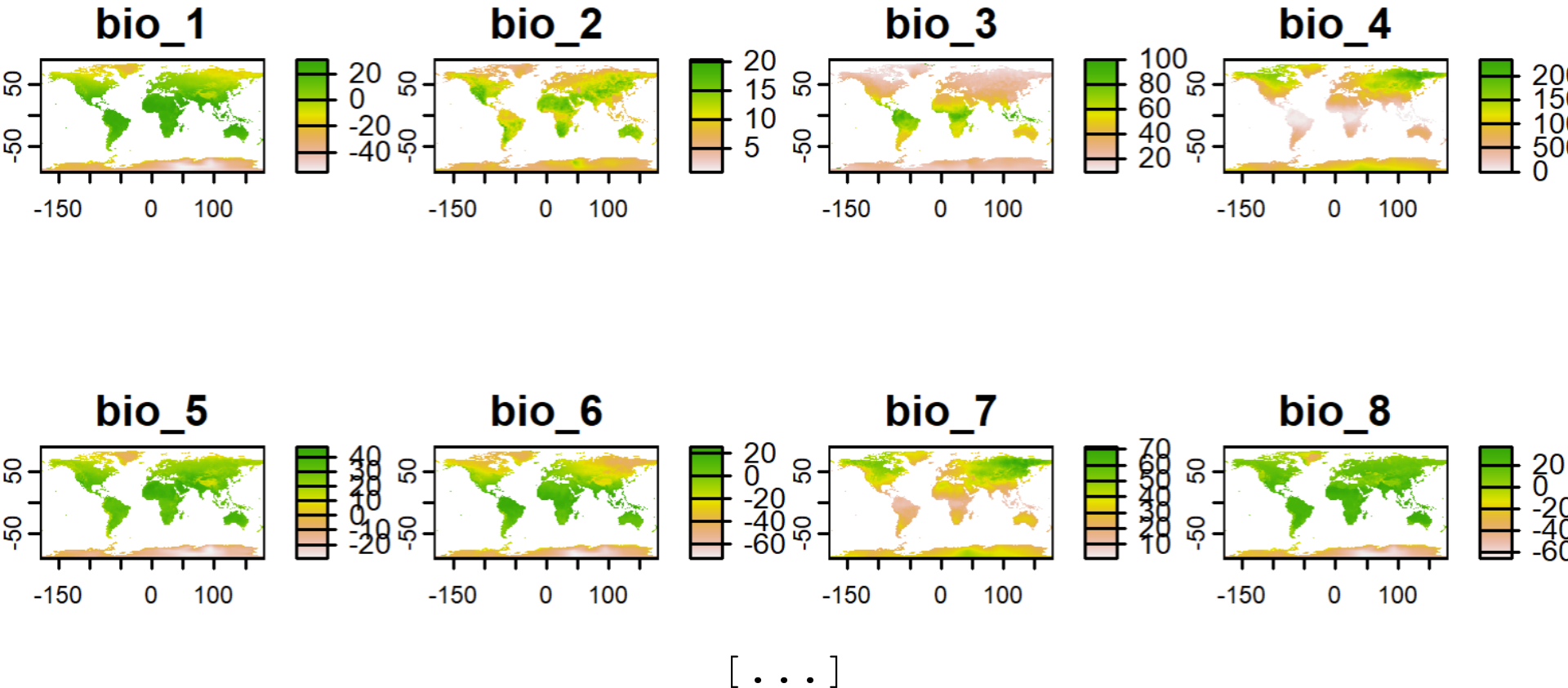
# marine:

layers <- rast(lapply(marine_layer_names,
function(v) bio_oracle(path = "../outputs",
var = v, stat = "Mean", benthic = FALSE)))
```

Import predictor variables

```
plot(layers)
```

```
# terrestrial:
```

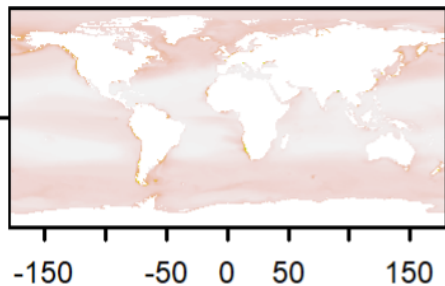


Import predictor variables

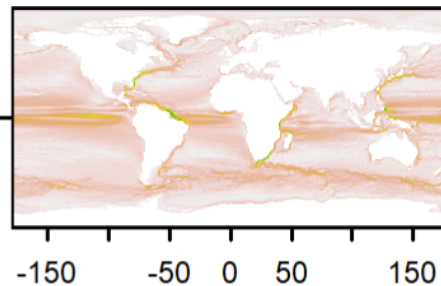
```
plot(layers)
```

