

Biodiversity Data Use

GBIF Secretariat

Version b34f741, 2022-03-30 08:40:29 UTC

Table of Contents

Description	1
Audience	1
Prerequisites	1
Learning objectives	2
Certification	2
Files for download	2
Data Processing	3
Assessing the Conservation Status of a Species	3
Data Processing	4
Fit for purpose data	4
Flags and Issues	6
Data Processing Tools	7
Handling Taxonomic Uncertainty	7
Species mis-identification	8
Synonymy	8
New names	8
Handling Data Quality	9
Geospatial Filters & Issues	9
Country centroids	9
Points along the equator or prime meridian	10
Uncertain location	10
Gridded datasets	10
Absence records	11
Establishment Means	12
Basis of Record	12
Old Records	12
Duplicates	13
Data Processing Pipeline	13
Preparing plotting variables	13
Downloading world map	14
Downloading map of Madagascar	14
Downloading GBIF datasets	14
Looking up species taxon keys	14
Sending request to download file	14
Actually downloading the file	15
Retrieving the citation of the downloaded dataset to use in your report	15
Data Visualization	15
Data cleaning step 1	16

Plotting raw records vs. cleaned records (step 1)	17
Data cleaning step 2	18
Plotting raw records vs. cleaned records (step 2)	18
Zooming in to madagascar	19
Data cleaning step 3	20
Plotting raw records vs. cleaned records (step 3)	21
Data cleaning step 4	22
Ecological Niche Models	23
What is an ecological niche model?	23
Uses	24
Commonly used algorithms	24
Environmental variables	26
Common sources of data	27
Selecting covariates (or environmental variables)	28
Training regions	28
Interpretation and Post-Processing of Niche Models	29
Model Evaluation	30
Thresholding a Niche Model	32
Projecting a Niche Model	33
The Big Caveat	33
Projection Uncertainty	34
Use Case - Modelling Species Distributions Under Climate Change	35
Exercise 1 - Starting Wallace	35
Exercise 2 - Loading occurrence points	36
Exercise 3 - Processing Occurrences	38
Exercise 4 - Loading environmental data in Wallace.	40
Exercise 6 - Partitioning Occurrence Data	41
Exercise 7 - Calibrating Niche Models with Maxent	41
Exercise 8 - Model Evaluation and Selection	43
Exercise 9 - Visualizing Model Results	44
Exercise 10 - Visualize model results in geographic space	44
Exercise 11 - Niche model projection	45
Exercise 12 - Calculating Environmental Similarity	47
Exercise 13 - Saving Your Session Code	48
Assessing the conservation status of a species	49
IUCN Red List of Threatened Species	49
IUCN Red List Categories and Criteria	50
Global vs National Red List Assessments	51
Red List assessment process	51
GBIF-mediated data and Red List assessments	52
Applying Criterion B - Restricted Geographic Range	52

Mapping standards for IUCN Red List Assessments	53
Minimum Documentation	54
Use Case - Red List Assessment	55
Scenario	55
Exercise - Applying IUCN Red List Criterion B	56
Key documentation	57
API	57
Cloud Computing	57
Darwin Core	57
Data publishing	58
Data publishing: IPT	58
Digitization	59
GBIF	59
Georeferencing	59
Invasive Species	59
Living Atlases	59
Miscellaneous	60
OpenRefine	60
Planning/Collaboration	60
Red List Assessments	61
Quality	62
R	62
Sensitive species	62
Taxonomy	62
Glossary	63
Acknowledgements	65
Course design and instruction	65
Translators	65
Spanish	65
French	65
Resources	65
Colophon	66
Suggested citation	66
Contributors	66
Licence	66
Persistent URI	66
Document control	66
Cover image	66

Description

This course takes a modular approach to training in the use of GBIF-mediated data and builds on the Data Use for Decision Making course developed as part of the Biodiversity Information for Development (BID) programme. Different use cases have been developed that will get you started on some of the many different ways you can use GBIF-mediated data.

This course is comprised of online content paired with quizzes and you should complete these first. You can then complete the relevant use case that will build on what you have learnt through a set of practical exercises. When offered as an onsite or virtual workshop, group work and social interaction are encouraged.

The course comprises of one compulsory course in Data Processing and then optional modules that you can follow based on the learning path that is most relevant to you. The first two optional modules that we have developed are:

- Ecological Niche Modeling
- Assessing the Conservation Status of a Species

Topics include:

- Data Processing
 - Processing a GBIF-mediated dataset and making it “fit-for-purpose”
- Ecological Niche Modeling
 - Introduction to running and interpreting a basic ecological model to determine the distribution of a species
 - Exploring niche modeling under different environmental conditions
- Assessing the Conservation Status of a Species
 - Using GBIF-mediated data for creating species distribution maps using IUCN mapping standards
 - Using GBIF-mediated data for assessing species conservation status using the IUCN Red List Categories and Criteria

Audience

This course is designed for individuals who work as researchers or technicians in biodiversity research or policy institutions. The instruction provided is particularly useful for those who have a need or desire to use GBIF-mediated data in their own research or analyses.

Prerequisites

- [Introduction to GBIF course](#)

Additionally, to make best use of the activities around this course, the participants should possess the following skills and knowledge:

- Basic skills in computer and internet use, and, in particular, in the use of spreadsheets.
- Basic understanding of computer-based geographical and statistical analysis tools e.g. GIS and R, and may have already run analyses using these tools.
- Basic knowledge about geography and biodiversity informatics: geography and mapping concepts, basic taxonomy and nomenclature rules.
- Willingness to disseminate the knowledge learned in the workshop with partners and collaborators in your project by adapting the biodiversity data use training materials to specific contexts and languages while maintaining their instructional value.
- A good command of English. While efforts are made to provide materials in other languages, instruction/videos will be in English.

Learning objectives

- Access GBIF mediated data through a range of access points
- Understand common data quality issues in GBIF downloads that may affect data use
- Apply data processing routines on GBIF downloads to create fit for purpose datasets
- Learn the difference between fundamental and realized niches
- Explain how to delimit a training region
- Generate a simple niche model
- Explain niche model results
- Identify areas of uncertainty in projection
- Develop a communication strategy and convincing arguments for the integration of biodiversity into decision making processes
- Apply criterion B of the IUCN Categories and Criteria for the assessment of a species conservation status using a fit-for purpose dataset
- Apply IUCN mapping protocols for the production of Red List species distribution maps

Certification

Certificates in the form of digital badges will be issued to participants that successfully complete a use case assignment. There are several use cases and you should complete the ones that are of most relevance to your work or that interest you.

Files for download

All files for the course may be downloaded from this page.

Data Processing

Presentations PDF versions of the accompanying presentation can be found below.

[Data Processing - ENGLISH](#)

[El procesamiento de datos - ESPAÑOL](#)

[Le traitement de données - FRANÇAIS](#)

Exercise data This [compressed file](#) (ZIP 7 KB) contains the exercise data.

Assessing the Conservation Status of a Species

Presentation

[GBIF-mediated data and the IUCN Red List Categories and Criteria - ENGLISH](#)

IUCN Red List Categories and Criteria

[IUCN Red List Categories and Criteria v3.1 - ENGLISH](#)

[Categorías y Criterios de la Lista Roja de la UICN v3.1 - ESPAÑOL](#)

[Catégories et Critères de la Liste Rouge de l'UICN v3.1 - FRANÇAIS](#)

Red List Guidelines

[Guidelines for Using the Red List Categories and Criteria version 15 - ENGLISH](#)

[Directrices de uso de las Categorías y Criterios de la Lista Roja de la UICN Versión 14 - ESPAÑOL](#)

[Lignes directrices pour l'utilisation des Catégories et Critères de la Liste rouge de l'UICN Version 14 - Français](#)

Criteria Summary Sheet

[Criteria summary sheet - ENGLISH](#)

[Resumen de los Criterios - ESPAÑOL](#)

[Résumé des Critères - FRANÇAIS](#)

Mapping Standards

[Mapping Standards - ENGLISH](#)

[Standard Attributes for Spatial Data - ENGLISH](#)

Regional and National Levels Guidelines

Guidelines for application of IUCN Red List Criteria at Regional and National Levels Version 4 - ENGLISH

Directrices para el uso de los Criterios de la Lista Roja de la UICN a nivel regional y nacional Versión 4 - ESPAÑOL

Lignes directrices pour l'application des Critères de la Liste rouge de l'UICN aux niveaux régional et national Version 4. - FRANÇAIS

Data Processing



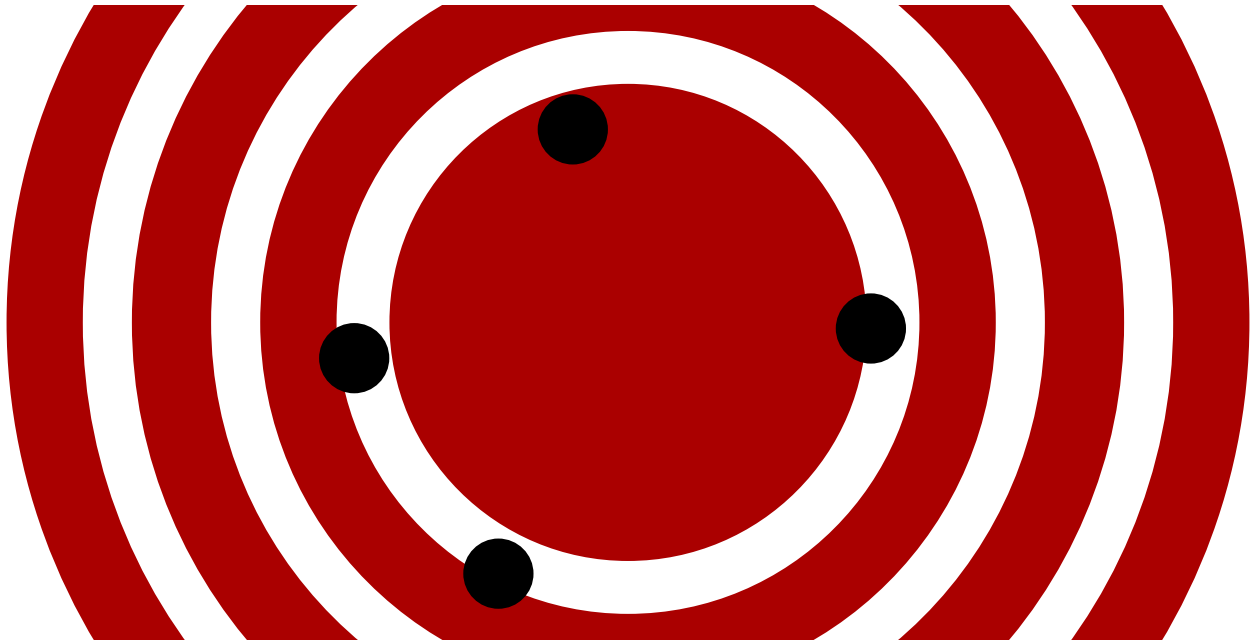
In the data processing module, you will familiarize yourself with the concept of fit-for purpose datasets and some commonly used data filters that you may want to consider for creating your own fit-for-purpose dataset.

Fit for purpose data

Almost always you will want to post-process your GBIF download in some way to fit your purposes. Sometimes you will have to make difficult judgement calls for your particular use-case. Whenever you are dealing with thousands or millions of records, you will never quite know the true quality of the source data. It is important to keep in mind that you are always just mitigating data quality issues, not eliminating them.

The data that we get in GBIF download, will contain data from a range of sources and the data will likely vary in its correctness and consistency. Correctness and consistency are two ways of documenting data errors and are measures of data quality. These are measures of how well the data gatherer was able to capture the true value being investigated. The nature of GBIF's data publication workflow means that the correctness and consistency of the data can vary dependent on the data publishers and the source of the data. Knowing these properties of the data you have, will help you to understand the ways in which you can and cannot clean, validate and process the data.

- Correctness (Accuracy) - closeness of measured values, observations or estimates to the real or true value e.g. has the species been identified correctly or the collection locality been identified correctly.



For instance, if we are studying plant biogeography in Indonesia, and want to do a specific analysis for only one of the islands within the archipelago, then an appropriate question might be - Have localities on the island been correctly georeferenced?

- Consistency (Precision) - level of resolution of the data e.g. precision of coordinates, taxonomic determination.

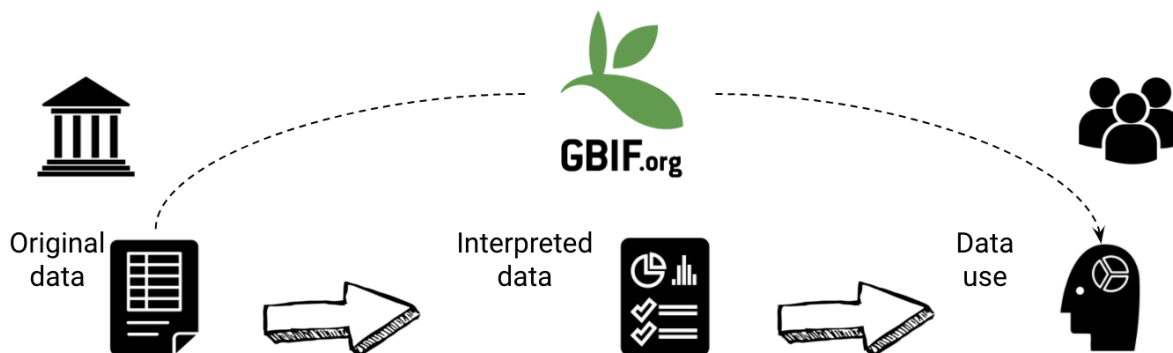


In the Indonesian example, an appropriate question might be - Does the uncertainty in the coordinate estimate allow for the occurrence record to not be on the island?

As a general rule, for most analyses you want highly accurate data although the level of precision may vary dependent on your analysis. GBIF can help you to determine the accuracy and precision of the data through, for example, filters and issue flags, however, you must always double-check!

Flags and Issues

The GBIF network publishes datasets, integrating them into a common access system. Here users can retrieve data through common search and download services. During the indexation process over the raw data, GBIF adds issues and flags to records with common data quality problems. These may contain useful information for you as a user to create fit-for-purpose datasets.



Remarks are shown on the individual occurrence pages to explain the process done after interpretation:

*Excluded means the original data couldn't be interpreted, so is excluded in the interpreted fields.

*Altered means the original data is modified in the interpretation process to be indexed in GBIF.org.

*Inferred means the Using other record information the data indexed is inferred, if the original is empty.

Excluding all records with a particular issue is not currently possible with the search interface. It is possible to filter all records you are not interested in with issues by selecting the particular issue and hitting the reverse button. However, reversing will still only give you all other flagged occurrences and not issue-free records. This is something that GBIF is working to improve. (at occurrence search)

The screenshot shows the GBIF Occurrences search interface. On the left, a sidebar lists various filters. The filter 'Basis of record invalid' is checked and highlighted with a red box, showing a count of 12,214,242. Below the list are 'CLEAR' and 'REVERSE' buttons, also highlighted with red boxes. The main content area shows a search for '12,214,242 RESULTS' with tabs for TABLE, GALLERY, MAP, TAXONOMY, METRICS, and DOWNLOAD. The table header includes columns for Scientific name, Country or area, Coordinates, and Month & year. The table body is currently empty, showing only the header row.

A full overview of all issues and flags can be here: <https://data-blog.gbif.org/post/issues-and-flags/>

Data Processing Tools

While GBIF filters will allow for some data processing i.e. select only those data for download that fulfil certain criteria, it is highly recommended that you use additional data processing tools. Your choice of which tool you use will be based on personal access, familiarity and utility of each of the tools. Tools that can be used for this are:

- Spreadsheet editing software e.g. Excel, Google Sheets (smaller datasets)
- OpenRefine
- R packages and scripts eg rgbif, CoordinateCleaner, scrubr and biogeo (automated data processing)
- Geographical Information Systems (GIS) e.g. ArcGIS, QGIS and MapInfo

It is always important to include a data visualisation step in your data processing so that you can identify anomalous data points that may have been missed during the processing stage. For those processing in R, this can be done within R, however, with OpenRefine or spreadsheet editing software you may have to use GIS software or even tools such as Google Earth for data visualisation.

Handling Taxonomic Uncertainty

Uncertainty surrounding the taxonomy of a data point can arise for several reasons:

- Species mis-identification
- Synonymy
- Novel names

Species mis-identification

Species identification is a complex process, with species typically described from a certain set of characters identified in a published species description and linked to a type specimen held within a scientific collection that be used for validation of species identification. Where taxa are very similar or a set of complex traits are required for correct identification, specific taxonomic expertise may be required that data publishers may not possess leading to a mis-identification of a species. As users, you must have a clear understanding of how taxonomic determinations for your interest group are made:

- What are the characters used for defining the species?
- Are these characters easily confused or captured when the species is observed or collected?
- Are there related species that could be easily confused with the species you are interested in?

If you think that there is a risk that species may be incorrectly identified, you can take a conservative approach to the data you use and only use those data linked to specimens in collections where taxonomic validation would be possible and eliminate other data sources. Another approach may be to use associated data such as collector information, media, DNA sequences etc to validate the taxonomic determination.

Synonymy

Synonymy can arise when the same species has been described several times and a new name is given to the species each time it is described, or, when there is a change in the taxonomy of a species, for example, a species is moved from one genus to another. Only one species name can be accepted, and other names are what we call synonyms. These synonyms may still be in use to a lesser or greater extent and you should be sure when getting data from GBIF to obtain data for the taxonomic name you need. GBIF's taxonomic backbone differentiates between accepted scientific names and synonyms, and unique identifiers in the form of taxon keys. Species searches <https://www.gbif.org/species/search> allow for filtering for accepted names and synonyms and taxon keys can be used for programmatic searches of GBIF.

Taxon Keys Scientific names can be messy. If you are accessing GBIF-mediated data programatically as opposed to via the website, taxon keys provide an effective way for defining searches based on taxonomy. Taxon keys are issued at the species, genus family, order, phylum and kingdom level. Unique identifiers are issued to accepted names with synonyms of those accepted names issued the same identifier. So, it may make sense to sort out the species by their unique taxon keys provided during the indexation of the dataset by GIBF.

New names

There may be instances where the scientific name does not match any name in the GBIF backbone, perhaps because the species is newly described, or is not within a checklist used by GBIF to construct its backbone. These names are flagged with the TAXON_MATCH_HIGHERRANK flag indicating that the scientific name has not been recognised but that the data point has matched at a higher taxonomic level eg. genus or family. This flag can be used for identifying and filtering for these data. When names have been misspelled or badly formatted, there is also a TAXON_MATCH_FUZZY flag

that can be used for identifying and filtering names that can only match the taxonomic backbone using a fuzzy, non exact match.

Handling Data Quality

Filtering the data allows you as a user to obtain the data that is most fit for purpose. All searches have a set of filters that can be used for finding the data you need, and occurrence searches have a set of additional 'Advanced' search filters for users that need to do more advanced filtering. While filters may allow you to filter out data that may not be relevant, or be of lower quality for your purposes, additional filtering may be required either manually or programmatically to deal with additional data quality issues that arise during the GBIF data publishing model. Below are some common data filters that you as a user might consider to make the data more fit-for-purpose.

Geospatial Filters & Issues

The data can be filtered spatially in an occurrence search in one of 3 ways:

- Country or area/Continent - data is filtered by country and will include data within the Exclusive Economic Zone (EEZ)
- Administrative area - this filter uses the GADM database <https://gadm.org/data.html> of administrative areas for all countries in the world to allow for GBIF removes common geospatial issues by default if you choose to have data with a location.
- Location - this filter allows you to filter for data with coordinates and/or draw your own polygon shape filters or use a GeoJSON file to delimit your own shape filter. If you filter for those data with coordinates, a number of geospatial issues associated with the data publishing workflow will be eliminated. These are:
 - Zero Coordinates- Coordinates are exactly (0,0) or what is sometimes called "null island". Zero-zero coordinate is a very common geospatial issue. GBIF removes (0,0) when hasgeospatialissue is set to FALSE.
 - Country coordinate mis-match - Data publishers will often supply GBIF with a country code (US,TW,SE,JP...). GBIF uses the two letter ISO 3166-1 alpha-2 coding system - https://en.wikipedia.org/wiki/ISO_3166-1_alpha-2. When a point does not fall within the country's polygon or EEZ, but says that it should occur within the country, it gets flagged as having "country coordinate mis-match" and will be removed if data are filtered for locations.
 - Coordinate invalid - If GBIF is unable to interpret the coordinates i.e. the coordinates.
 - Coordinate invalid - The coordinates are outside of the range for decimal lat/lon values (-90,90), (-180,180).

Country centroids

Country centroids are where the observation is pinned to the centre of the country instead of where the taxon was observed or recorded. Country centroids are usually records that have been retrospectively given a lat-lon value based on a textual description of where the original record was located. Geocoding software uses gazetteers, geographical dictionaries or directories used in conjunction with a map or atlas, to attribute coordinates to place names. So, if the record simply says

"Brazil", some publishers will put the record in the center of Brazil. Similarly if the record simply says "Texas" or "Paris" the record will go in the center of those regions. This is almost exclusively a feature of museum data (PRESERVED_SPECIMEN), but it can also happen with other types of records as well.

Identifying country centroid data is currently not possible using GBIF filters, however, the R package CoordinateCleaner can be used for identifying and filtering for country centroids.

Points along the equator or prime meridian

Some publishers consider zero and NULL to be equivalent so that empty latitude and longitude fields for a record are given a zero value. As a result, records end up being plotted along the equator and prime meridian lines.

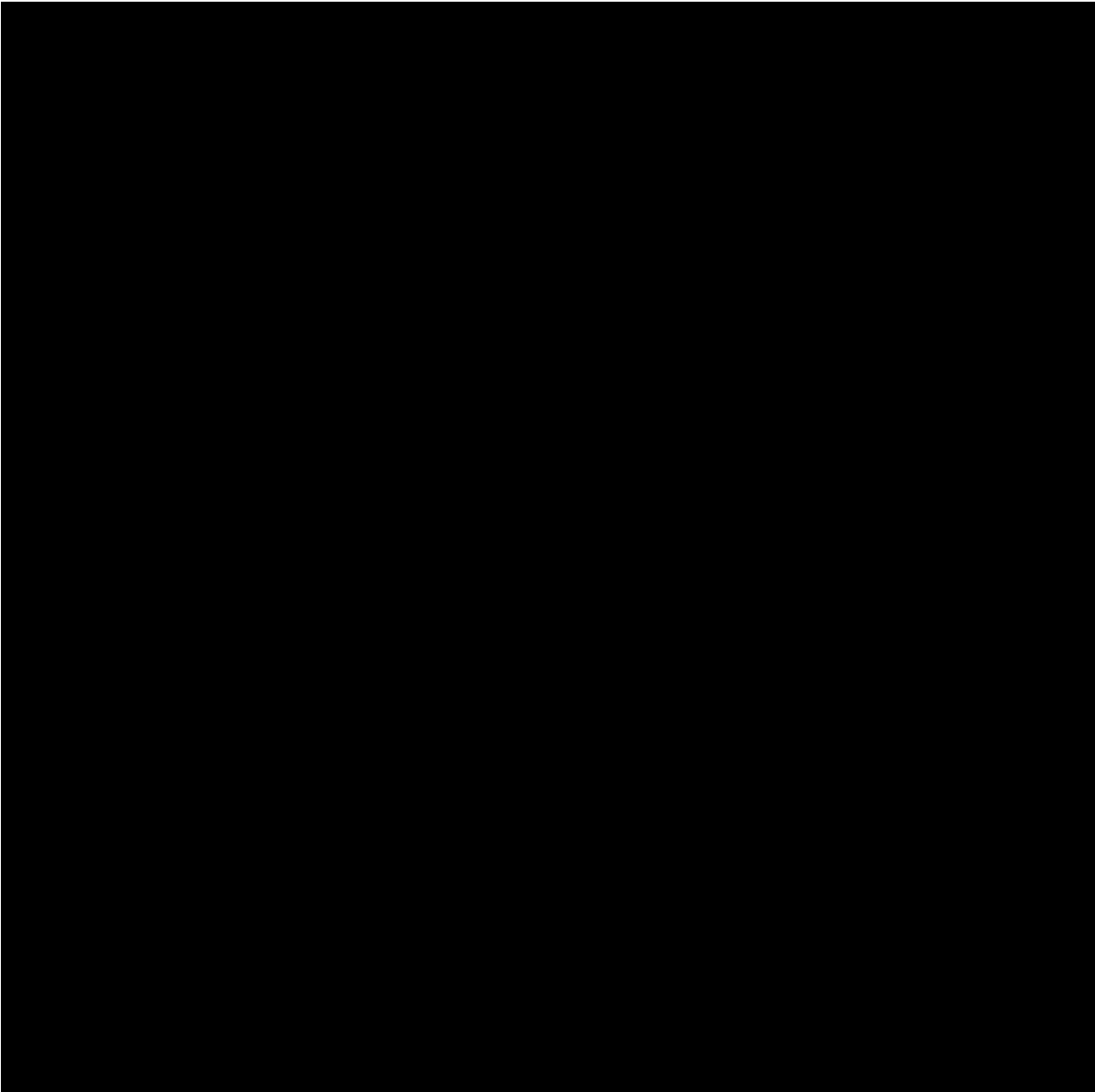
Uncertain location

Often you will want to be sure that the coordinates give a certain location and are not really 1000s of km away from where the organism was observed or collected. There are two fields - coordinate precision and coordinateUncertaintyInMeters - in Darwin Core that you get with a SIMPLE CSV download. that you can use to filter by "uncertainty". However, these fields are not used very often by publishers who feel that their records are fairly certain (from a GPS) and we would recommend not filtering out missing values.

There are also a few "fake" values for coordinate uncertainty that you should be aware of. These values are errors produced by geocoding software and do not represent real uncertainty values. These "fake" values are 301, 3036, 999 and 9999. In the case of the value 301, the uncertainty is often much-much greater than 301 and actually represents a country centroid.

Gridded datasets

Gridded datasets are a known problem at GBIF. Many datasets have equally-spaced points in a regular pattern. These datasets are usually systematic national surveys or data taken from some atlas ("so-called rasterized collection designs"). Georeferenced occurrences are snapped to a central point



Most publishers of gridded datasets actually fill in one of the following columns: coordinateuncertaintyinmeters, coordinateprecision, footprintwkt So filtering by these columns can be a good way to remove gridded datasets. The R package Coordinate cleaner also has a function for removing gridded datasets. GBIF has an experimental API for identifying datasets which exhibit a certain amount of "griddiness". You can read more here: <https://data-blog.gbif.org/post/finding-gridded-datasets/>

Absence records

By default, both presence and absence records are shown when you search www.gbif.org. Absence records confirm that a species was not found at a specific locality when that area was surveyed and this information can be useful in, for example, developing ecological niche models. However, you may only be interested in presence records and in this instance you can filter for only presence records using the Occurrence Status filter.

Establishment Means

The Darwin Core term `establishmentMeans` identifies the process by which the biological individual(s) represented in the Occurrence became established at the location. As such, it can serve as a useful filtering tool for identifying records that are outside of a species native range with accepted terms for this field being `native`, `nativeReintroduced`, `introduced`, `introducedAssistedColonisation`, `vagrant` and `uncertain`. Currently, GBIF records can be searched using the older vocabulary terms `native`, `introduced`, `naturalized`, `native`, `managed` and `uncertain` - https://rs.gbif.org/vocabulary/gbif/establishment_means.xml, and these will be updated in late 2022. In some instances, removing "MANAGED" records will remove zoo records.

Use this filter cautiously, however, as most records do not contain this information and so would be excluded from a search with this filter on. We would recommend to use the information within the `Establishment Means` term for filtering after download.

Basis of Record

Basis of record is a Darwin Core term that refers to the specific nature of the record and can refer to one of 6 classes:

- **Living Specimen** - a specimen that is alive, for example, a living plant in a botanical garden or a living animal in a zoo.
- **Preserved Specimen** - a specimen that has been preserved, for example, a plant on an herbarium sheet or a cataloged lot of fish in a jar.
- **Fossil Specimen** - a preserved specimen that is a fossil, for example, a body fossil, a coprolite, a gastrolith, an ichnofossil or a piece of petrified tree.
- **Material Citation** - A reference to, or citation of, one, a part of, or multiple specimens in scholarly publications, for example, a citation of a physical specimen from a scientific collection in taxonomic treatment in a scientific publication or an occurrence mentioned in a field note book.
- **Human Observation** - an output of human observation process eg. evidence of an occurrence taken from field notes or literature or a records of an occurrence without physical evidence nor evidence captured with a machine.
- **Machine Observation** - An output of a machine observation process for example a photograph, a video, an audio recording, a remote sensing image or an occurrence record based on telemetry.

Basis of record should allow users to filter out those individuals in ex-situ collections such as zoos and botanic gardens or fossils as well as filter for those records based on whether the record is based on a specimen or an observation, which can support taxonomic validation. You should note that, even though this can be a useful filter, data publishers do not always fill the basis of record field correctly, or, there may be nuances in the data that may not be immediately obvious to a user e.g. <https://data-blog.gbif.org/post/living-specimen-to-preserved-specimen-understanding-basis-of-record/> and you should always double check your data before use.

Old Records

GBIF has many museum records that might be older than what is desired for some studies.

Duplicates

Duplication of records can occur when several records of the same individual are made. This can occur from for instance, a researcher depositing several specimens from an individual tree in herbaria around the world who all then publish these data on GBIF, or when an individual has been deposited in a natural history collection and the individual was also sampled for its DNA. In this instance, there will be a record for the specimen in the collections and one for the DNA sequence.

GBIF has recently introduces a clustering function in its advanced search that allows users to identify clusters of records i.e. records that appear to be derived from the same source. This allows users to identify potential duplicated data and filter for these out of your download. Note that if you filter out those records that are in a cluster, you will lose all records found within that cluster and will lose potentially useful data. The filter may be better used to indicate the extent to which there is duplication in the dataset, or for independent downloads of the clustered and non-clustered datasets for comparison.

Data Processing Pipeline

Author: Arman Pili

This is an example script that can be used for developing your own data processing pipelines. The script takes you through the process of downloading, visualizing and cleaning GBIF-mediated data. Remember that this script is only a tool and that, ultimately, it is you the user to ensure that the data is fit for purpose.

You can download the R Markdown document for your own use here - [Lemur catta project](#). You will be redirected to another page. Right click and "Save as..." to save the file to your hard drive. (If you save it as an .Rmd file instead of a .txt file you will benefit from the additional functionality that comes with that format when you open it in RStudio)

The following packages will be needed

```
library(rgbif)           #notes
                          #for downloading datasets from gbif
library(countrycode)    #for getting country names based on countryCode important
library(rnaturalearth)  #for downloading maps
library(sf)             #for manipulating downloaded maps
library(tidyverse)      #for tidy analysis
library(CoordinateCleaner) #for quality checking of occurrence data
```

Preparing plotting variables

These will return the variables you will need for plotting your data later

Downloading world map

```
world_map <- ne_countries(returnclass = "sf")
```

Downloading map of Madagascar

```
country_map <- ne_countries(country = "Madagascar",  
                             scale = "medium",  
                             returnclass = "sf")
```

Downloading GBIF datasets

Looking up species taxon keys

I chose to use the Ring-Tailed Lemur (**Lemur catta**) for demonstrative purposes.

```
name_backbone(name = "Lemur catta", rank = "species")
```



A tibble: 1 x 24

	usageKey <int>	scientificName <chr>	canonicalName <chr>	rank <chr>	status <chr>	confidence <int>	matchType <chr>	kingdom <chr>	phylum <chr>
1	2436412	Lemur catta Linnaeus, 1758	Lemur catta	SPECIES	ACCEPTED	98	EXACT	Animalia	Chordata

1 row | 1-10 of 24 columns

In the next step, you will need a taxon key. Here, the taxon key is equivalent to the usageKey.

What is the usageKey of Lemur catta?

```
usageKey <- 2436412  
# alternatively  
usageKey <- name_backbone(name = "Lemur catta", rank = "species") %>%  
  pull(usageKey)
```

Looks like there are a lot of possible taxon keys for **Lemur catta**. I searched it in GBIF taxonomic backbone instead: 2436412

Sending request to download file

Remember to fill in your own GBIF login credentials for "user", "pwd" and "email"

```
gbif_download_key <- occ_download(pred("taxonKey", 2436412),
  format = "SIMPLE_CSV",
  user = "*****",
  pwd = "*****",
  email = "*****")
```

Actually downloading the file

```
occ_download_wait(gbif_download_key)

gbif_download <- occ_download_get(gbif_download_key,
  overwrite = TRUE) %>%
  occ_download_import() %>%
  setNames(to_lower(names(.)))

head(gbif_download, 10) #view first ten rows
```



A tibble: 10 x 50

gbifid	datasetkey	occurrenceid	kingdom	phylum	class	order	family	genus	species
3468953484	50c9509d-22c7-4a22-a47d-8c48425ef4a7	https://www.inaturalist.org/observations/106150988	Animalia	Chordata	Mammalia	Primates	Lemuridae	Lemur	Lemur catta
3468952451	50c9509d-22c7-4a22-a47d-8c48425ef4a7	https://www.inaturalist.org/observations/106150987	Animalia	Chordata	Mammalia	Primates	Lemuridae	Lemur	Lemur catta
3465922984	50c9509d-22c7-4a22-a47d-8c48425ef4a7	https://www.inaturalist.org/observations/105662768	Animalia	Chordata	Mammalia	Primates	Lemuridae	Lemur	Lemur catta
3465911165	50c9509d-22c7-4a22-a47d-8c48425ef4a7	https://www.inaturalist.org/observations/105662767	Animalia	Chordata	Mammalia	Primates	Lemuridae	Lemur	Lemur catta
3456788913	50c9509d-22c7-4a22-a47d-8c48425ef4a7	https://www.inaturalist.org/observations/102259497	Animalia	Chordata	Mammalia	Primates	Lemuridae	Lemur	Lemur catta
3455434832	50c9509d-22c7-4a22-a47d-8c48425ef4a7	https://www.inaturalist.org/observations/102259223	Animalia	Chordata	Mammalia	Primates	Lemuridae	Lemur	Lemur catta
3436826156	be614ad1-d0d1-4e27-a3e4-212aca8b5da6	http://d.luomus.fi/MY.10264315	Animalia	Chordata	Mammalia	Primates	Lemuridae	Lemur	Lemur catta
3398989158	50c9509d-22c7-4a22-a47d-8c48425ef4a7	https://www.inaturalist.org/observations/99359760	Animalia	Chordata	Mammalia	Primates	Lemuridae	Lemur	Lemur catta
3386809779	d31840e3-06ed-4ab7-bbed-783d0da5a869		Animalia	Chordata	Mammalia	Primates	Lemuridae	Lemur	Lemur catta
3384587415	50c9509d-22c7-4a22-a47d-8c48425ef4a7	https://www.inaturalist.org/observations/97270744	Animalia	Chordata	Mammalia	Primates	Lemuridae	Lemur	Lemur catta

1-10 of 10 rows | 1-10 of 50 columns

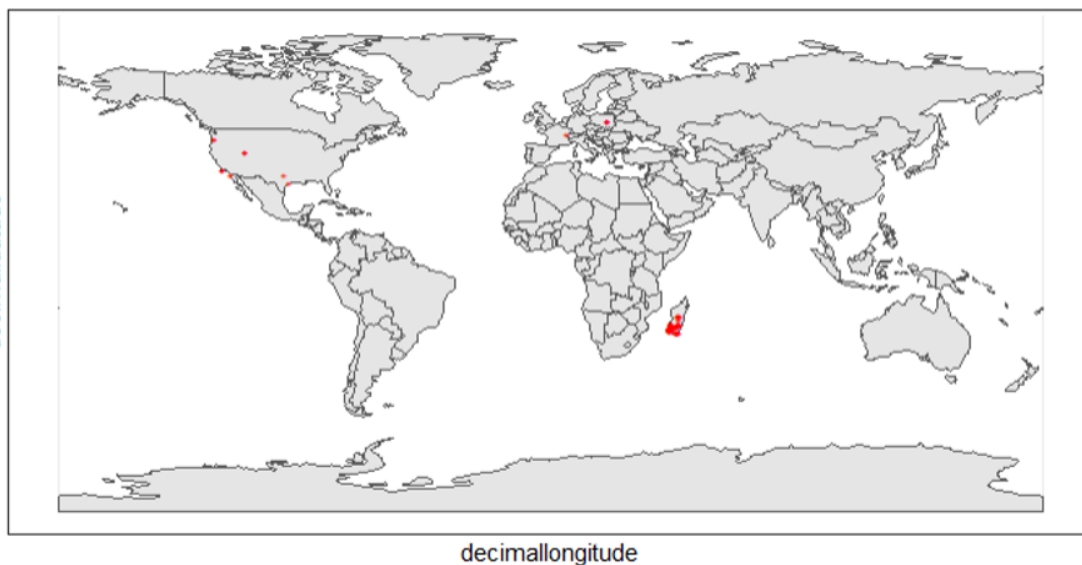
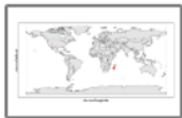
Retrieving the citation of the downloaded dataset to use in your report

```
print(gbif_citation(occ_download_meta(gbif_download_key))$download)
```

Data Visualization

Lemur catta is native to Madagascar; but just to make sure, let's check where else it can be found

```
ggplot() +
  geom_sf(data = world_map) +
  geom_point(data = gbif_download,
            aes(x = decimallongitude,
                y = decimallatitude),
            shape = "+",
            color = "red") +
  theme_bw()
```



From initial look, what's wrong with the distribution of the Lemur?

Whelp! seems like there are unusual occurrences outside its native range. Let's check further.

```
table(gbif_download$countrycode)
```

```

      CN  CZ  DE  DK  FR  IL  MG  PL  US  ZA  ZW  ZZ
259    2   1  31   1   2   1 437   6  37   2   1  38

```

Data cleaning step 1

With each step note the number of records that you are removing

Removing data recorded based on fossil or living specimens, and records from alien/invasive populations

```

clean_step1 <- gbif_download %>%
  as_tibble() %>%
  filter(!basisofrecord %in% c("FOSSIL_SPECIMEN",
                              "LIVING_SPECIMEN"),
         !establishmentmeans %in% c("MANAGED",
                                    "INTRODUCED",
                                    "INVASIVE",
                                    "NATURALISED"))
print(paste0(nrow(gbif_download)-nrow(clean_step1), " records deleted; ",
            nrow(clean_step1), " records remaining."))

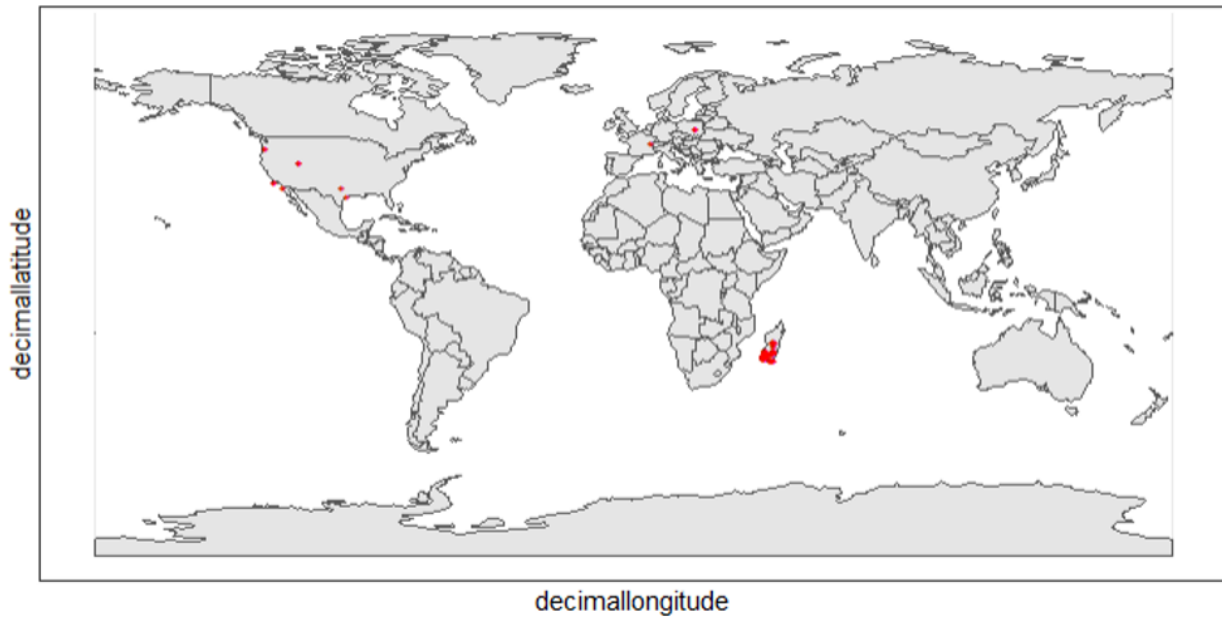
```

Plotting raw records vs. cleaned records (step 1)

```

ggplot() +
  geom_sf(data = world_map) +
  geom_point(data = gbif_download,
            aes(x = decimallongitude,
                y = decimallatitude),
            shape = "+",
            color = "black") +
  geom_point(data = clean_step1,
            aes(x = decimallongitude,
                y = decimallatitude),
            shape = "+",
            color = "red") +
  theme_bw()

```



Data cleaning step 2

Flagging records with problematic occurrence information using functions of the `coordinatecleaner` package.