```
# marine:
```

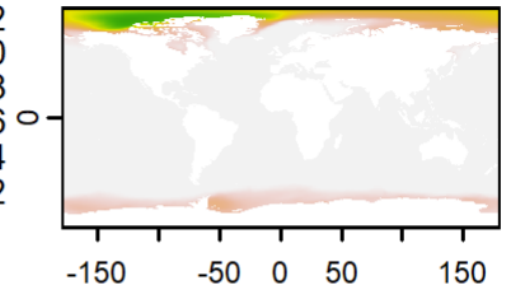
Chlorophyll.Mean



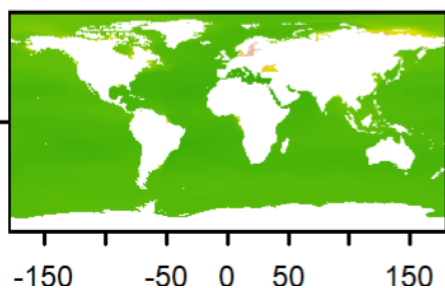
Current.Velocity.Mean



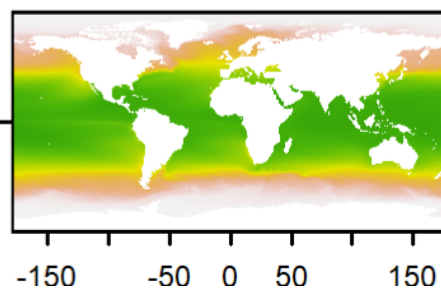
Ice.thickness.Mean



Salinity.Mean



Temperature.Mean



Data prep in general

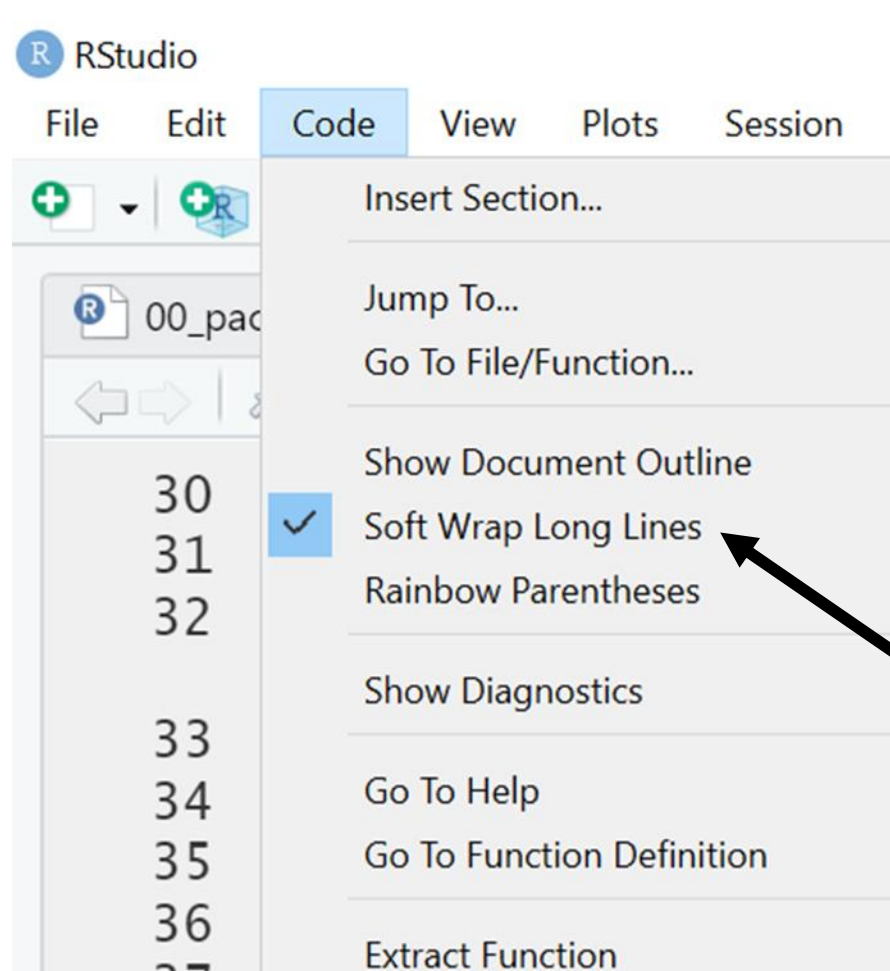
- gathering and formatting data (species occurrences + predictor variables) is usually the most **laborious and time-consuming** part of modeling work
- we'll use general data sources as examples but, for real work, you should take the time to gather data that are more complete and more appropriate for your region and study system

PRACTICAL

- download the course materials; **don't move or rename files** within the '*practicals*' folder (unless you're prepared to change all scripts accordingly!)
- the scripts set the working directory automatically and use **relative file paths**, so that they are **reproducible** on any computer (this is good practice, adopt it!)
- if you choose a different species, the scripts will automatically use a dedicated folder for its outputs, with **no need to manually create new folders**
- do the **assignments** so I can keep you on a good path!

PRACTICAL

run "**1a_data.R**" (in "*practicals/R*") one command at a time – observe each result before moving on!



- some parts of the scripts need to be run, bypassed or edited, depending on your species and region
- to avoid doing senseless things, you need to **read the script annotations!**
- **tick this** to see full lines without having to constantly scroll right

PRACTICAL

- if you have **limited experience with R**, better **stick to the scripts (changing only the species)** and keep **focus on the course subject**
- if you're more **experienced with R** and have **time left**, you can try to **adapt the script** e.g. to add different variables, compare with other modelling methods, etc.
- ask me for **help** whenever you need it, but remember this is not an R course and we can't get too distracted from the course subject