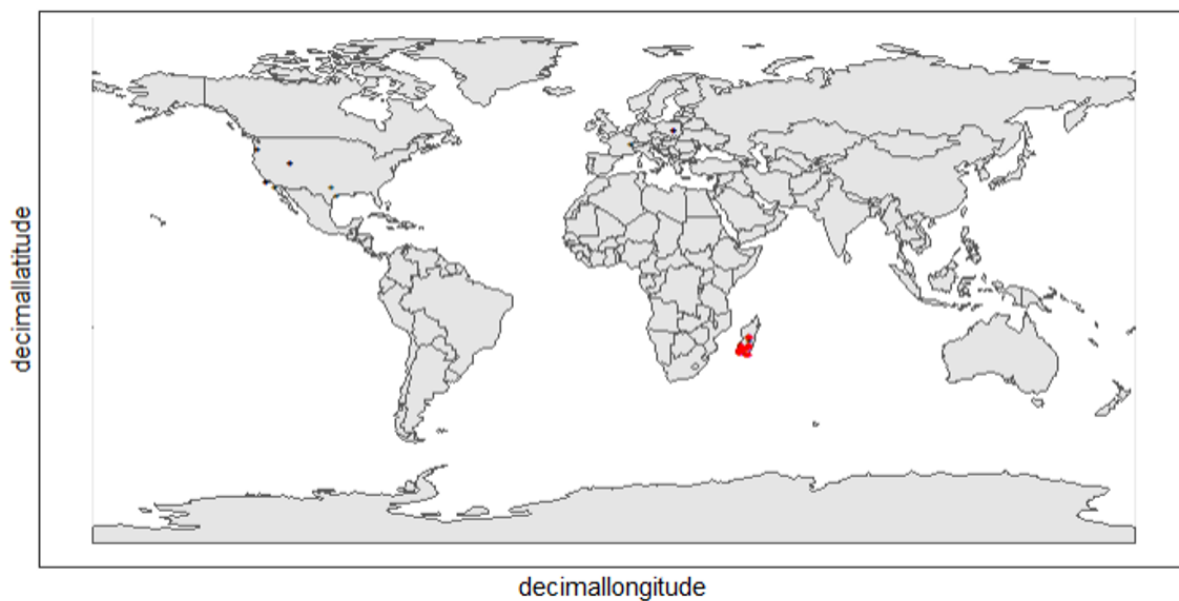
```
clean_step2 <- clean_step1 %>%  
  filter(!is.na(decimallatitude),  
         !is.na(decimallongitude),  
         countrycode == "MG") %>% # "MG" is the iso code for Madagascar  
  cc_dupl() %>%  
  cc_zero() %>%  
  cc_equ() %>%  
  cc_val() %>%  
  cc_sea() %>%  
  cc_cap(buffer = 2000) %>%  
  cc_cen(buffer = 2000) %>%  
  cc_gbif(buffer = 2000) %>%  
  cc_inst(buffer = 2000)  
print(paste0(nrow(gbif_download)-nrow(clean_step2), " records deleted; ",  
            nrow(clean_step2), " records remaining."))
```

Plotting raw records vs. cleaned records (step 2)

```

ggplot() +
  geom_sf(data = world_map) +
  geom_point(data = gbif_download,
            aes(x = decimallongitude,
                y = decimallatitude),
            shape = "+",
            color = "black") +
  geom_point(data = clean_step2,
            aes(x = decimallongitude,
                y = decimallatitude),
            shape = "+",
            color = "red") +
  theme_bw()

```



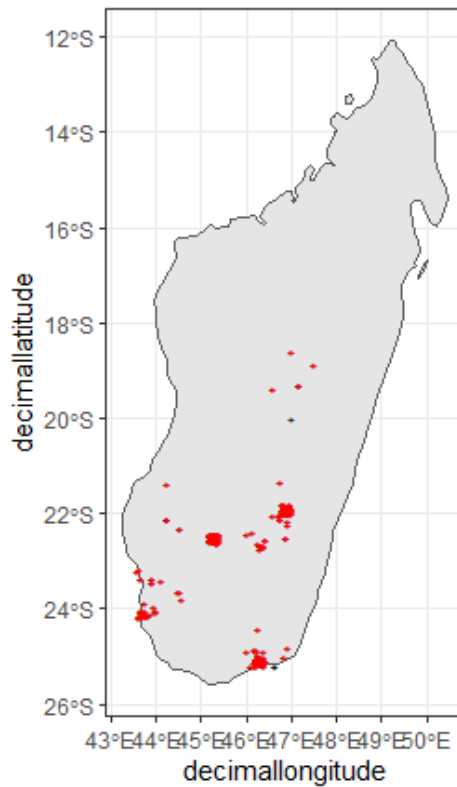
The black "+" markers indicate the occurrences of the raw dataset; whereas the red "+" markers indicate the occurrences of the cleaned dataset.

Zooming in to madagascar

```

ggplot() +
  geom_sf(data = country_map) +
  geom_point(data = gbif_download,
            aes(x = decimallongitude,
                y = decimallatitude),
            shape = "+",
            color = "black") +
  geom_point(data = clean_step2,
            aes(x = decimallongitude,
                y = decimallatitude),
            shape = "+",
            color = "red") +
  coord_sf(xlim = st_bbox(country_map)[c(1,3)],
           ylim = st_bbox(country_map)[c(2,4)]) +
  theme_bw()

```



Data cleaning step 3

Removing records with coordinate uncertainty and precision issues


```

clean_step3 <- clean_step2 %>%
  filter(is.na(coordinateuncertaintyinmeters) |
         coordinateuncertaintyinmeters < 10000,
         is.na(coordinateprecision) |
         coordinateprecision > 0.01)

print(paste0(nrow(gbif_download)-nrow(clean_step3), " records deleted; ",
            nrow(clean_step3), " records remaining." ))

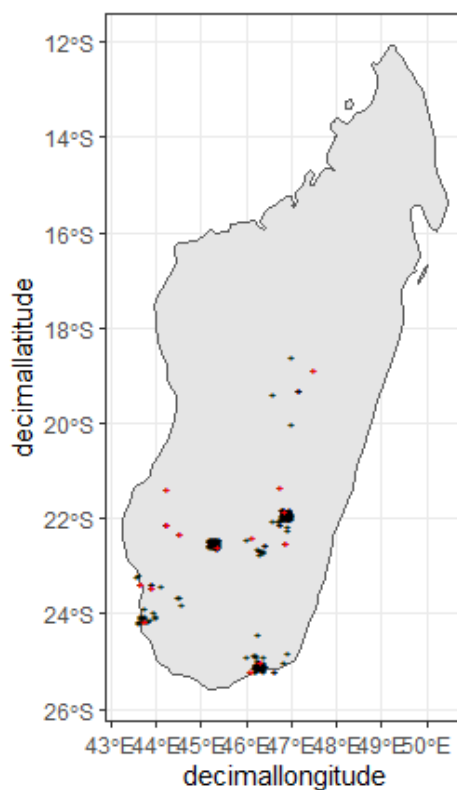
```

Plotting raw records vs. cleaned records (step 3)

```

ggplot() +
  geom_sf(data = country_map) +
  geom_point(data = gbif_download,
            aes(x = decimallongitude,
                y = decimallatitude),
            shape = "+",
            color = "black") +
  geom_point(data = clean_step3,
            aes(x = decimallongitude,
                y = decimallatitude),
            shape = "+",
            color = "red") +
  coord_sf(xlim = st_bbox(country_map)[c(1,3)],
           ylim = st_bbox(country_map)[c(2,4)]) +
  theme_bw()

```



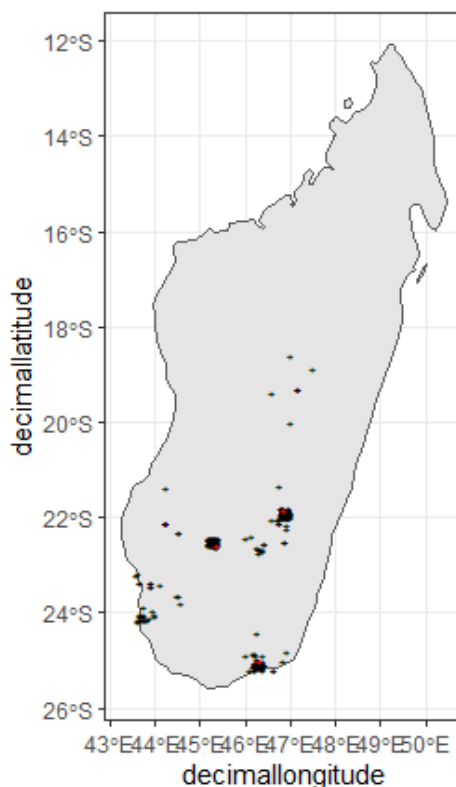
Oh no! we only have 14 records left.

Data cleaning step 4

Further removing records with temporal range outside that of our predictor variables

```
clean_step4 <- clean_step3 %>%  
  filter(year >= 1955) # WorldClim temporal range is 1970 to 2000s tho  
  print(paste0(nrow(gbif_download)-nrow(clean_step3), " records deleted; ",  
              nrow(clean_step4), " records remaining." ))
```

```
ggplot() +  
  geom_sf(data = country_map) +  
  geom_point(data = gbif_download,  
            aes(x = decimallongitude,  
                y = decimallatitude),  
            shape = "+",  
            color = "black") +  
  geom_point(data = clean_step4,  
            aes(x = decimallongitude,  
                y = decimallatitude),  
            shape = "+",  
            color = "red") +  
  coord_sf(xlim = st_bbox(country_map)[c(1,3)],  
           ylim = st_bbox(country_map)[c(2,4)]) +  
  theme_bw()
```

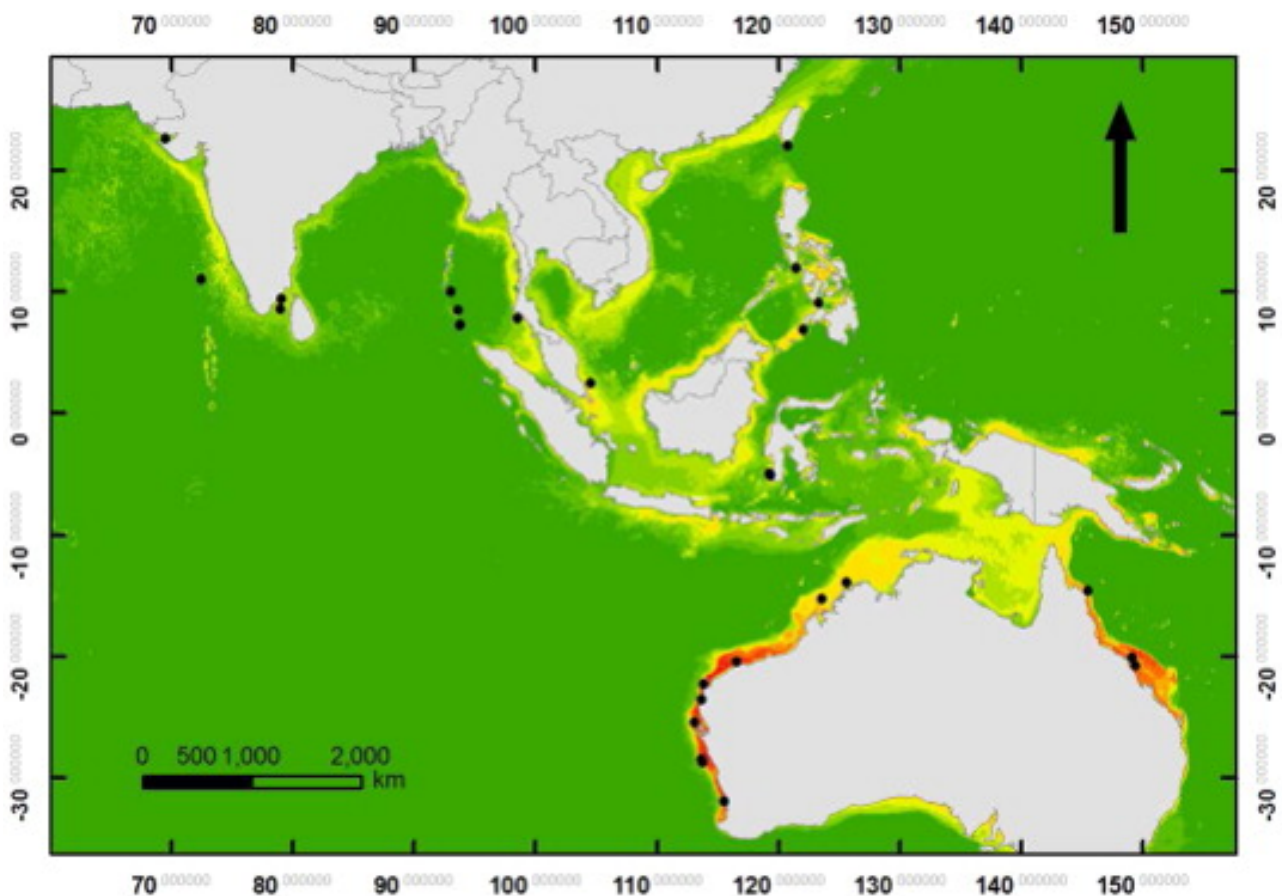


Oh nooooo there are just three records left! You may not have enough data points for what you want to do. You can always go back to your pipeline and refine.

Ecological Niche Models



This module will introduce you to the basic concepts and terminology associated with developing your own ecological niche models. By the end of the module, you should be able to understand what it is you are modelling and the different approaches available for developing your model, what kind of data you will need to model ecological niches, and how you can assess the quality of the models you produce. You will finish with an introduction on how to you can project your models into novel environments.



Carlos-Júnior LA, Barbosa NP, Moulton TP, Creed JC. Ecological Niche Model used to examine the distribution of an invasive, non-indigenous coral. *Mar Environ Res.* 2015 Feb;103:115-24. doi: 10.1016/j.marenvres.2014.10.004.

What is an ecological niche model?

An ecological niche model is an equation or set of equations that describe the ecological niche of a species. Niche modeling essentially does quantitatively, what natural historians do qualitatively, predicting suitable habitat for a species in geographic space. Models identify the characteristics of where a species has been found or observed and use that information to predict where else might be suitable habitat. Modelling approaches allow for the inference of a species without the need for experimental manipulation that may be too costly, logistically impossible and/or unethical. These

approaches also allow us to use existing data that we may have collected or observed ourselves, harvested from literature or obtained from data aggregators such as GBIF.



In this video (07:14), you will review the different concepts of a niche.

▶ <https://vimeo.com/662201339> (Vimeo video)

Uses

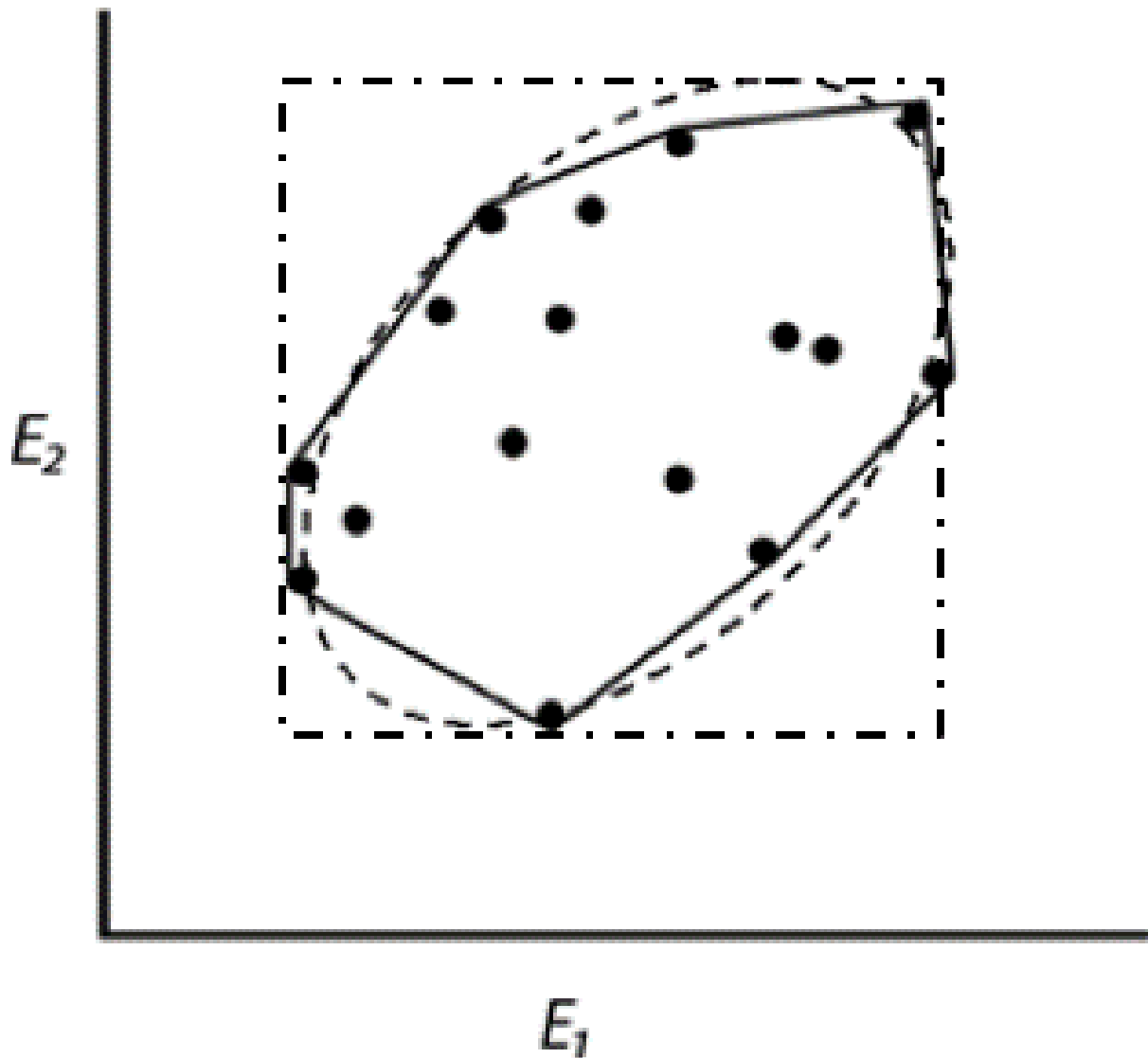
Beyond defining the limits of a species niche, ecological niche models have a number of practical uses including:

- The designation of protected areas - comparisons of protected area coverage and niche models of interest species can identify gaps in protected areas coverage that merit increased protection. By stacking niche models for different species, we can also identify areas where biodiversity is concentrated and maximise the protection of as many species as possible e.g. Leathwick et al., 2008
- Invasive species management - We can build niche models of known occurrences of an invasive species within its native range and project this information into regions of potential invasion to determine the potential invasive threat of the species e.g. Carlos-Júnior et al., 2015.
- Climate change mitigation - using models to define the characteristics of a species' niche, we can then predict the availability of suitable habitat under different climate scenarios and thus infer the impacts of climate change e.g. Aguilar et al. 2015

Commonly used algorithms

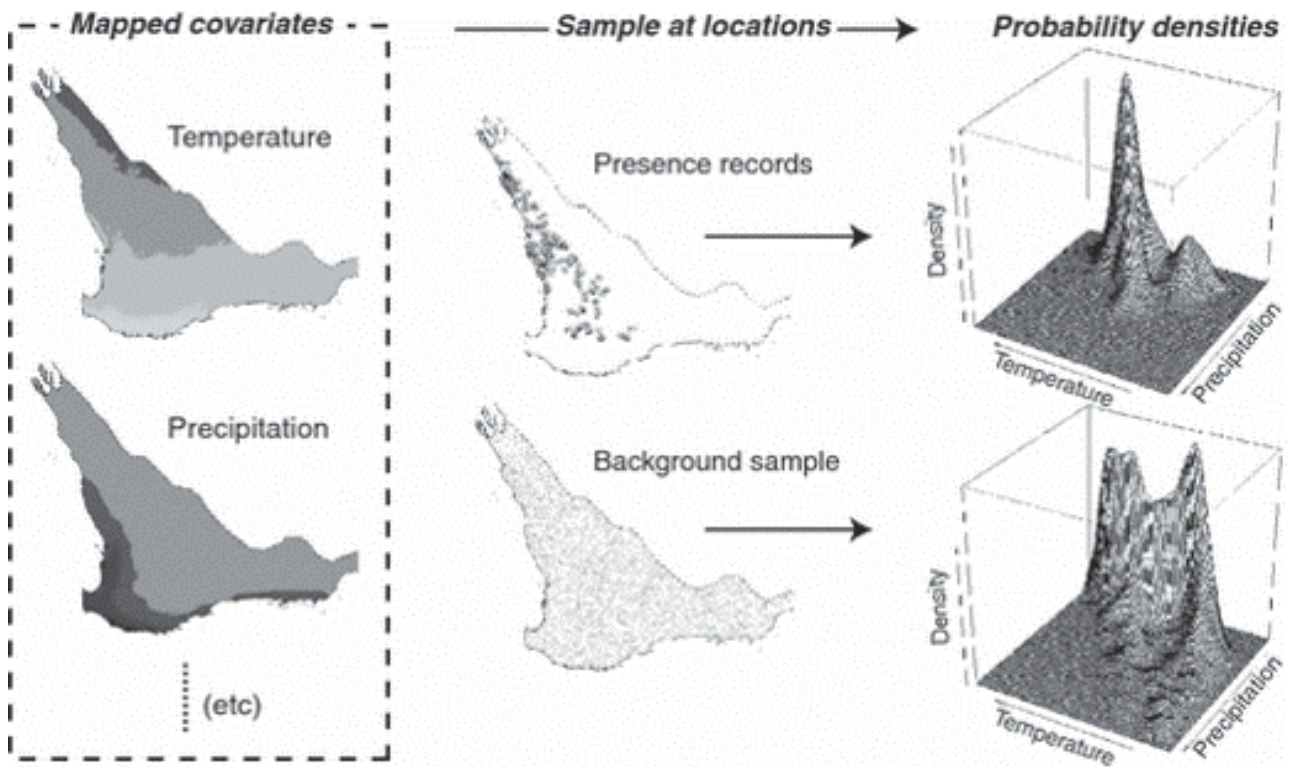
Algorithms to develop ecological niche models can be divided into three categories: presence-only, presence-background and presence-absence.

Presence-only algorithms focus solely on the environmental values linked to each occurrence record for calibration and create environmental envelopes, which are ellipsoids, squares, or convex-hull that surround the occurrences in an environmental space. These models are not very predictive Calibration of these modes is insensitive to changes in the extent of the study area. Commonly used algorithms include Bioclim and NicheA.



Presence-absence algorithms need a set of localities where the organism occurs (i.e., presence) and a set of localities where the organisms does not occur (i.e., absence). Presence-absence models are calibrated by comparing environmental conditions where the organism is present vs. where it is absent and are generally useful to reconstruct the distribution at fine scale and short periods, resulting in the need of accurate localities and high-resolution environmental variables. These models, however, have limited capacities to be projected to different areas or periods, instead, their signals are space and time specific. Many algorithms are available including regression (e.g., Generalized Linear Models and Generalized Additive Models) and classification (e.g., Boosted Regression Trees, Random Forest, and Support Vector Machines) algorithms, with protocols described in detail elsewhere

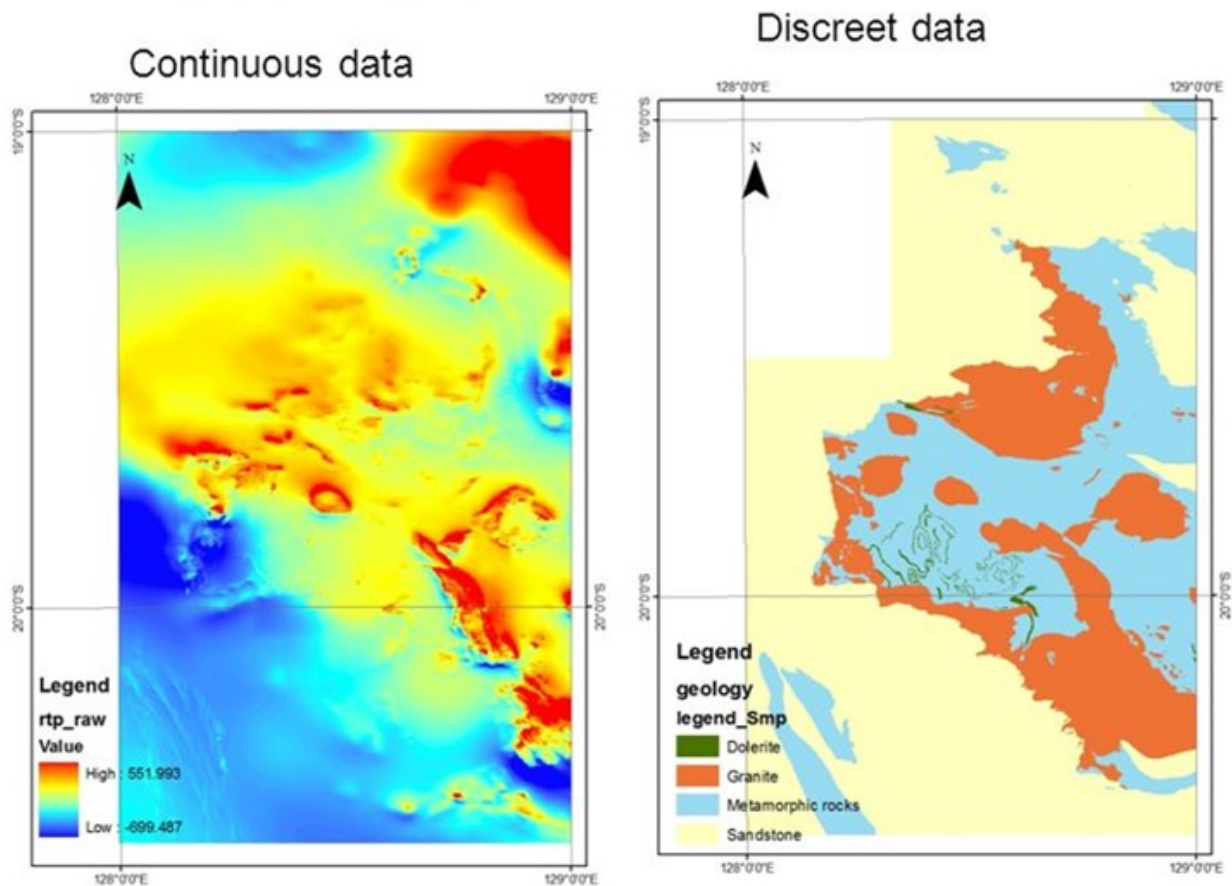
Absence data is limited in availability and is of questionable quality, as it is difficult to confirm the absence of a species from a particular area. To overcome the issues, presence-background models simulate absence data by generating random points (i.e., fake absence data) across the study area to be able to use presence-absence algorithms. Because the background corresponds to the study area, calibration of these algorithms is highly sensitive to variations in the extent of the study area extent selected. A popular ecological niche modeling algorithm using this approach is Maxent, and will focus on this algorithm throughout this course.



Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E. and Yates, C.J. (2011), A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17: 43-57. <https://doi.org/10.1111/j.1472-4642.2010.00725.x>

Environmental variables

Environmental variables, also known as environmental data, explanatory variables, bioclimatic data or covariates are anything that can be summarized by a raster (gridded dataset). These variables are used to characterize the niche of a species. The data can be either continuous or categorical (i.e. data expressed as vectors), direct measurements or derived products, static or dynamic or terrestrial, aquatic or atmospheric.



The tables below give examples of how these data can be classified.

Continuous		Categorical	
Elevation, bathymetry		Geology, Ecosystem	
Direct Measurement		Derived Product	
Remotely sensed data (raw), weather station data		climatology data, GCMs, derived remotely sensed data	
Static		Dynamic	
Altitude, bathymetry, slope, aspect, soil characteristics		temperature, precipitation, sea surface height	
Terrestrial	Aquatic	Freshwater	Atmospheric
Climate, terrain, vegetation/land cover, soil	Sea surface temperature, bathymetry, pH, salinity	Flow rates, accumulation, temperature	Wind (UV), radiation

Common sources of data

- WorldClim (Terrestrial)
- EarthEnv (Terrestrial and Freshwater)

- Bio-Oracle (Marine)
- National Geophysical Data Center (Terrestrial and Marine)
- National Snow and Ice Data Center (Terrestrial and Marine)
- World Ocean Atlas (Marine)
- Raw GCM outputs (ALL)

WorldClim is the most commonly-used climate data consisting of 19 derived bioclimatic variables (“BioClim”). These are typically divided into “quarters” (warmest quarter, driest quarter) and are related to seasonality. WorldClim also produces past and future modeled climate * Past: HCO, LGM, LIG * Future: to 2100 AD

But there are other sources e.g. <http://ecoclimate.org/> that stretch back farther. These are often not just climate models but also models of land position/amount. These past and future models differ in that past models are parameterized and testable using direct evidence, whereas future models are based on forcing variables (e.g. CO₂)

Selecting covariates (or environmental variables)

More environmental data isn’t always better. You want to balance to achieve a balance between the number of data points and the number of environmental variables so that you do not overfit your model. When selecting variables we want to be sure that:

- our variables are biologically relevant - they should reflect the species of study’s biology e.g. solar radiation may not be a relevant environmental variable for soil dwelling species
- our variables are not highly correlated - for instance, if we take the two variables: elevation and temperature. Temperature is not independent of elevation so we may want to remove one of these variables. In this instance, elevation would be preferably removed as it is more accurately measured.
- we do not use all 19 Bioclim variables

Importantly, spatio-temporal resolution and covariate data extent should align with:

- the limitations of other input data (e.g., available usable occurrence data)
- the scope of the base question(s)/hypotheses

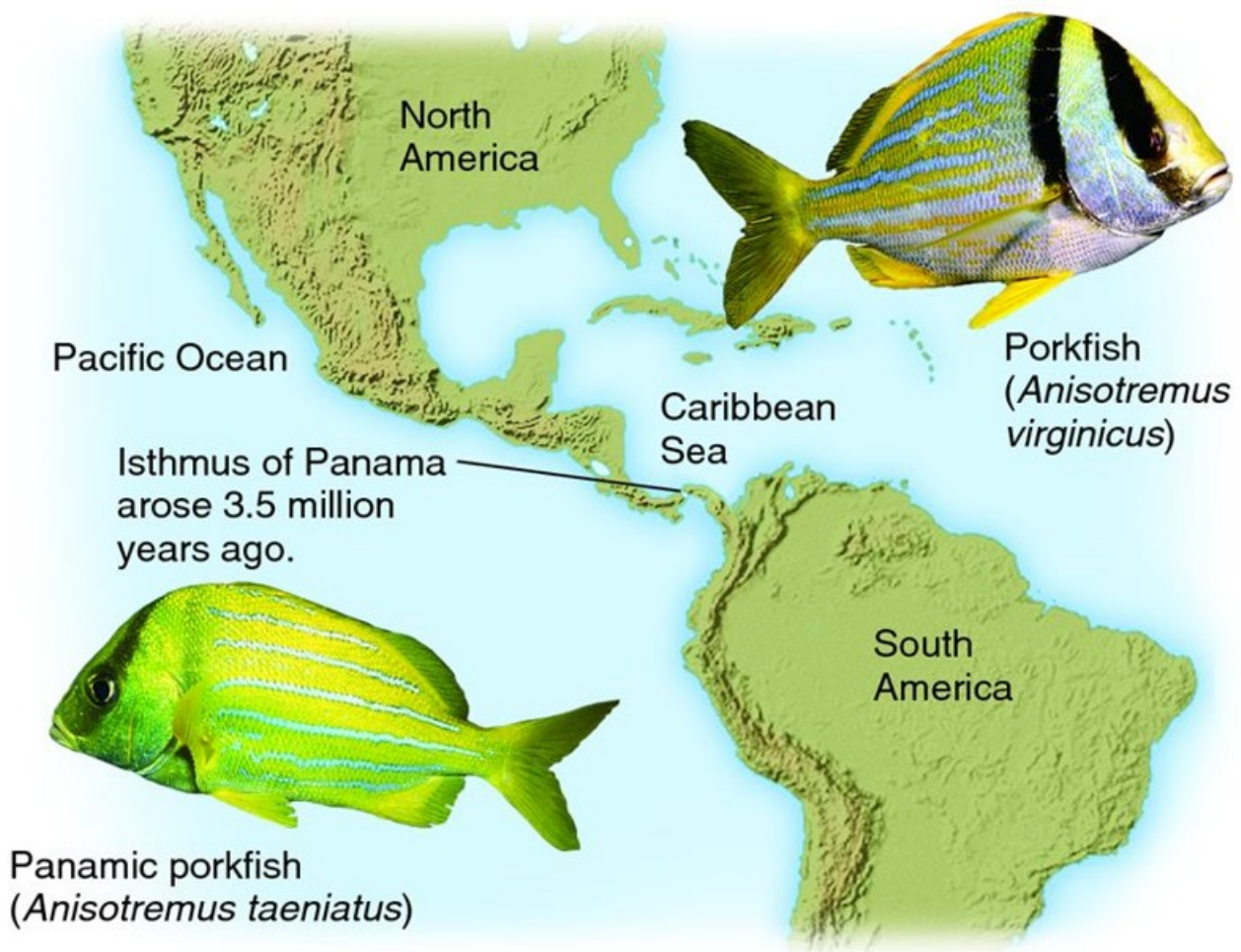
For example, if your environmental data have a spatial resolution of 10 Arc Minutes and a temporal resolution between 1955 and 2006, then the temporal and spatial resolution of the GBIF-mediated data you are going to use should correspond to those resolutions.

Training regions

Training regions (or study areas) are the areas from which model algorithms sample the background for model inference. In the case of presence-background models such as Maxent, this will be the area from which the model will randomly pick pseudoabsences that are used for calibrating the model. The training area can be thought of as the areas where the species could potentially experience environmental conditions. The species may not actually occur there, but it is possible

that the species can reach those areas. Points to consider when delimiting your training regions are:

- Where did the species originate?
- How far can the species disperse?
- Are there any biogeographic barriers that would prevent the dispersal of the species?
- it should not be a rectangle
- it should not correspond to political boundaries
- it should not be a coarse range delimitation (e.g. range map)
- bigger is not better



In the above example, the isthmus of Panama acts as effective barrier to the isolation of the Panamic porkfish to the Pacific and the Porkfish to the Caribbean. Training regions for each species would not contain areas on the opposite side of the Isthmus from where the species was found.

Interpretation and Post-Processing of Niche Models

You are now ready to build your model and this means deciding on the level of complexity of your model. This is done through two key factors: feature classes and the regularization multiplier.

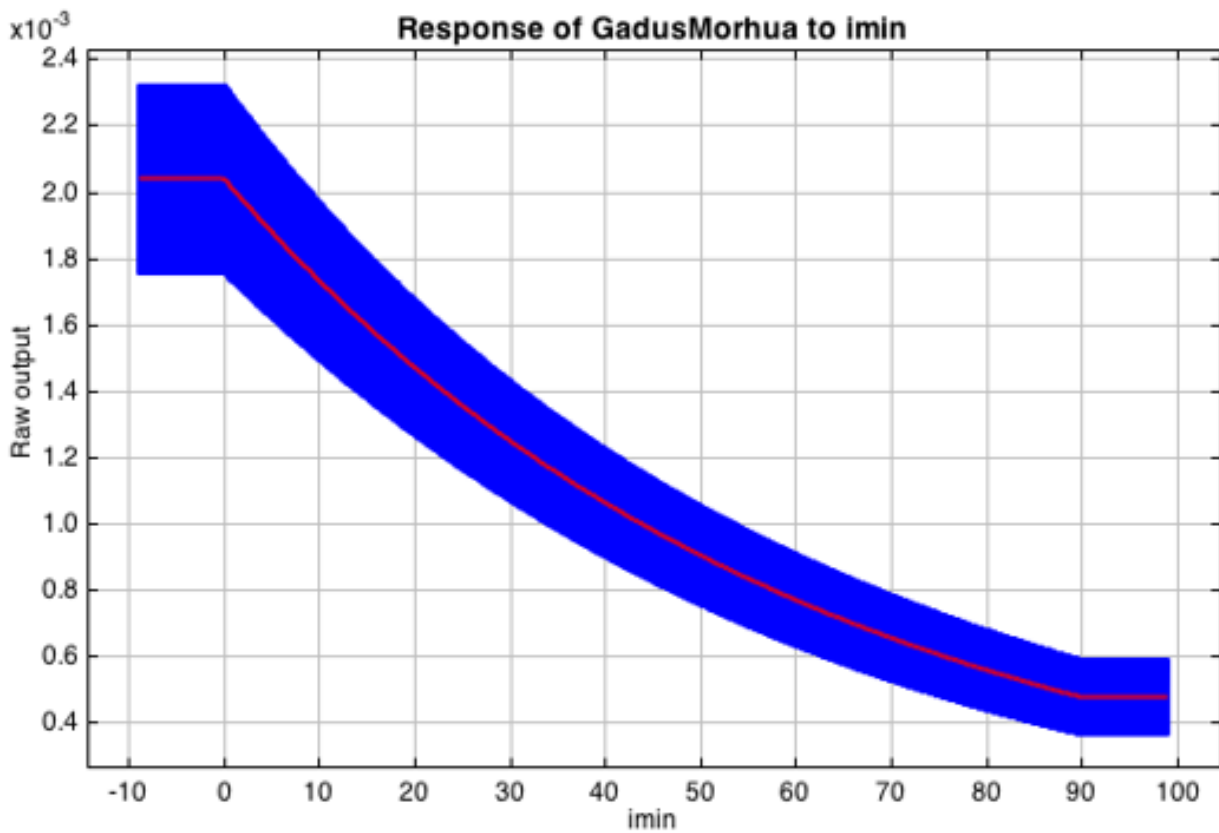
Feature classes determine the shape of available modeled relationships in environmental space and the more feature classes chosen, the higher the potential for model complexity. The regularization multiplier penalizes complexity to a greater degree, with higher values leading to simpler models with fewer variables. For these reasons, evaluating model performance and estimating optimal model complexity constitute important elements of a niche/distributional modeling for examples simultaneously varying the feature classes allowed and the regularization multipliers applied to each of them. Phillips, S.J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*. 31: 161-175.

Model Evaluation

You will have to assess the model's precision and significance — that is, whether the model can correctly predict independent presence (or absence) data and whether the model prediction is better than null expectations. Outputs for your model will include variable response curves and a number of statistics that can be used for assessing the performance of your model.

Variable Response Curves

Variable response curves are model outputs that describe how well your model has characterised how the species responds to the variable. Approximately normal curves may indicate better estimates of the fundamental niche of the species e.g.



Curves that deviate from normal distributions or are flat, may indicate that the variable may not be a good estimator of a species's fundamental niche. However, some variables such as ice concentrations, the lower curve in the diagram above, do not work like that - very few species can live enclosed in ice!

Statistics

In the ideal modeling scenario... You would seek to identify the ideal model calibration for your data and modeling intent, by comparing:

- multiple calibration scenarios for an individual algorithm and
- the best model calibration scenario across multiple algorithms

In the use cases, where you will be dipping your toes into the major theoretical concepts underpinning ENM/SDM, you'll be looking at only 1 algorithm.

Many options exist for evaluating model calibration scenarios.

Common and accepted approaches are:

- Akaike Information Criterion (AIC) - AIC is a log likelihood based evaluation metric, commonly used within regression methods. It compares and identifies the best model calibration scenario for an individual statistical algorithm. It balances model fit with model complexity but can NOT be used to compare between different algorithms. We can evaluate the performance of a model i.e. "which model performed better" by choosing the model with the lowest AIC. However, when AICs are only within 2 points of each other, these do not differ significantly and you will need to look at other factors (e.g., variable contribution through variable response curves) that may suggest which (if any) of the equivalent models is more ideal
- Omission Rate (OR) - compares model performance across algorithms. It is a method of evaluating a model's ability to accurately predict to test data (typically after applying a threshold). When $OR = 0$, then no presences were predicted as absent and the model has performed well.

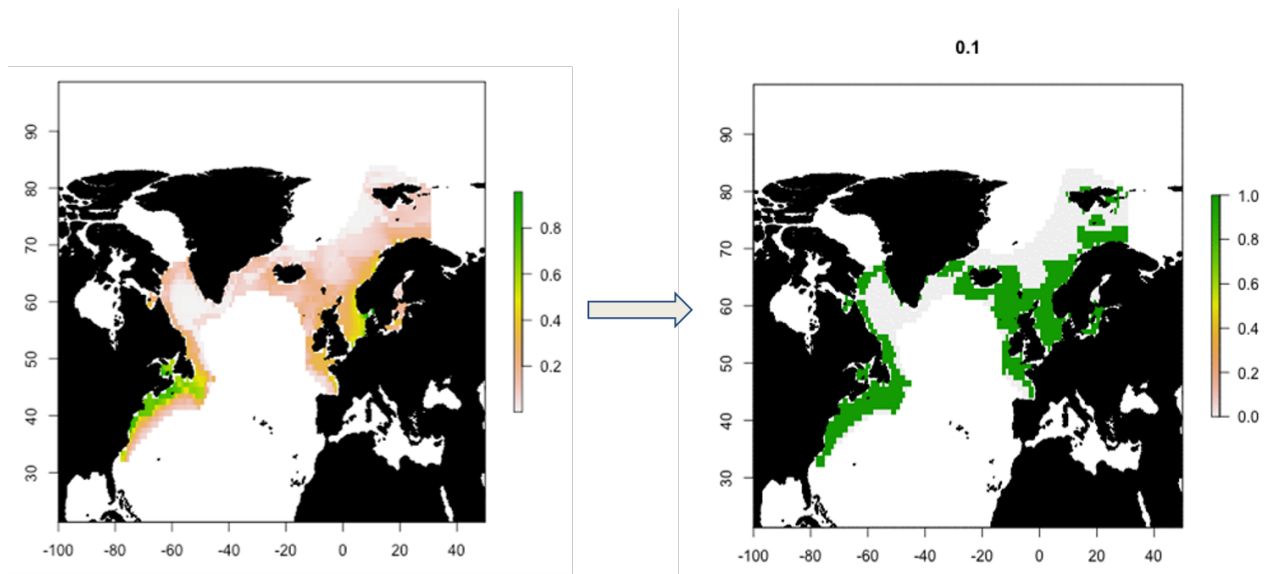
Thresholding a Niche Model

Thresholding is the process by which we convert the continuous (raw) output, or continuous suitability surface, from a statistical model to a binary output. The binary output is generally interpreted as areas that are suitable/not suitable for the species. Models are rarely perfect and it is likely that they will predict species as being present where they are not actually present (commission errors) and, conversely, absent where they actually occur (omission errors). When we threshold out model we want to decide on a threshold at which we are minimising both commission and omission errors. If we have threshold value of 100 then all areas are suitable for the species and we will have a high number of commission errors and the number of omission errors will approach 0.

	Species is present	Species is absent
Model predicts species as present	Accurate	Type 1 Error (commission)
Model predicts species as absent	Type 2 Error (omission)	Accurate

We choose the "threshold" value that determines a presence versus an absence of the species using the: - Minimum Training Presence (MTP) - this threshold assumes that the least suitable habitat at which the species is known to occur is the minimum suitability value for the species - $MTP + user-$

selected error rate (e.g., E=5%, E=10%) - a user-selected threshold that omits all regions with habitat suitability lower than the suitability values for the lowest 5% or 10% of occurrence records. It assumes that the percentage of occurrence records in the least suitable habitat do not occur in regions that are representative of the species overall habitat, and thus should be omitted. This threshold omits a greater region than the MTP.



Precise method by which you do this depends on the quality of the data that you used to build the model.

Projecting a Niche Model

You project a niche model when you map your model onto the training region to find additional suitable habitat. You can also map your model into the past or the future or into novel environments. You are asking, where can the species persist?

Projecting to your training region is the most common and simplest form. However, you can also project into different contemporaneous geographies too, for example:

- target sampling in undersurveyed regions for rare organisms e.g. de Siqueira et al. 2009
- predicting the existence of sister species e.g. Owens et al. 2013
- predicting the invasive potential of introduced species.

We can also project into the past and the future, for example: * to hindcast distributions in the case of determining paleodistributions of modern taxa for identifying refugia e.g. Peterson and Nyári, 2007 * to forecast species distributions to identify range shifts due to climate change e.g. Wang et al., 2016.

The Big Caveat

Models are built using a specific set of occurrence data and environmental data and we do not know how our model will behave in new environments. Transferring a model across space and/or time may lead to extrapolation if the projected environments are novel relative to training environments. Model algorithms have three strategies for dealing with extrapolation of response curves into

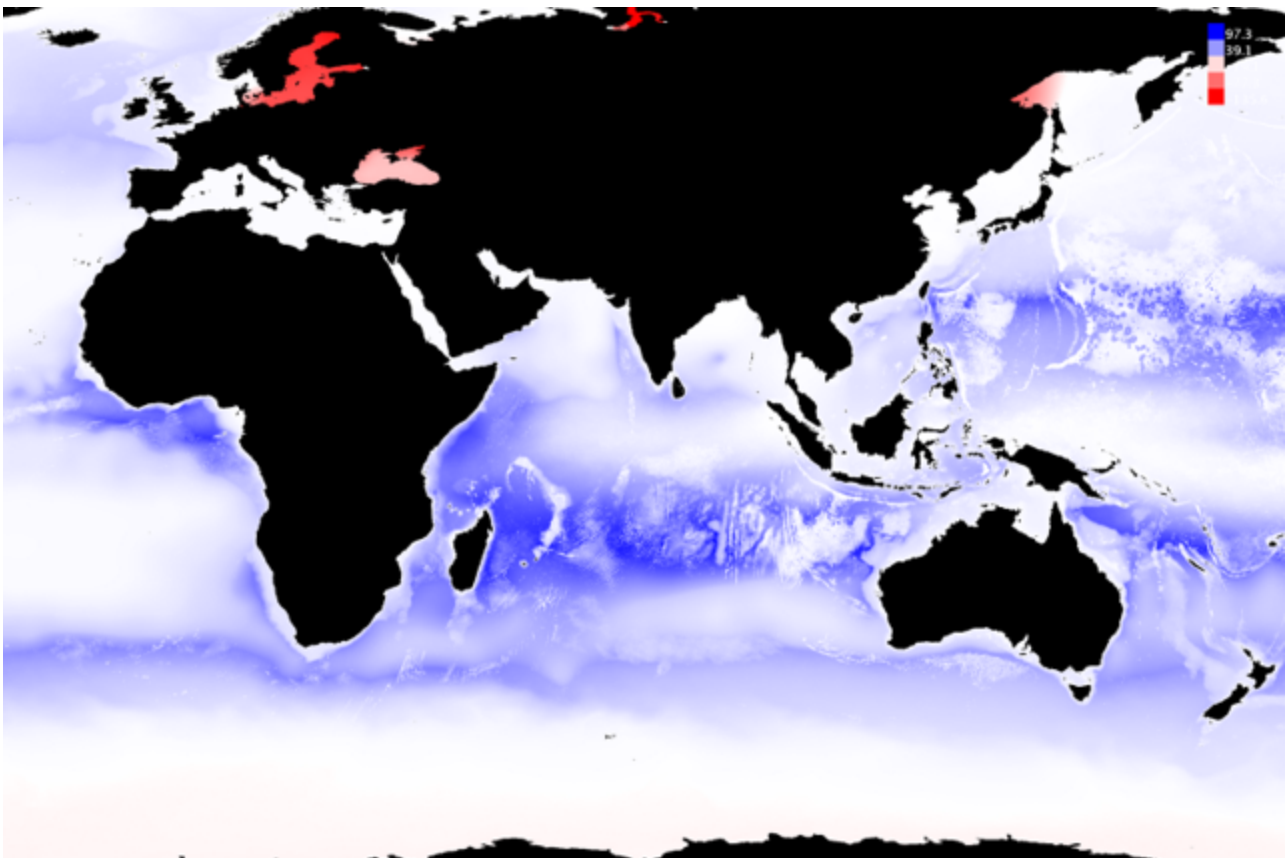
environmental conditions different than those existing in the region of model calibration, they can:

Truncate - designate all conditions outside of the calibration data range as unsuitable and thus not project beyond the training region
Clamp - use the marginal values in the calibration area as the prediction for more extreme conditions in transfer areas thus potentially under predicting the full extent of the projected niche
Extrapolate - extend the response curve based on trends obtained from calibration conditions or assumptions about the niche

It is left to the user whether they want their model to clamp or not.

Projection Uncertainty

MESS: Multivariate Environmental Suitability Surface is a measure of the similarity between the new environments and those in the training sample. They allow modelers to identify areas of model extrapolation in novel environments. It measures the similarity of any given point to a reference set of points, with respect to the chosen predictor variables. It reports the closeness of the point to the distribution of reference points, gives negative values for dissimilar points and maps these values across the whole prediction region. The map below is an example of a MESS with areas in red on the map highlighting areas of model extrapolation where into potentially unsuitable environments for the species.



Use Case - Modelling Species Distributions Under Climate Change

This is a practice use case for the ecological niche modeling module developed by Dr. Hannah Owens, University of Copenhagen. For this use case, you will use the modular, R-based platform Wallace link <https://wallacecomod.github.io/> for reproducible modeling of species niches and distributions.

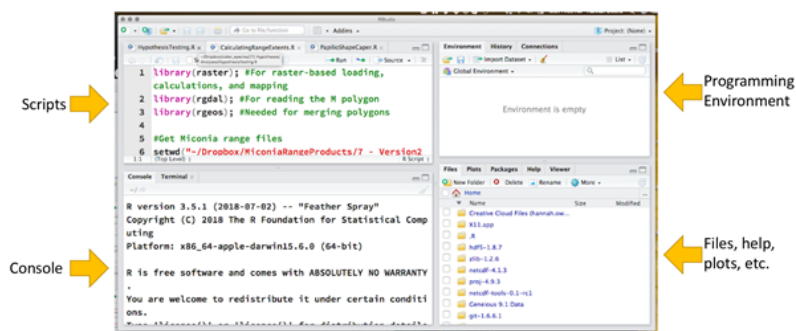
The platform is powered by a variety of R packages including dismo and enmeval allows and lowers the threshold for access to basic modeling functions through a user friendly graphical user interface (GUI). It is a great platform for getting you started in in niche modeling and will allow to learn the basics of modeling, explore your data and generate a backbone R script that will allow for the repeatability of your analyses. What it will not be able to do is generate final, publishable models but it should give you a good grasp of the principles of niche modeling.

Exercise 1 - Starting Wallace

First thing's first. We'll start by launching Wallace and giving you a quick overview of the steps you will go through to generate a niche model.

- Launch RStudio.

RStudio is a helpful platform for writing and executing R code. You can write and save scripts, execute commands, and keep track of datafiles. We don't have time to give you an exhaustive overview of all the RStudio features, but if you are curious, check out the link at the end of this exercise. For our purposes, the important thing for you to know is that there are four panes in RStudio, as follows (note that depending on your operating system and the version of RStudio, it may not appear exactly identical):



- Launch Wallace

a) If you have installed Wallace, proceed to step b. If you have not installed Wallace, type the following in the console window of RStudio and hit "Enter":

```
install.packages("wallace")
```

```
Untitled10* x Espace.R* x GAM
Source on Save
1 # install and run wallace
2 install.packages("wallace")
3 library(wallace)
4
5 run_wallace()
6
```

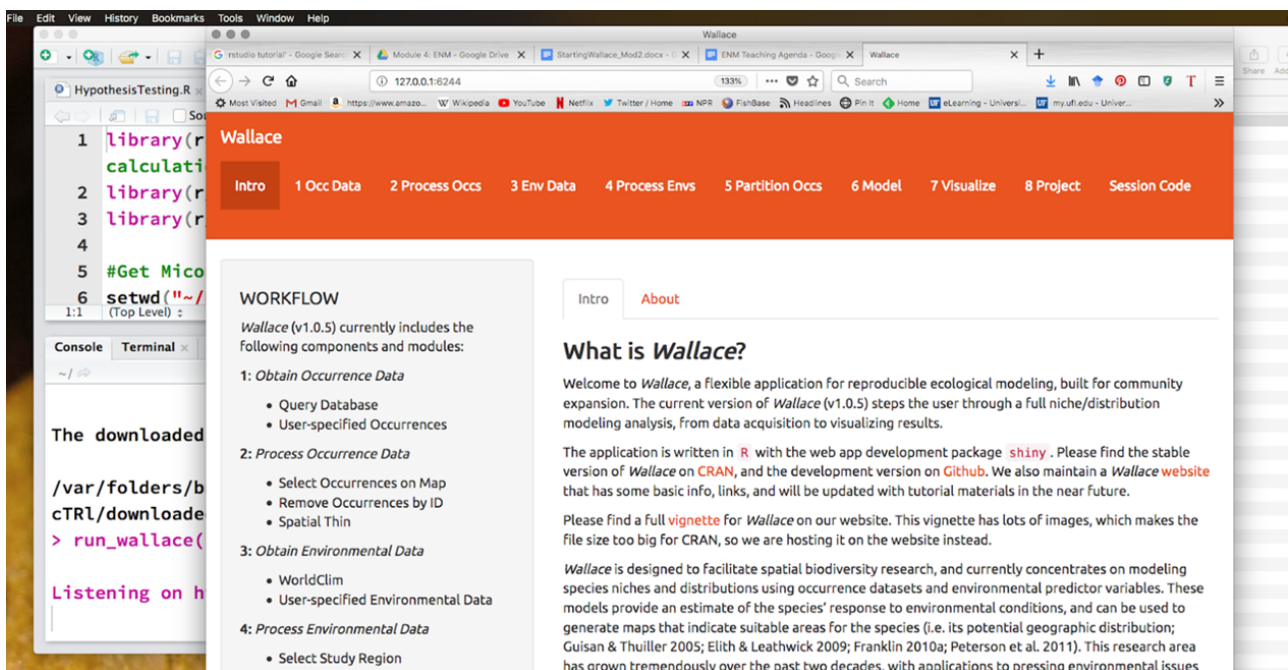
b) Type the following in console window of RStudio and hit "Enter":

```
library("wallace")
```

c) Type the following in console window of RStudio and hit "Enter":

```
run_wallace()
```

This should launch Wallace in an internet browser window.

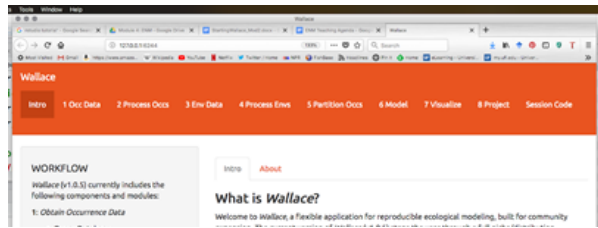


Along the top of the browser is a menu bar with each of the steps that go into generating a niche model. You can also see this as a workflow in the left-hand panel. We will follow the general outline of this workflow, but with some slight modification. This will be explained more as we work through generating our own niche models.

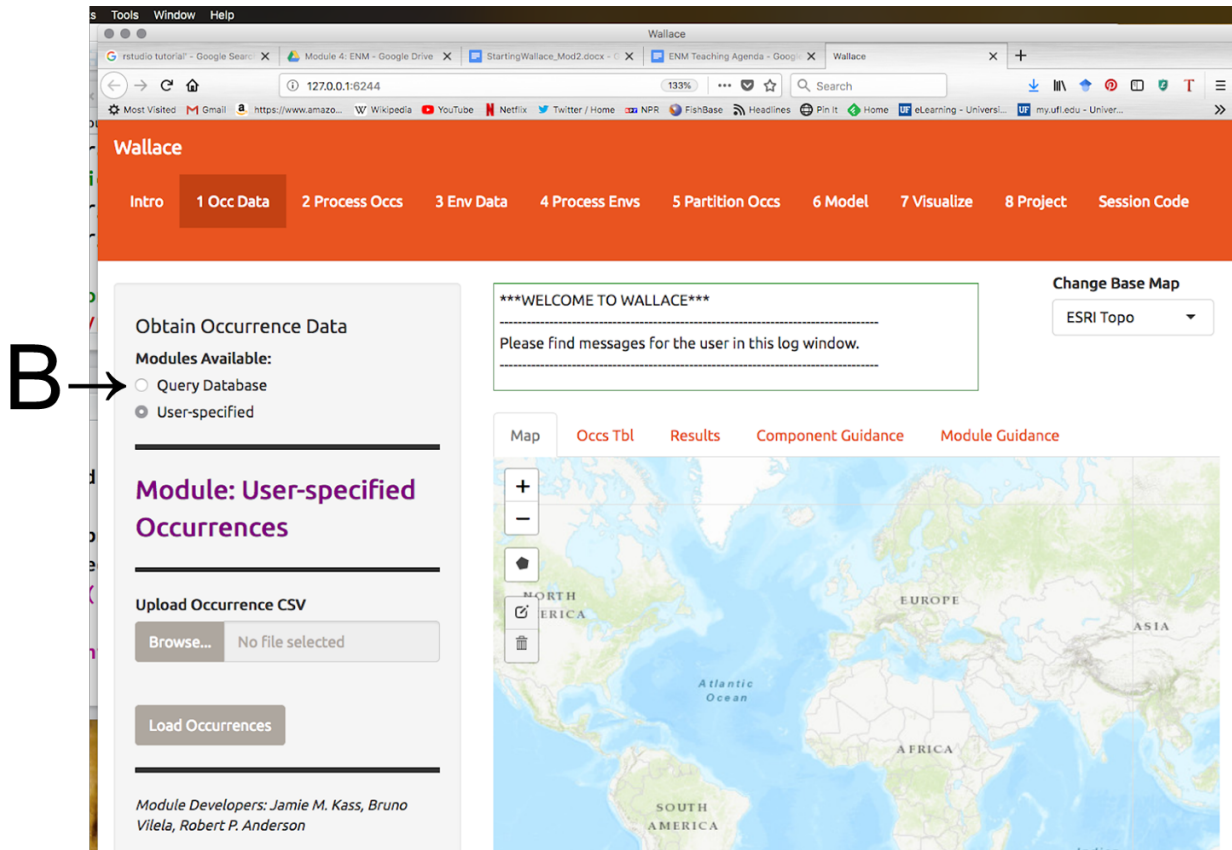
Exercise 2 - Loading occurrence points

There are two options for loading occurrence data in Wallace. The first option is to directly query several biodiversity databases including GBIF. This is good for generating quick and dirty niche models to explore data, but Wallace does not allow you to clean data in all the ways we have discussed in the data processing module. For this reason, you can also load in your own occurrence data that you have processed using your own processing pipeline or data that is not yet served through a database connected to Wallace. For this exercise, we will query GBIF directly using Wallace.

a) Click on “1 Occ Data” in the menu bar at the top of the Wallace window



b) In the panel on the left, select the “Query Database” radio button.



Notes on the Wallace interface:

- As we begin importing and processing data, a detailed record of what has been done will be recorded in the log window above the map.
- If you need a refresher on the background of a particular modeling step, you can click on the “Component Guidance” tab
- If you need help understanding the elements of a particular modeling step, you can click on the “Module Guidance”.

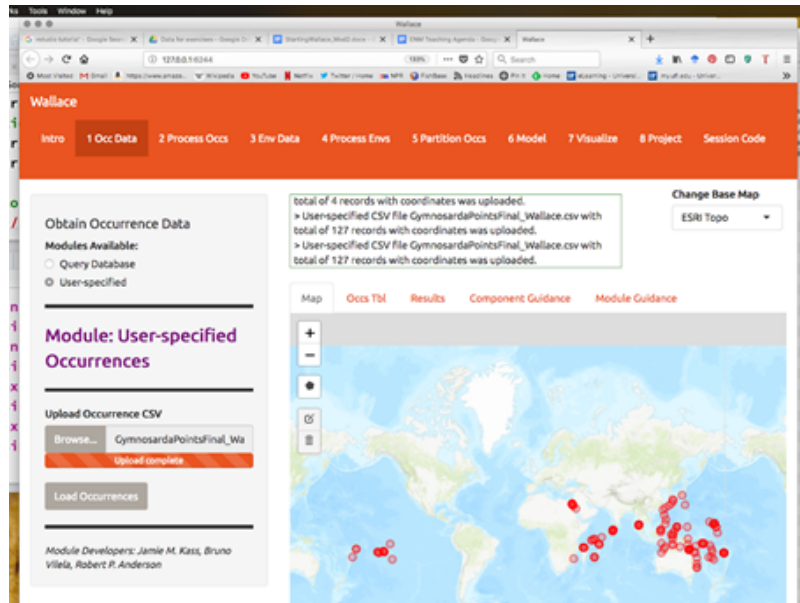
b) Select the “GBIF” radio button and enter a species name of your choice. I will be using *Protea cynaroides* for this example, in case you want to follow along exactly. I also increased the number of occurrences I am searching for to 10,000, because I want all the points.

Choose Database
 GBIF VertNet BISON

Enter species scientific name

Set maximum number of occurrences

c) Click "Query database". Your occurrence points should all show up as red dots on the map.



Note: You can interactively explore your points by clicking on them on the map (see below). You should see all the information associated with that record from the table you uploaded. This is helpful for verifying your occurrence points before progressing through the workflow.

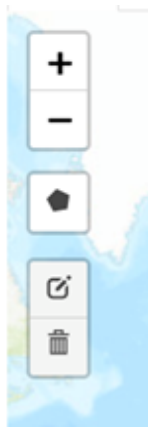
Second Note: You can inspect the point data as a spreadsheet under the "Occs Tbl" tab next to the "Map" tab.

Exercise 3 - Processing Occurrences

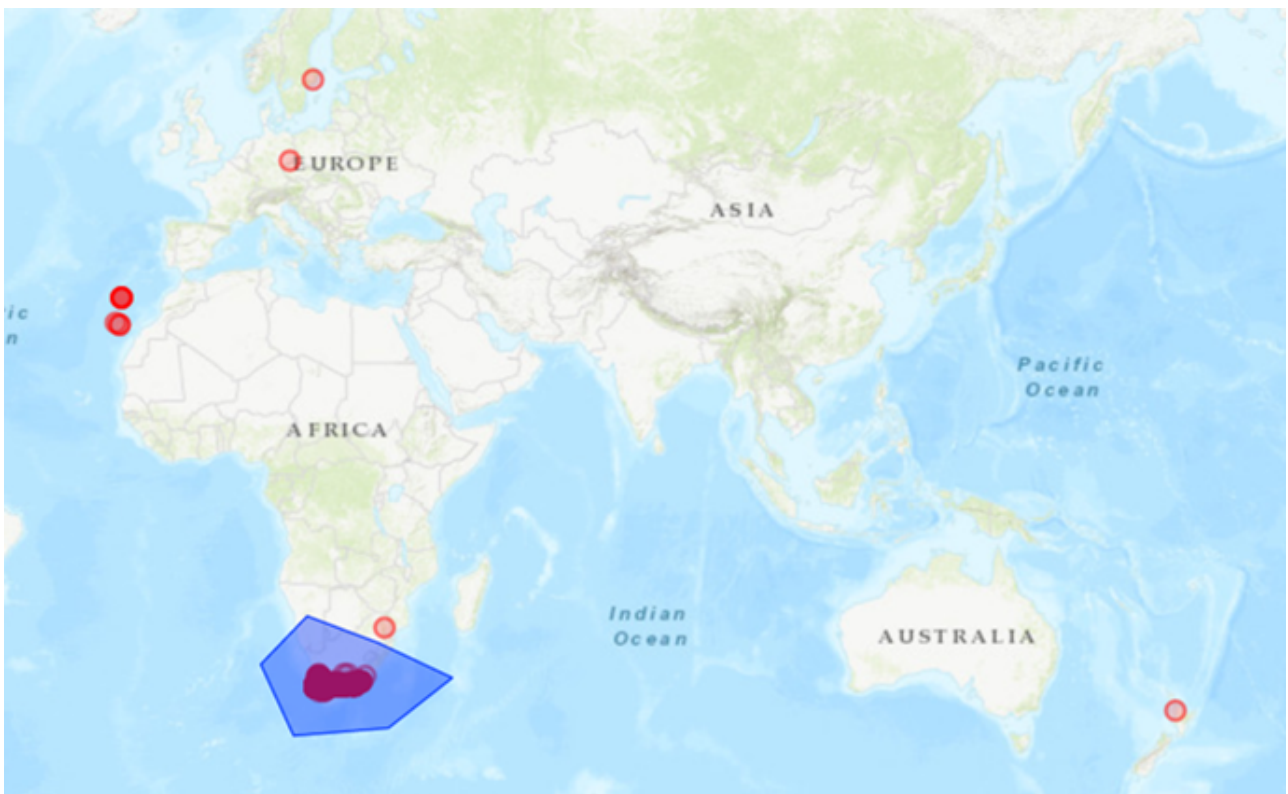
At this point, you may notice that there are some points that do not look correct when they are mapped. This could be due to a clerical error, a human-mitigated introduction, or a natural vagrancy. Whatever the reason, these points can cause errors in your model and must be removed.

a) Click on "2 Process Occs" in the menu bar at the top of the Wallace window.

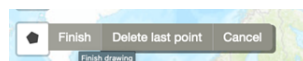
b) Select the "Select Occurrences on Map" radio button under "Modules Available". c) Click on the "Draw a polygon" button in the map window



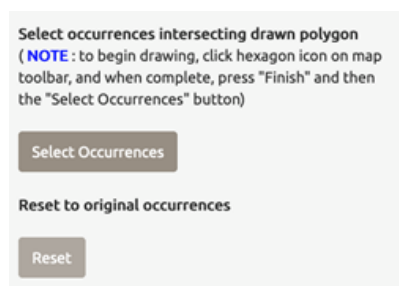
d) Draw a polygon around the points you want to keep. In my case, I only want to keep the *Protea cynaroides* occurrences in its native range in South Africa.



e) Click the "Draw Polygon" button again and click "Finish".



f) Click the "Select Occurrences" button. If you don't do this, all the points will be kept!



Note: For your own practical projects, you may also want to consider spatially thinning your data (this

can remove some risk of sampling bias effecting your model results). We are not doing this step now because it takes a fair bit of time for some datasets.

Exercise 4 - Loading environmental data in Wallace.

- Click on “3 Env Data” in the browser window in which Wallace is running.
- Select the “WorldClim Bioclims” radio button. Choose the 2.5 arcmin resolution (or whatever resolution you feel is most appropriate given your data) and check the “Specify variables to use in analysis?” box. Select the variables you think will be most informative for your model. Under “Module Guidance” there is an explanation of what the different variables are. Never use all 12 BioClim variables. It leads to overfit models with low predictive power.



Module: WorldClim Bioclims
raster - Geographic Data Analysis and Modeling

Select WorldClim bioclimatic variable resolution
2.5 arcmin

Specify variables to use in analysis?

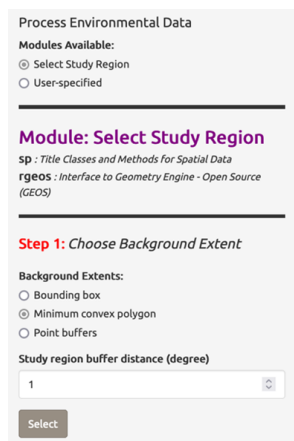
Select

bio1 bio2 bio3 bio4 bio5
 bio6 bio7 bio8 bio9 bio10
 bio11 bio12 bio13 bio14 bio15
 bio16 bio17 bio18 bio19

Using map center coordinates as reference for tile download.
Using map center 32.476, -39.104

Load Env Data

- Click the “Load Env Data” button. Your view should change to something similar to that shown below. The gray box will show metadata on the environmental data you have uploaded.



Process Environmental Data

Modules Available:

Select Study Region
 User-specified

Module: Select Study Region
sp : Title Classes and Methods for Spatial Data
rgeos : Interface to Geometry Engine - Open Source (GEOS)

Step 1: Choose Background Extent

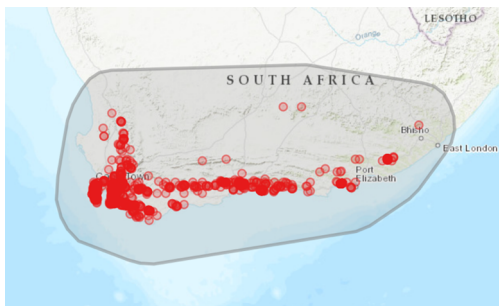
Background Extents:

Bounding box
 Minimum convex polygon
 Point buffers

Study region buffer distance (degree)

1

Select



- Sample background points.

The screenshot shows the Wallace software interface. At the top, it says "STEP 1: Upload polygon with field order: longitude, latitude (Lcsv)". Below this, there is a "Browse..." button and a text field containing "GymnosardusTrainingRegion.csv". An "Upload complete" button is visible. Underneath, there is a "Study region buffer distance (degree)" input field with the value "0" and a "Load" button. A horizontal line separates this from "STEP 2: Sample Background Points". Below this, it says "Mask predictor rasters by background extent and sample background points". There is a "No. of background points" input field with the value "10000" and a "Sample" button.

The number of background points shown (10,000) is fine. This is the number of points that will be sampled randomly from the training region you have uploaded. Values of predictor variables for these background points can then be compared to those at the occurrence points to improve model fit. Click "Sample" and be patient. This takes a little time.

Exercise 6 - Partitioning Occurrence Data

Ideally, you will have two completely independent occurrence datasets to determine the strength of the model's predictive ability. Unfortunately, this rarely reality. When no independent datasets exist, one solution is to partition your data into subsets we assume are independent of each other, then sequentially build a model on all the subsets but one and evaluate this model on the left-out subset. This is known as k-fold cross-validation (where k is the total number of subsets). After this sequential model- building step is complete, Wallace summarizes (averages) the statistics over all the partitions and builds a consensus model using all the data.

- a) Click on "5 Partition Occs" in the browser window in which Wallace is running.
- b) Select the "Spatial Partition" radio button.
 - From the "Options Available" dropdown menu, select "Checkerboard 1 (k = 2)".
 - Click "Partition". This may take a few minutes depending on the amount of occurrence data you have and the partition option selected.

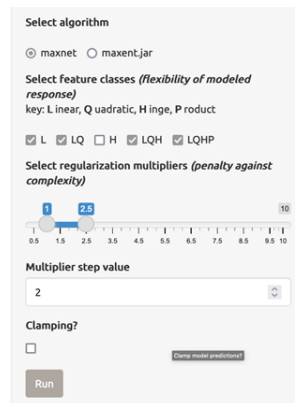
The screenshot shows the "Partition Occurrence Data" module in Wallace. It has two radio buttons: "Non-spatial Partition" (unselected) and "Spatial Partition" (selected). Below this, it says "Module: Spatial Partition" and "ENMeval - Automated Runs and Evaluations of Ecological Niche Models". Under "Options Available:", there is a dropdown menu with "Checkerboard 1 (k = 2)" selected. Below that, there is an "Aggregation Factor" input field with the value "2". At the bottom, there is a "Partition" button.

Exercise 7 - Calibrating Niche Models with Maxent

'Wallace' allows for very few opportunities to set the parameters of your models (as compared to using the Maxent GUI), but Wallace WILL run several model iterations with different parameter combinations and tell you which set fit the data best.

- a) Click on "6 Model" in the browser window in which Wallace is running.

b) Select the “Maxent” radio button at the top left. Under “Select algorithm”, select the “maxnet” radio button. “Maxnet” and “maxent.jar” use the same underlying math, but “maxnet” does not use Java. This means it runs more readily on a wider range of computer operating systems than “maxent.jar”; Maxent was developed in the early 2000s using Java so that it had a graphical user interface. Now Java often causes more problems than it solves.



c) Under “Feature classes” uncheck “Hinge”. Feature classes refer to the sorts of equations Maxent will use to try to model the data (linear equations, quadratic equations, and equations involving products). “Hinge” equations use two linear equations that “hinge” at a particular value of an explanatory variable, which isn’t a very natural response to an environmental variable.

d) Select regularization multipliers from 1 to 3. Set the “Multiplier step value” to 1. The regularization multiplier sets how closely our model fits the data that we have used. A smaller value than 1 will result in a more localized output distribution that is a closer fit to the presence records. Overfitting the model in this way may mean that it doesn’t generalize well to independent data. A larger multiplier will give a more spread out, less localized prediction. The multiplier step value sets the intervals at which regularization multiplier will be tested. So with multiplier values of 1-3 and a multiplier step value of 1, test models will be run for regularization multiplier values of 1, 2, and 3.

e) Press ‘Run’. Be patient, this process can take a few minutes.

f) When the process is complete, the ‘Results’ tab will open and display both the full model and partition evaluation statistics and the individual partition evaluation statistics. Remember, modeling algorithms are stochastic, so results displayed may be a little different each time you run the models.

Map Occs Tbl Results Component Guidance Module Guidance

Evaluation Lambdas

Full model and partition bin average evaluation statistics

	rm	fc	tune.args	auc.train	cbi.train	auc.diff.avg	auc.diff.sd	auc.val.avg	auc.v
12	3	LQHP	rm.3_fc.LQHP	0.911	0.985	0.005	0.001	0.905	
11	2	LQHP	rm.2_fc.LQHP	0.915	0.991	0.006	0.003	0.908	
9	3	LQH	rm.3_fc.LQH	0.903	0.966	0.004	0.002	0.896	
4	1	LQ	rm.1_fc.LQ	0.9	0.951	0.003	0.002	0.897	
5	2	LQ	rm.2_fc.LQ	0.899	0.946	0.003	0.001	0.896	
6	3	LQ	rm.3_fc.LQ	0.898	0.942	0.003	0	0.895	
8	2	LQH	rm.2_fc.LQH	0.906	0.967	0.006	0.001	0.898	
1	1	L	rm.1_fc.L	0.887	0.946	0.003	0.001	0.885	
2	2	L	rm.2_fc.L	0.886	0.939	0.003	0.001	0.884	
3	3	L	rm.3_fc.L	0.886	0.945	0.003	0.001	0.884	

Previous 1 2 Next

Exercise 8 - Model Evaluation and Selection

Wallace provides a fairly broad suite of evaluation metrics to use in determining which model to utilize. For our purposes, we will use AICc. Typically, the model with the lowest AICc score (or a delta AICc of 0) is considered to be the best model (balancing goodness-of-fit with simplicity). But, omission rate is also a common and effective method of evaluating binary predictions, so we will look at these as well.

a) Look at the “Full model and partition bin average evaluation statistics” table in the Results section (the top table).

b) Sort the AICc scores lowest to highest. Which model has the lowest AICc score? The name of the model tells you what the parameter settings are. RM = randomization multiplier, FC = feature class.

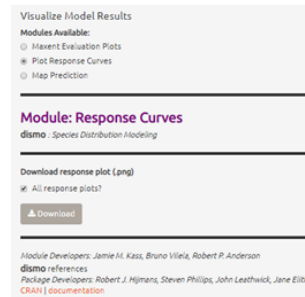
c) Now look at the “Individual partition bin evaluation statistics” table (the bottom results table). You’ll see that data have been evaluated using binning based on two threshold levels: the 10 percentile training (or.10p) and the minimum presence training thresholds (or.MTP). Which model has the lowest omission rate?

d) Based on this information, choose the model you think is the best fit. This will likely be a compromise—one model that outperforms the others on all evaluation metrics is quite rare. Use your best judgement, and ask for help if you’re stuck.

Exercise 9 - Visualizing Model Results

Now that we've seen the numbers, let's get an idea of what our niche models look like in terms of inferred response curves and geography. NOTE: Remember to click on the "Component Guidance" tab if you need a refresher overview on niche/distributional models and the "Module Guidance" tab if you need additional information about the occurrence data partitioning methods.

a) Click on "7 Visualize" in the browser window in which Wallace is running.



b) Under "Visualize Model Results" select the "Plot Response Curves" radio button.

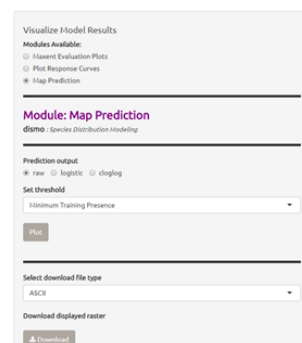
c) In the "Results" window to the right, you'll see a single plot for the first environmental variable. To view variable responses for the model you have decided is the best fit, select that model under "Current Model" at the far right side of the plot. To view a different response curve for another environmental variable, click on the "Current Env Variable" dropdown menu and select the variable you want to view. If you wish to view the response curves for all variables simultaneously, we will need to download the plots.

d) Save the response plots. Under "Download response plot (.png)", check the box next to "All response plots?" then Download. Open the plots to examine all response plots side-by-side. How do they look? Are they roughly bell-shaped, suggesting the model has completely characterized suitability of all the variables you used? Are the responses fairly smooth, or are they jagged, like the model is overfit?

Exercise 10 - Visualize model results in geographic space

a) Under "Visualize Model Results" select the "Map Prediction" radio button.

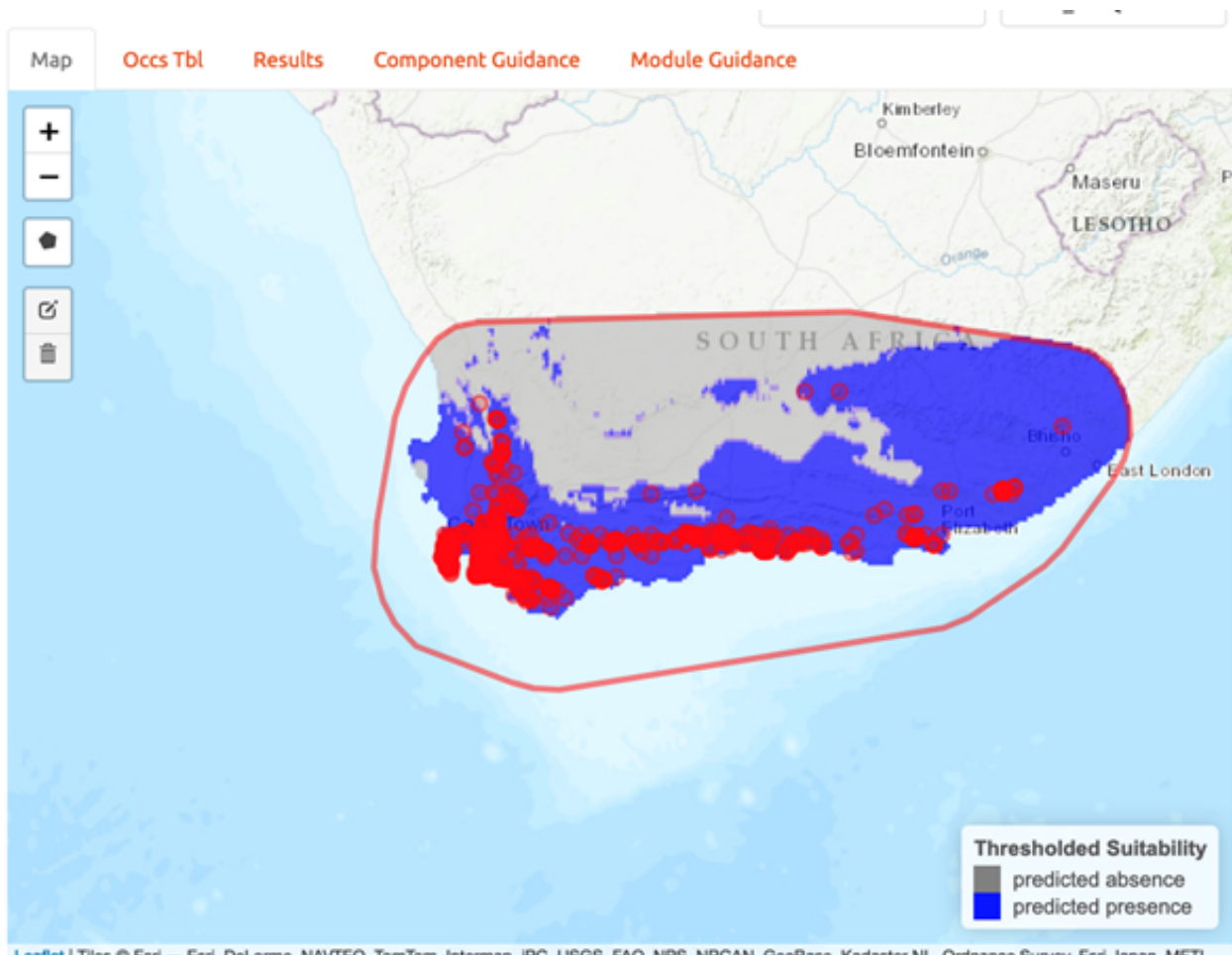
b) If you look to the right above your map, you'll see 3 drop down menus. Click on the "Current Model" dropdown menu and select the model that performed best according to your model evaluation statistics.



c) Under “Prediction Output” select the “raw” radio button.

e) From the “Set threshold” dropdown menu, choose the threshold (minimum training presence or 10 percentile training) that yielded the best omission rate accord to the model evaluation in Exercise 6.

f) Click on “Plot”. Your thresholded binary model results for the calibration/training region should appear in the display window with the extent of the training region denoted in red (an example is below).



g) Now, take a few minutes to explore the three alternate model projection options. That is, if your best evaluated model was LQHP_2 with a MTP threshold, then take a minute to visualize LQHP_2 with a 10 Percentile Training Threshold, LQHP_1 with a MTP, and LQHP_2 with a 10 Percentile Training Threshold. What similarities do you see across the visualizations? Are there major differences?

h) Save your model prediction. First, be sure to return all settings to reflect your selected model and threshold. Then, select “ASCII” from the “Select download file type” dropdown menu, and press “Download”. Save the file to your working project folder.

Exercise 11 - Niche model projection

REMEMBER: if you want more information on the background of model projection, click on the “Component Guidance” tab; if you need additional information about the model projection process, click on the “Module Guidance” tab.

a) Click on “8 Project” in the browser window in which Wallace is running.

b) Under “Modules Available”, select the “Project to New Time” radio button. Select “2070” under “New Time Period”, and choose your favorite global circulation model and RCP scenario. The higher the number of the RCP scenario, the more CO₂ in the simulated atmosphere.

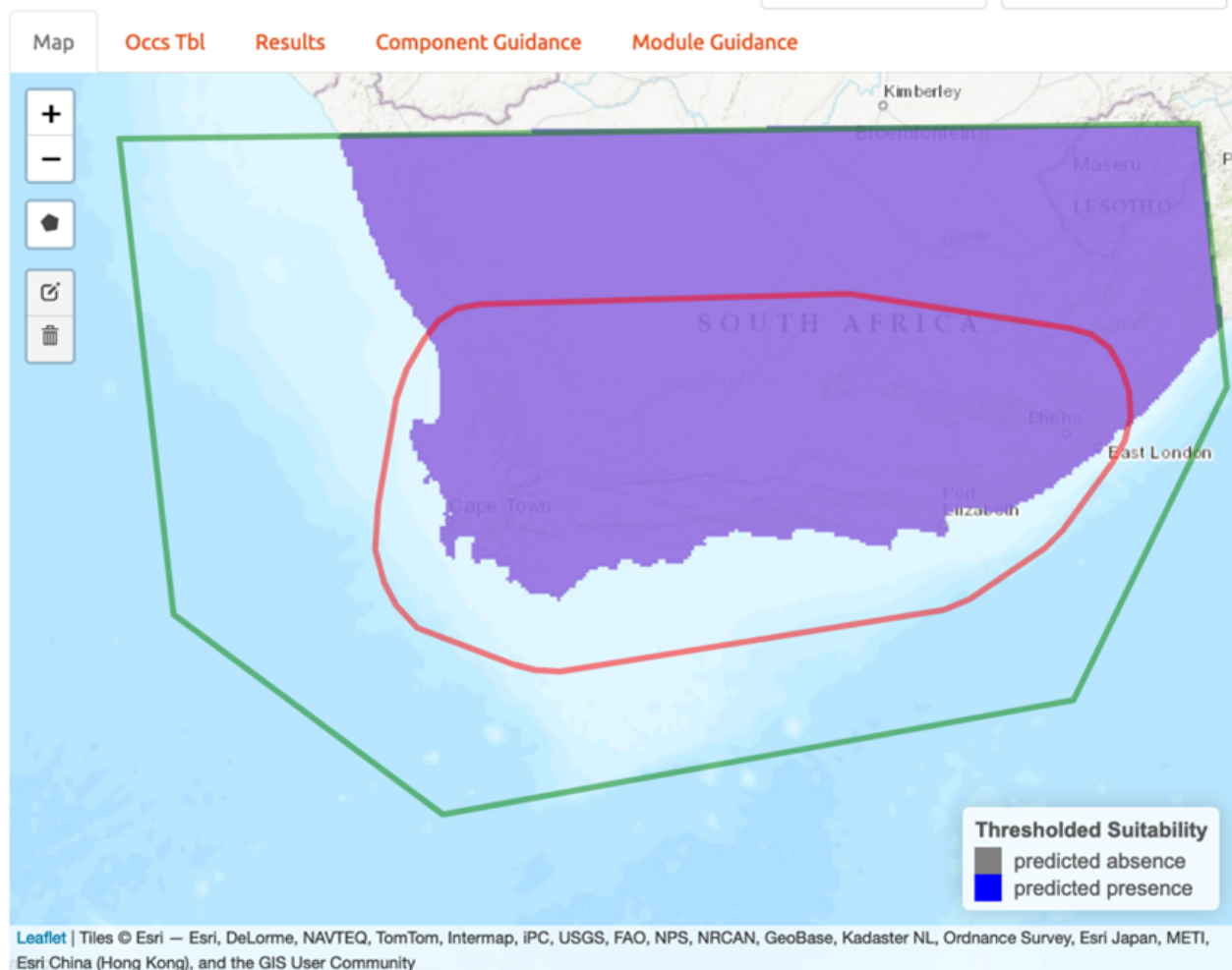
c) Click on “Draw a Polygon” on the left-hand of the map, then draw a polygon of the desired new extent of your projection. NOTES: Only project to the region you’re interested in. Global projections take a long time, and a lot of computing power. Also, a limitation of Wallace is that the new extent (the projection region) must include the full extent of the calibration region.



d) From the “Set threshold” dropdown menu, select the model threshold you want.

e) Press “Project” under “Project model to current extent”. Be patient; it takes time to mask environmental grids to the new extent and project the model to this new area.

f) Once the model projection is complete, delete the projection polygon you drew. To do this, click on the garbage can icon on the left side of the map (circled in black) and press “Clear all”. This should leave the polygon outline but remove the gray fill so you are able to view the model projection results.



g) Save your model projection. Under the “Select download file type” dropdown menu, select “ASCII”. Press “Download”. Save the file to your working project folder. NOTE: The file name automatically generated by Wallace is the exact same as the file name produced for the model training and projection files (the format includes the feature class selection of the model, the model number, and the selected threshold). As such, be sure to add “_proj” to the end of the file name. For example, the projection file name for the example provided here would be “LQHP_2_thresh_mtp_proj.asc”.

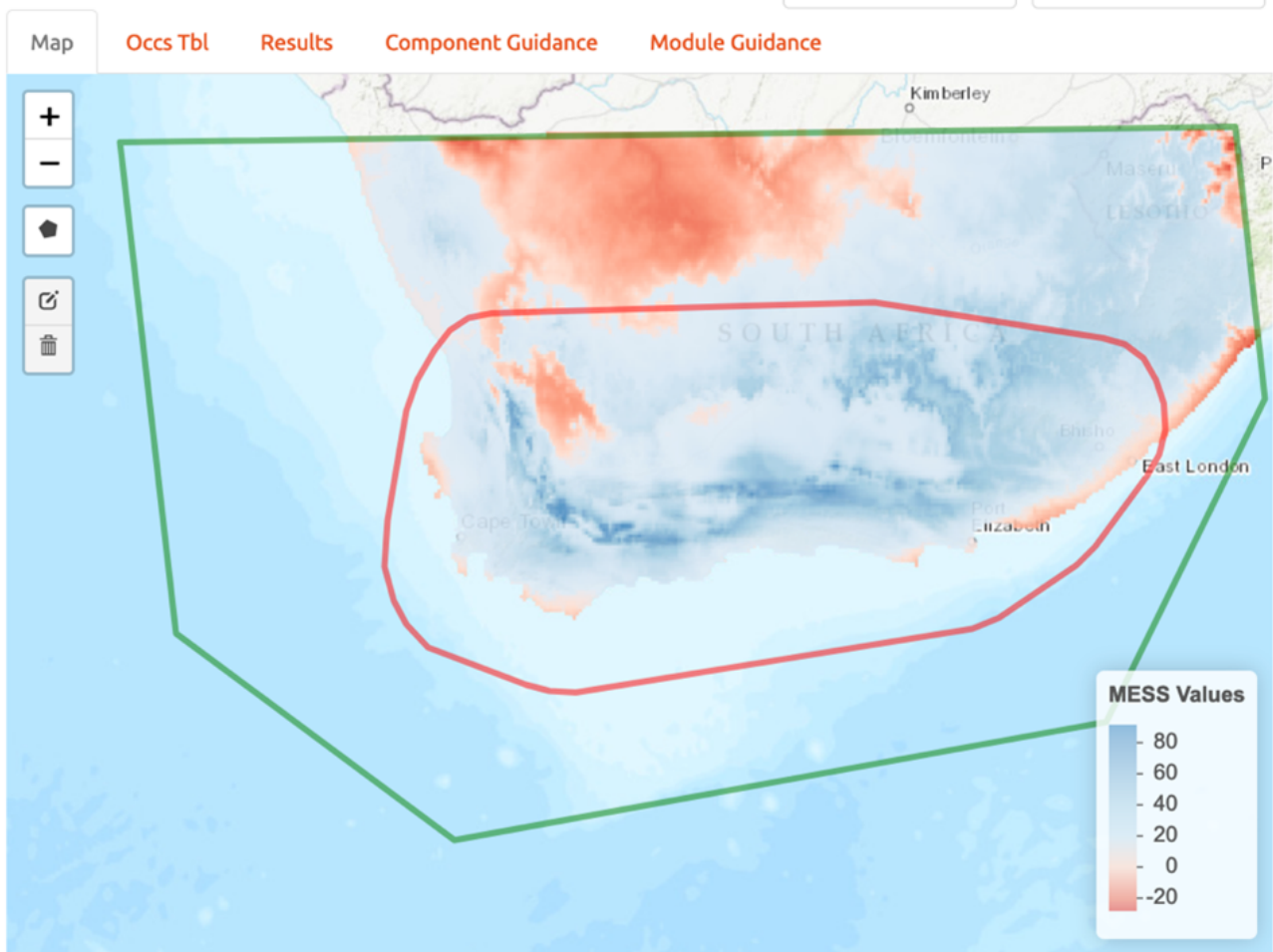
Exercise 12 - Calculating Environmental Similarity

MESS analyses allow us to characterize the degree to which the model projection region differs from the bioclimatic conditions of the model calibration region.

a) Under “Project Model: Modules Available” select the “Calculate Environmental Similarity” radio button.

b) Press the “Calculate MESS” button under “Calculate MESS for current extent”. Be patient; this process can take a fair bit of time depending on the geographic extent and spatial resolution of your data.

c) Look at the resulting map. What stands out most? High positive values indicate increasing similarity with the conditions used to train the model, and low negative values indicate increasing difference relative to the model calibration bioclimatic conditions.



d) Save MESS evaluation. Under the “Select download file type” dropdown menu, select “GeoTIFF”. Then press “Download”. Save the file to your working project folder. NOTE: The file name automatically generated by Wallace is the exact same as the file name produced for the model training and projection files (the format includes the feature class selection of the model, the model number, and the selected threshold). As such, be sure to add “_MESS” to the end of the file name. For example, the file name for the example provided here would be “LQHP_2_thresh_mtp_MESS.asc”.

Exercise 13 - Saving Your Session Code

It’s best practice to always maintain detailed records of the specific steps taken during research. Conveniently, Wallace provides us with the option to download a record of actions taking during the modeling session.

- a) Click on “Session Code” in the browser window in which Wallace is running.
- b) Under “Select download file type” dropdown menu, choose “Rmd”.
- c) Click on “Download Session Code”. Save the file to your working project folder. The default file name should work just fine.
- d) Congratulations! You have now successfully completed (maybe) your first niche model, and have the code to reproduce the whole analysis in R. If you open the RMD file in Rstudio, you will be able to see the code (and tweak and re-run the analysis, if you want!). You can also use this code to generate a pdf report detailing your analysis.

Assessing the conservation status of a species



This module will introduce you to the basic concepts and terminology associated with assessing the conservation status of a species using the IUCN Red List Categories and Criteria. By the end of the module, you should understand the conceptual framework of the IUCN Red List of Categories and Criteria and how you can apply them using GBIF-mediated data. This module does not cover the entirety of the Red List assessment process but focuses on how GBIF-mediated data can be used within these processes. For a complete training you should complete the [online IUCN Red List training course](#).

IUCN Red List of Threatened Species

The International Union for Conservation of Nature (IUCN) Red List of Threatened Species provides a robust and transparent framework in the form of the [IUCN Red List Categories and Criteria](#) for estimating the risk of extinction of all taxa (excluding microorganisms and taxa below subspecies) across all systems - marine, terrestrial and freshwater. It is an internationally recognised standard that can be applied at global, regional and national scales and is acknowledged as a key tool for assessing progress towards achieving biodiversity targets as set out in the Convention on Biological Diversity.

The screenshot shows the IUCN Red List website interface. At the top, there is a navigation bar with the IUCN logo and the text "THE IUCN RED LIST OF THREATENED SPECIES™". Below the navigation bar is a search bar with the placeholder text "Names - common, scientific, regions etc...". To the right of the search bar is an "Advanced" button. Below the search bar is a grid of four species cards, each featuring a photograph of the species, its name, scientific name, and conservation status. The species shown are: Scaly-foot Snail (Chrysomallon squamiferum), Aneгада Rock Iguana (Cyclura pinguis), Araripe Manakin (Antilophia bokemanni), and Giant Kangaroo Rat (Dipodomys ingens). Below the grid is a link for "Amazing species". At the bottom of the screenshot is a red banner with the text "More than 32,000 species are threatened with extinction. That is still 27% of all assessed species." Below the banner is a bar chart showing the percentage of threatened species in different taxonomic groups: Amphibians (41%), Mammals (26%), Conifers (34%), Birds (14%), Sharks & Rays (30%), Reef Corals (33%), and Invertebrates (28%). At the very bottom is a "Take action" button and the text "Help us make The IUCN Red List a more complete barometer of life."

Taxonomic Group	Percentage of Threatened Species
Amphibians	41%
Mammals	26%
Conifers	34%
Birds	14%
Sharks & Rays	30%
Reef Corals	33%
Invertebrates	28%

All species have a probability of going extinct due to random events. However, some species have a higher probability of extinction due to a number of determining factors such as population trends, range and threats faced by the species. The IUCN Red List Categories and Criteria provides a framework against which this extinction risk can be assessed.

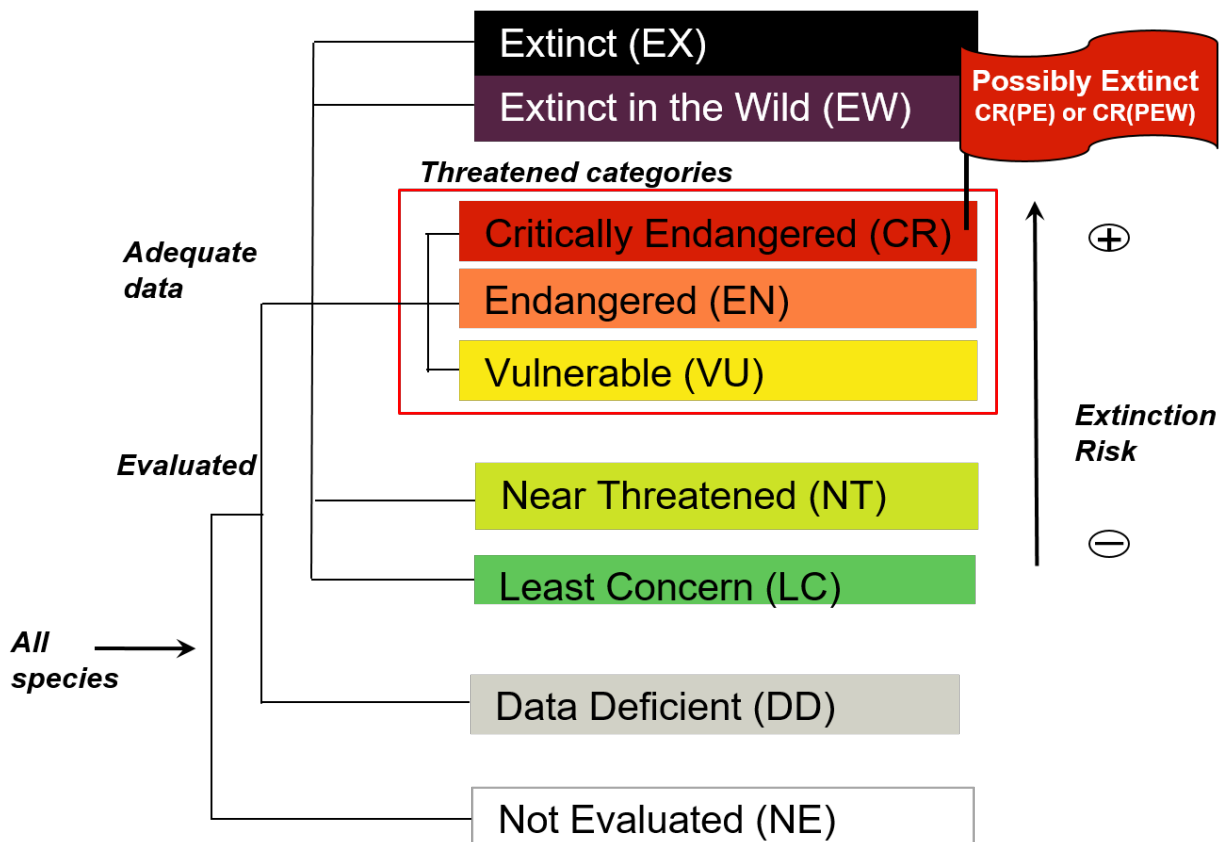
IUCN Red List Categories and Criteria

Assessments are based on 5 objective, scientific criteria, each containing a set of quantitative thresholds against which the risk of extinction of a species can be assessed. Species must be assessed against all criteria during an assessment.

These 5 criteria are:

- Criterion A - Population reduction
- Criterion B - Restricted geographic range
- Criterion C - Small population size and decline
- Criterion D - Very small or restricted population
- Criterion E - Quantitative analysis

Each of these criteria has a set of associated thresholds for these biological traits that allow assessors to assess the risk of extinction for that species and apply one of 9 categories.



All criteria should be applied to a taxon during an assessment and it is the criterion with the highest threat category that is used as the final Red List assessment.

Global vs National Red List Assessments

The IUCN Red List of Categories and Criteria were developed for applying at a global level i.e. to take into account a species entire global distribution that may cross international borders. The majority of species currently on the [IUCN Red List](#) are global assessments. However, species can be assessed at a regional, national or local level and for these, assessors should use the [Guidelines for Application of IUCN Red List Criteria at Regional and National Levels](#), an adaptation of the global Categories and Criteria. These regional guidelines provide additional guidance on:

- Deciding on which species should be assessed at a regional, national or local level
- Additional categories for assessments at a regional, national and a local level
- Assessing breeding vs non-breeding species at a regional, national and a local level
- Integrating information on the species from across its global distribution in regional, national and a local assessments.

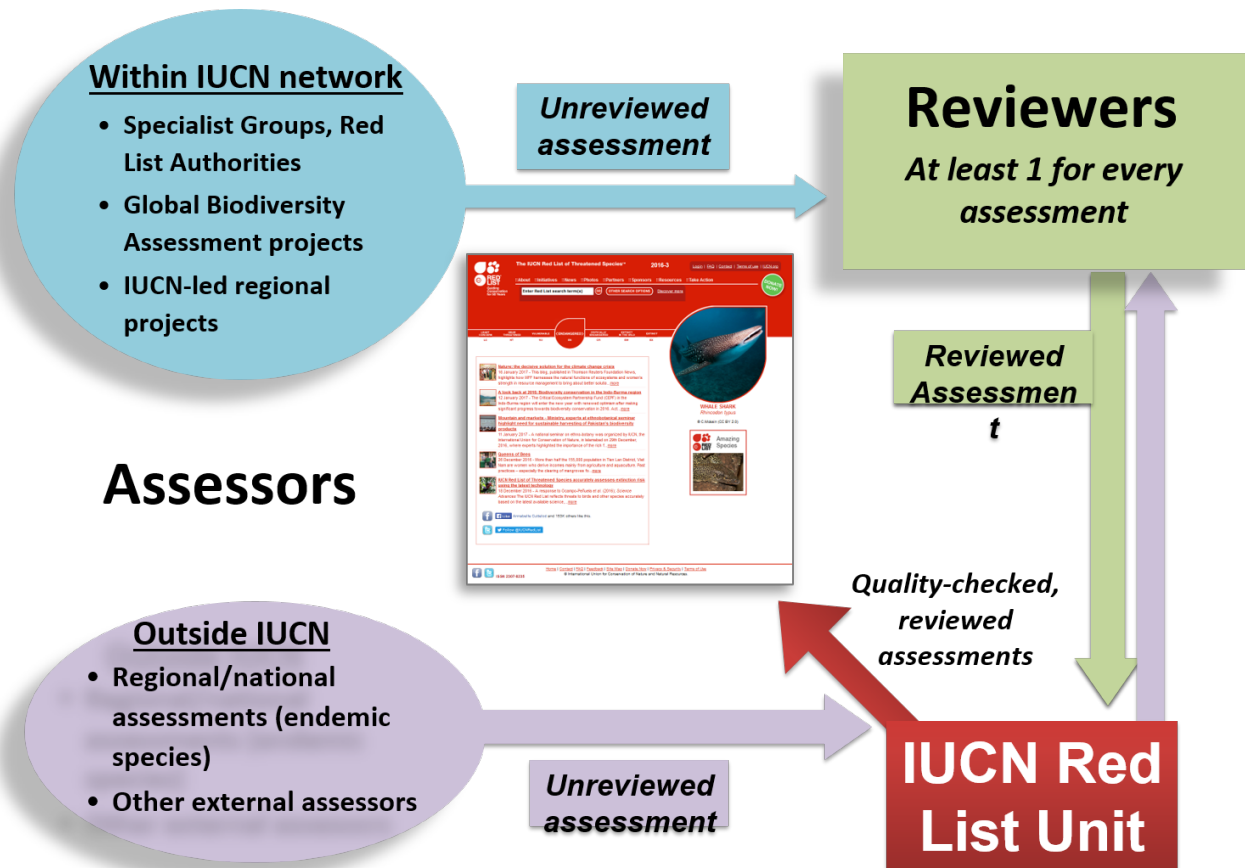
It should be noted that regional, national or local assessments of species that are endemic to those regions are, by default, global assessments and no additional regional correction is needed.

It is important to realise that while these guidelines are in place to assist with regional and national assessments, national and regional red lists may not have followed the IUCN guidelines for assessing species. For more information about national Red Lists published by countries around the world, see the [National Red List](#) website.

Red List assessment process

Assessments that are to be integrated into the global IUCN Red List of Threatened Species can come from coordinated efforts within the IUCN network e.g. [IUCN Species Survival Commission Specialist Groups](#) or from other processes such as the development of national Red Lists, if there are species that are being assessed that are endemic to that country. National Red Listing processes may differ from those prescribed by IUCN, but ANY assessment to be submitted to the global IUCN Red List of Threatened Species will have to go through the following stages:

- **Assessment** - Assessors are experts who have sufficient knowledge of a taxon to be able to apply the criteria in an informed way, these experts can come from within the IUCN network such as the [IUCN Species Survival Commission Specialist Groups](#) or from national or regional taxonomic experts.
- **Review** - Reviewers of global assessments are generally from the network of Red List Authorities (RLAs), which are mostly IUCN Specialist Groups, but where there are gaps, other organisations act as RLAs (e.g., Project Seahorse, BirdLife International, etc) and agree that they are appropriate based on all data currently available for the species. Assessments coming to the IUCN Red List Unit from within IUCN should already have been through the review process (the RLAs nest within the SGs). Assessments coming from outside the IUCN network need to go through the peer review process.



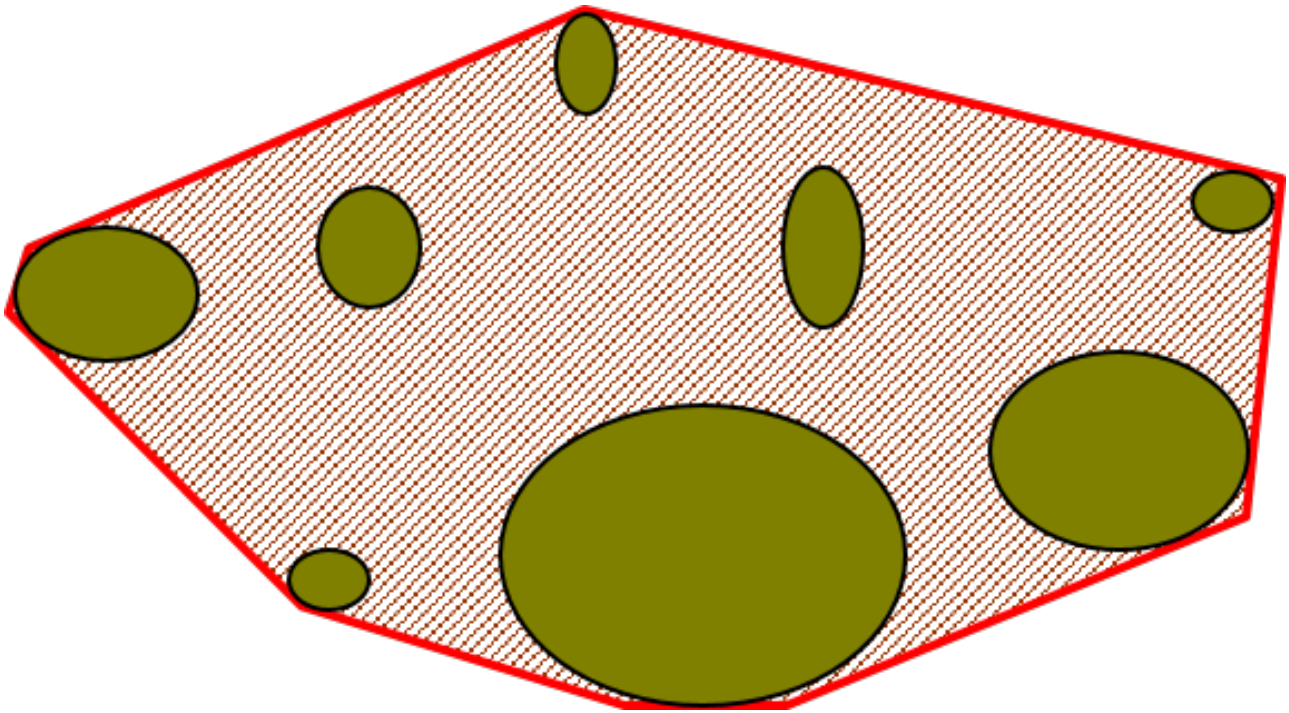
GBIF-mediated data and Red List assessments

Key to the Red List assessment process is data and the Categories and Criteria allow for the use of a range of data of heterogeneous quality within an assessment. These data can be observations, estimations, projections, inferences or suspicions. Processed GBIF-mediated data is a source of observation data providing georeferenced locality data that can be used to calculate key metrics in the assessment process, particularly for Criterion B and for producing species distribution maps that are required to accompany assessments. Remember, that ALL criteria should be applied during an assessment, which is why you will ideally have additional information on population sizes and trends along with information on threats to the species.

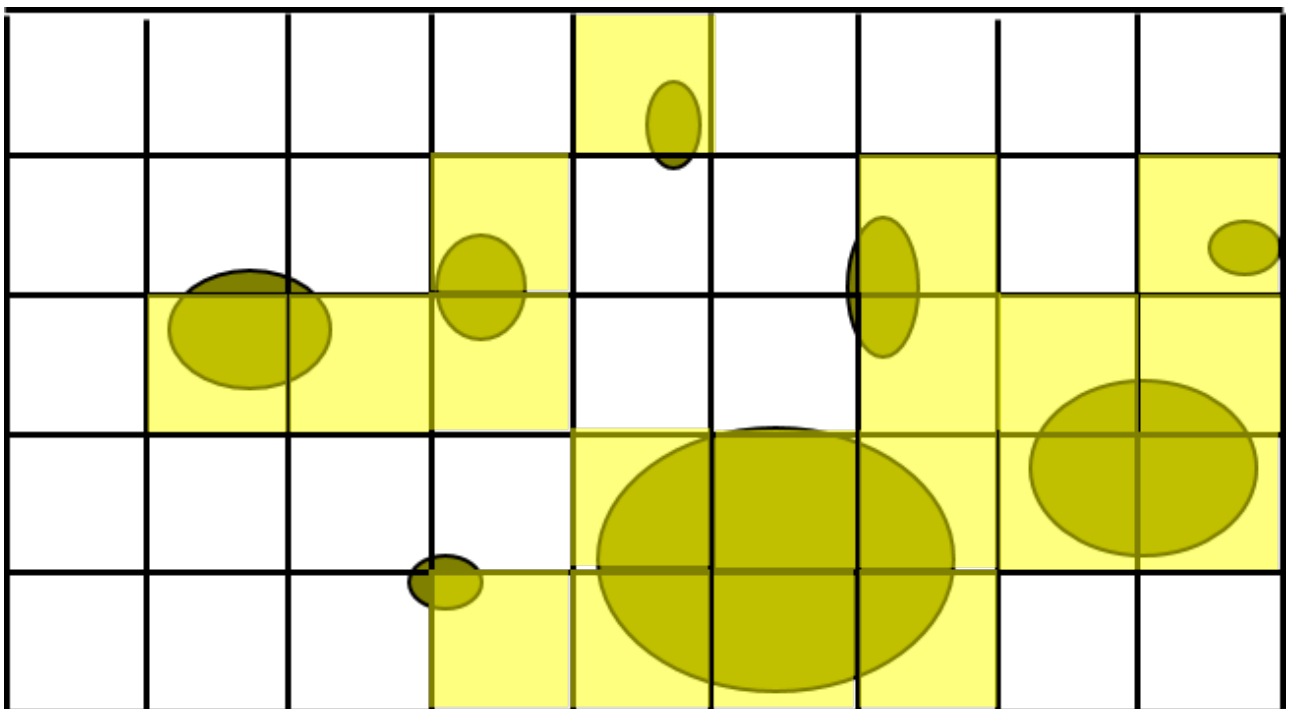
Applying Criterion B - Restricted Geographic Range

Criterion B identifies populations with restricted distributions that are also severely fragmented or occur in a small number of locations, are experiencing continuing decline, or are exhibiting extreme fluctuations. Taxa with very large ranges will generally have a lower risk of extinction than a species with a highly restricted distribution, which is likely to be more at risk from localised threats.

Two of the metrics within criterion B that are used for identifying these restricted distributions are Extent of Occurrence (EOO) and Area of Occupancy (AOO). Extent of Occurrence is the area within the shortest continuous imaginary boundary drawn around all known, inferred, or projected sites presently occupied by the taxon. It is not the species range and is drawn as minimum convex polygon around the limits of a species known range.



Area of Occupancy is the area within the extent of occurrence that is actually occupied by the taxon. It is measured by overlaying a 2x2 km grid and counting the number of occupied cells.

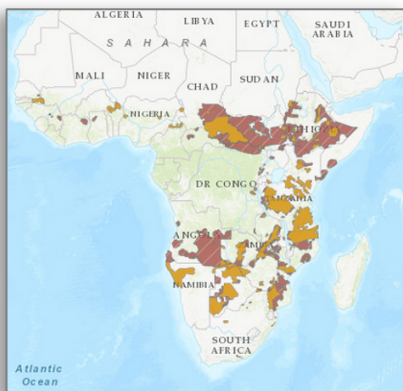


Both these metrics require georeferenced locality data and GBIF-mediated data can be used for calculating both EOO and AOO of species. A number of tools have been developed for calculating these measurements including ArcGIS toolboxes, the R package red and GeoCat. The latter provides users with little programming or GIS experience, the ability to take GBIF-mediated data and calculate EOO and AOO measurements.

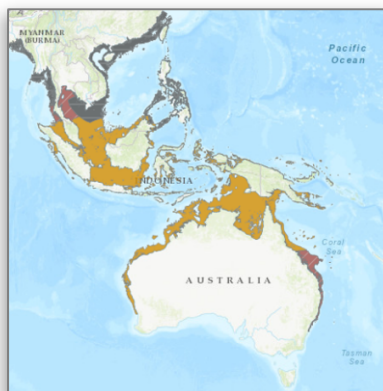
Mapping standards for IUCN Red List Assessments

All assessments should be accompanied by a distribution map. Maps are included on the Red List for

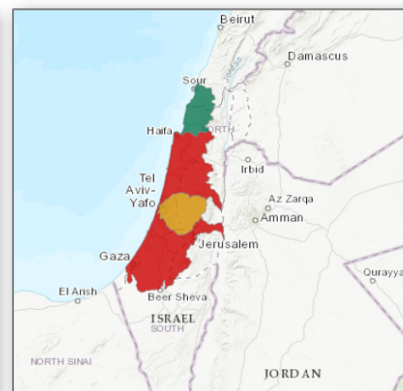
several reasons. Primarily, the maps provide a visual representation of the taxon’s distribution, so people can see where the taxon is found and help to identify priority areas for conservation and inform conservation policy. Different mapping standards are applied for different taxonomic groups and for whether the species is terrestrial, marine or freshwater. Full guidance on the application of these standards can be found on the [IUCN Mapping Standards](#) webpage.



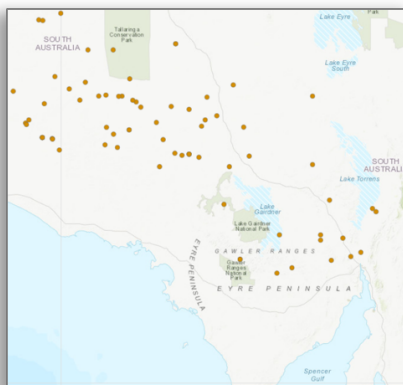
Terrestrial



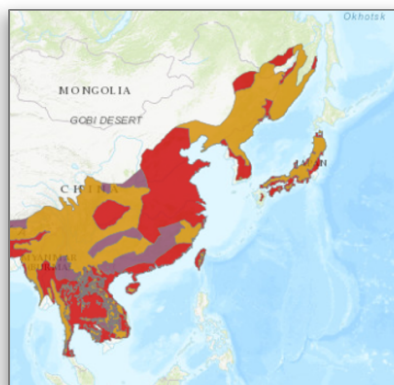
Marine



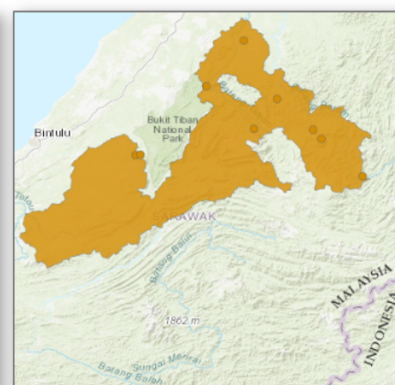
Freshwater



Plants



Vertebrates



Invertebrates

In many cases the distribution is depicted as polygons, but it may also be represented by data points (collection records), or a mixture of points and polygons. Polygon maps, commonly referred to as “limits of distribution” or “field guide” maps, aim to provide the current known distribution of the species within its native range i.e. the species probably only occurs within that polygon. The taxon may not be distributed equally within that polygon or occur everywhere within that polygon. These limits of distribution are determined by using known occurrences of the species, along with knowledge of habitat preferences, remaining suitable habitat, elevation limits, and other expert knowledge of the species and its range.

Minimum Documentation

Assessors should provide with their maps, whether they are points, polygons or a combination of both, a set of accompanying attributes i.e. data attached to points and polygons. Some of these attributes are required as part of minimum documentation supporting assessments. A full overview of these minimum documentation requirements can be found in the mapping standards guidelines

on the [IUCN Mapping Standards](#) webpage. You can also find a downloadable Excel file at the above link where attribute fields have been mapped to Darwin Core fields to highlight those fields in your GBIF downloads that fulfil minimum documentation requirements when submitting Red List assessment maps.

Use Case - Red List Assessment



This Use Case is a practice use case for the data processing and assessing conservation status using the IUCN Categories and Criteria module. Please note that this use case is fictitious and built for instructional purposes only. Any reference to countries and structures/organisations, real or otherwise, within those countries are used merely to facilitate the use of the data and may not reflect the reality within those countries.

Scenario



Brachypelma smithi (F.O.Pickard-Cambridge, 1897) observed in Mexico by adrianita (licensed under <http://creativecommons.org/licenses/by-nc/4.0/>) - <https://www.gbif.org/occurrence/1453068554>

Mexico is a mega-diverse country with high numbers of species and high levels of endemism across all taxonomic groups. The government is committed to assessing the conservation status of all its species and has agreed to develop a national red list applying IUCN Categories and Criteria to all vertebrate and invertebrate taxonomic groups by the year 2030. As part of this process, they have started the assessment of arachnid groups and are now assessing the conservation status of all species within the family Theraphosidae – the tarantulas.

National and international arachnid specialists have come together at a Mexican arachnid red list workshop and they will now be assessing the conservation status of the tarantula species *Brachypelma smithi* (F.O.Pickard-Cambridge, 1897) and *Aphonopelma anax* (Chamberlin, 1940). Assessors want to use GBIF-mediated data to assess the species using Criterion B and provide maps that comply to IUCN documentation requirements.

Description

Brachypelma smithi (F.O.Pickard-Cambridge, 1897) is a large spider species in the family Theraphosidae. It has a very restricted range within the state of Guerrero, requiring clearings in tropical dry forests on sandy soil where it can build its tunnel/den. As with other tarantula species, males and females have different lifespans with averages of 6 and 25 years respectively. Both males and females reach sexual maturity on average after 4 years. The species plays a role in local indigenous culture where it is considered a vessel for the souls of community members and is thus afforded protection where the species is found near local, indigenous communities. The species is threatened by the collection of individuals for the pet trade and is listed under CITES Appendix II; it is highly prized by collectors due to its rarity.

Aphonopelma anax (Chamberlin, 1940) is a large spider species in the family Theraphosidae. It is found in Southern Texas, USA and North Mexico. They are found in semiarid climates in grasslands and shrub forests as well as within cities. They live in burrows that can be created by themselves or modifying suitable habitats such as dead trees, empty rodent burrows, stacks of wood or natural crevices. They have slow growth rates and live for several years before maturing – females can live up to 40 years while males rarely live over 2 years once they have matured. The species is common across its range and although there has been increasing urban and agricultural development across its range in both the USA and Mexico, the extent of this development is limited to only a small portion of the species entire range.

Exercise - Applying IUCN Red List Criterion B

In this exercise, you will calculate the Extent of Occurrence and Area of Occupancy for *Brachypelma smithi* and *Aphonopelma anax* using GeoCat - www.geocat.kew.org and then do Red List assessments for the species. For the purposes of this exercise, we will apply the global IUCN Categories and Criteria to both species across their range in Mexico. In reality, the regional Categories and Criteria should be applied as these are national assessments.

<i>Brachypelma smithi</i> (F.O.Pickard-Cambridge, 1897)		
Extent of occurrence		
Area of Occupancy		
		Justification (please state whether this is observed, estimated, projected, inferred or suspected)
Severe Fragmentation	Yes or No	
Number of Locations		

<i>Brachypelma smithi</i> (F.O.Pickard-Cambridge, 1897)		
Continuing Decline	Yes or No	
Extreme Fluctuations	Yes or No	
Final Assessment		

<i>Aphonopelma anax</i> (Chamberlin, 1940)		
Extent of occurrence		
Area of Occupancy		
		Justification (please state whether this is observed, estimated, projected, inferred or suspected)
Severe Fragmentation	Yes or No	
Number of Locations		
Continuing Decline	Yes or No	
Extreme Fluctuations	Yes or No	
Final Assessment		

Key documentation



The following references provide further detail on the topics covered in this course. All links open in a new window/tab.

API

- [GBIF API beginners guide](#)
- [\(Almost\) everything you want to know about the GBIF Species API](#)

Cloud Computing

- [GBIF and Apache-Spark on Microsoft Azure tutorial](#)
- [GBIF and Apache-Spark on AWS tutorial](#)

Darwin Core

- [Darwin Core Terms: A quick reference guide](#)
- [Simple DarwinCore](#)
- [Darwin Core Questions & Answers](#)

- Darwin Core extensions registered with GBIF

Data publishing

- Quick guide to publishing data through GBIF
- How to publish biodiversity data through GBIF.org
- Become a data publisher with GBIF
- Best Practices for Publishing Biodiversity Data from Environmental Impact Assessments
GBIF Secretariat & IAIA: International Association for Impact Assessment (2020).
- Guidance for private companies to become data publishers through GBIF: Template document to support the internal authorization process to become a GBIF publisher
Rui Figueira, Pedro Beja, Cristina Villaverde, Miguel Vega, Katia Cezón, Tainan Messina, Anne-Sophie Archambeau, Rukaya Johaadien, Dag Endresen & Dairo Escobar (2020).
- Publishing DNA-derived data through biodiversity data platforms
Anders F. Andersson, Andrew Bissett, Anders G. Finstad, Frode Fossøy, Marie Grosjean, Michael Hope, Thomas S. Jeppesen, Urmas Kõljalg, Daniel Lundin, R. Henrik Nilsson, Maria Prager, Cecilie Svenningsen & Dmitry Schigel (2020).
- Classes of datasets supported by GBIF
- GBIF data quality requirements for publishing
- GBIF data licenses
- Checklist core templates
- Occurrence core templates
- Sampling event core templates
- Sampling event data best practices
- Sharing images, sounds and videos on GBIF
- Data papers
- Published data papers

Data publishing: IPT

- The GBIF Integrated Publishing Toolkit: Facilitating the Efficient Publishing of Biodiversity Data on the Internet
Robertson et al. (2014)
- To install IPT or not to install IPT
- IPT data hosting centres
- IPT Install / Set up webinar
- Installing the IPT video
- IPT in Practice demonstration video

Digitization

- [iDigBio Digitization Resources](#)
- [iDigBio Collections Digitization Workflows](#)
- [iDigBio Digitization Workflows and Protocols](#)
- [iDigBio specimen image capture guide](#)
- [Canadensys 10-step guide to managing images with your biodiversity data](#)

GBIF

- [What is GBIF](#)
- [Strategic Plan](#)
- [Become a member](#)
- [Science Review](#)
- [Establishing an Effective GBIF Participant Node: Concepts and general considerations](#)
GBIF Secretariat (2019).

Georeferencing

- [Georeferencing Best Practices](#)
Arthur D. Chapman & John R. Wieczorek (2020).
- [Georeferencing Quick Reference Guide](#)
Paula F. Zermoglio, Arthur D. Chapman, John R. Wieczorek, Maria Celeste Luna & David A. Bloom (2020).
- [Georeferencing Calculator Manual](#)
David A. Bloom, John R. Wieczorek & Paula F. Zermoglio (2020).
- [Georeferencing resources](#)

Invasive Species

- [GRISS - Global Register of Introduced and Invasive Species](#)
- [TrIAS - Tracking Invasive Alien Species](#)

Living Atlases

- [Living Atlases](#)
- [ALA key technical documentation](#)

Miscellaneous

- [VertNet Guide to opening text files in Excel](#)
- [VertNet data licensing guide](#)

OpenRefine

- [OpenRefine documentation](#)
- [OpenRefine regular expressions](#)
- [Guía para la limpieza de datos sobre biodiversidad con OpenRefine](#)
Paula F. Zermoglio, Camila A. Plata Corredor, John R. Wieczorek, Ricardo Ortiz Gallego & Leonardo Buitrago (2021).
- [Using Google Refine and taxonomic databases \(EOL, NCBI, uBio, WORMS\) to clean messy data](#)
iPhylo blog post. Rod Page 2012.
- [Reconciling author names using Open Refine and VIAF](#)
iPhylo blog post. Rod Page 2013.
- [Validating scientific names with the GBIF Portal web service API](#)
Guest post was written by Gaurav Vaidya, Victoria Tersigni and Robert Guralnick 2013.
- [iDigBio Cleaning data with OpenRefine](#)
iDigBio 2013.
- [Have We Got the Names “Right”?](#)
Canadensys 2014.
- [Cleaning data with OpenRefine](#)
Desmet and Brosens 2016 TDWG.
- [EasyOpen Redlist](#)
Querying the IUCN Red List, using a species list, OpenRefine, and some pre-written code. Olly Griffin July 2019.

Planning/Collaboration

- [Agile](#)
- [What is SCRUM](#)
- [SCRUM Framework](#)
- [Kanban methodology](#)
- [Scrum Guide](#)
- [GitHub](#)

Red List Assessments

- **IUCN Red List Categories and Criteria**

- [IUCN Red List Categories and Criteria v3.1 - ENGLISH](#)
- [Categorías y Criterios de la Lista Roja de la UICN v3.1 - ESPAÑOL](#)
- [Catégories et Critères de la Liste Rouge de l'UICN v3.1 - FRANÇAIS](#)

- **Red List Guidelines**

- [Guidelines for Using the Red List Categories and Criteria version 15 - ENGLISH](#)
- [Directrices de uso de las Categorías y Criterios de la Lista Roja de la UICN Versión 14 - ESPAÑOL](#)
- [Lignes directrices pour l'utilisation des Catégories et Critères de la Liste rouge de l'UICN Version 14 - Français](#)

- **Criteria Summary Sheet**

- [Criteria summary sheet - ENGLISH](#)
- [Resumen de los Criterios - ESPAÑOL](#)
- [Résumé des Critères - FRANÇAIS](#)

- **Mapping Standards**

- [Mapping Standards - ENGLISH](#)
- [Standard Attributes for Spatial Data - ENGLISH](#)
- [Attribute codes for Presence](#)

- **Regional and National Levels Guidelines**

- [Guidelines for application of IUCN Red List Criteria at Regional and National Levels Version 4 - ENGLISH](#)
- [Directrices para el uso de los Criterios de la Lista Roja de la UICN a nivel regional y nacional Versión 4 - ESPAÑOL](#)
- [Lignes directrices pour l'application des Critères de la Liste rouge de l'UICN aux niveaux régional et national Version 4. - FRANÇAIS](#)

- **Spatial Tools and Data for Red List Assessments**

Includes brief overviews of some tools developed to aid mapping of spatial data and estimation of Red List metrics such as EOO and AOO, plus developer information, links to help files, support networks and associated research publications.

- **Supporting information guidelines**

The Documentation Standards and Consistency Checks for IUCN Red List Assessments and Species Accounts (also known as the Supporting Information Guidelines) provides guidance on the required and recommended supporting information for Red List assessments. It also provides guidance on the writing style and format that should be used for all IUCN Red List assessments.

Quality

- [Principles of Data Quality](#)
Arthur Chapman 2005.
- [Principles and Methods of Data Cleaning: Primary Species and Species-Occurrence Data](#)
Arthur Chapman 2005.
- [Be careful with dates in Excel](#)
DataOne 2014.
- [Character encoding for beginners](#)
- [MVZ Guide for Recording Localities in Field Notes](#)

R

- [Data Carpentry-Data Analysis and Visualization in R for Ecologists](#)
- [DataCamp- Range of courses in R, Python and SQL](#)
- [Introduction to R Manual](#)
- [rgbif Manual](#)
- [CoordinateCleaner Manual](#)
- [Downloading occurrences from a long list of species in R and Python](#)
- [Common things to look out for when post-processing GBIF downloads](#)
- [Finding gridded datasets & Gridded Datasets Update](#)
- [Country centroids](#)
- [Using shapefiles on GBIF data with R](#)
- [Not a bird download](#)

Sensitive species

- [Current Best Practices for Generalizing Sensitive Species Occurrence Data](#)
Arthur D. Chapman 2020.

Taxonomy

- [GBIF checklist datasets and data gaps](#)
- [GBIF Labs - Names Parser](#)
- [GBIF Labs - Species Matching](#)
- [Global Names Resolver](#)
- [Marine Name Matching Strategy for taxonomic quality control](#)
- [Nomenmatch](#)

Glossary

ALA

Atlas of Living Australia. The Australian node of GBIF, who developed an open source data portal now widely used within the GBIF community & partners for biodiversity national portals.

API

Application Programming Interface. A set of clearly defined methods of communication between various software components.

BID

Biodiversity Information for Development. An EU funded project co-ordinated by GBIF whose aim is to increase data mobilization capacity in the Africa, Caribbean and Pacific regions.

BIFA

Biodiversity Fund for Asia.

CC Licences

Creative Commons. These are a series of licenses set up by the Creative Commons organization that enable sharing and reuse of creativity and knowledge through the provision of free legal tools. Three of them can be assigned to GBIF-shared datasets: CC0, CC BY and CC BY-NC.

Controlled Vocabulary

This is a restricted set of terms that are used as possible values for a given field. One can think of it as a lookup list or dropdown for a particular field. For example the DwC field `basisOfRecord` should only contain one of these values: "PreservedSpecimen", "FossilSpecimen", "LivingSpecimen", "HumanObservation", "MachineObservation". We would say that list of values is a controlled vocabulary for that field.

DwC

Darwin Core is a biodiversity data standard, maintained by TDWG & widely used within the GBIF community and partners. It is a set of standardized terms (or field names) and their definitions, which are used to share biodiversity information.

DOI

Digital Object Identifier. A persistent identifier or handle used to uniquely identify objects. DOIs are in wide use mainly to identify academic, professional, and government information, such as journal articles, research reports and data sets, and official publications.

DwC-A

Darwin Core Archive. A compressed (zipped) file containing all the information needed to share with GBIF, for a particular resource. Each zip contains three types of files:

1. the actual data, in one or more text files: `occurrence.txt/event.txt/measurmentoffact.txt` etc
2. a mapping file: `rtf.xml`
3. a metadata (EML) file: `eml.xml` When you publish using the IPT, it creates a Darwin Core Archive, which is shared with GBIF. Also, when you download data from the GBIF website you

can choose a DwC-A format as well.

GUID

Globally Unique Identifier

IPT

Integrated Publishing Toolkit. It is a free and open source web application (software) for publishing biodiversity data. The software itself lives on a server (either at your institution or elsewhere) that must have access to the internet 24/7. It is used to create and handle Darwin Core Archive files that can be shared and used by anyone including GBIF.

Loan

In the context of natural history collections, this is the procedure of lending specimens between institutions.

LSID

Life Sciences Identifier. They are persistent, globally unique identifiers for biological objects.

Data Publishing

With regards to GBIF we have a very specific definition of data publishing. It refers to making biodiversity datasets publicly accessible and discoverable, in a standardized form, via an access point, typically a web address (a URL).

Resource

A Resource is the collective term used to refer to a particular dataset and its metadata once it has been uploaded to an IPT instance.

TDWG

Taxonomic Databases Working Group, now renamed Biodiversity Information Standards.

URN

Uniform Resource Number

UUID

Universally Unique Identifier

Acknowledgements

Course design and instruction

The success of this course depends heavily on the support provided to participants from GBIF's network of capacity enhancement mentors. Visit the GBIF page on [capacity enhancement mentoring](#) to read more about these individuals and their contributions.

The following individuals are recognized for their significant contributions to the course design, materials and instruction:

- Nadine Bowles Newark
- Andrea Baquero
- Kate Ingenloff
- Hannah Owens
- Melianie Raymond
- Andrew Rodrigues
- Laura Anna Russell
- John Tayleur
- John Waller

Special acknowledgement ...

Translators

Spanish

- Anabela Plos
- Paula Zermoglio

French

- Patricia Mergen
- Anne Sophie Archambeau

Resources

- R
- [Wallace Ecological Niche Modelling Vignette](#)
- [IUCN Red List of Threatened Species](#)

Colophon

Suggested citation

GBIF Secretariat (2021) GBIF Biodiversity Data Use Course. 6th edition. GBIF Secretariat: Copenhagen. <https://doi.org/10.15468/ce-wkk4-2w26>. [Date of course.]

Contributors

The *GBIF Biodiversity Data Use Course* was originally developed as part of **Biodiversity Information Development**, a programme funded by the **European Union**. The original curriculum was created by Andrew Rodrigues, Hannah Owens and John Tayleur with additional contributions by GBIF trainers, mentors and students.

Licence

Course Name is licensed under **Creative Commons Attribution 4.0 Unported License**.

Persistent URI

<https://doi.org/10.15468/ce-wkk4-2w26>

Document control

Fifth edition, November 2021

Cover image

Di Marco M, Ferrier S, Harwood TD, Hoskins AJ and Watson JEM (2019) Wilderness areas halve the extinction risk of terrestrial biodiversity. *Nature*. Springer Science and Business Media LLC 573(7775): 582–585. Available at: <https://doi.org/10.1038/s41586-019-1567-7>