Sample Data and Training Modules for Cleaning Biodiversity Information

Marlon E. Cobos, **Laura Jiménez***, Claudia Nuñez-Penichet, Daniel Romero-Alvarez, Marianna Simões

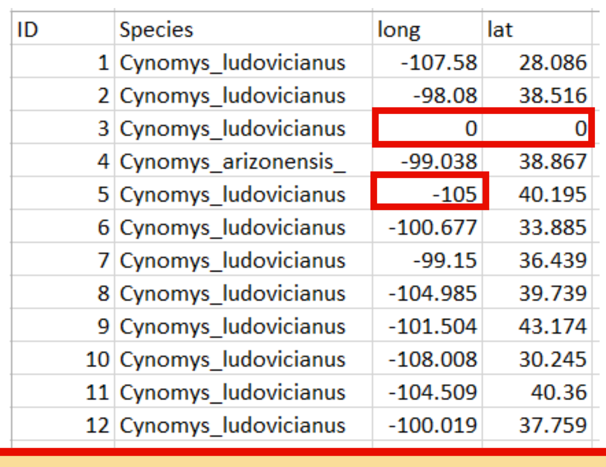
Department of Ecology and Evolutionary Biology and Biodiversity Institute, University of Kansas. *E-mail: laura_jimenez@ku.edu

ABSTRACT

The recent appreciation of rapid global losses of biodiversity has created a growing demand for quick and reliable access to high-quality primary biodiversity data, essential for conservation and evolutionary science, among other applications. As a consequence, biodiversity databases have become increasingly available, allowing large-scale assessment of patterns and processes influencing the evolution of life on Earth. However, data quantity is often compromised by low data quality, and, even though working with plenty of records could be tempting, building ecological models with inaccurate data mislead researchers in driving conclusions about reality. The most common types of errors in biodiversity data are those related to georeferencing, which vary from obvious to barely noticeable, making them challenging to recognize. Fortunately, Geographic Information Systems (GIS) have increasingly allowed identification of georeferencing mistakes. Here, we provide a hands-on exercise for data cleaning that allows easy and prompt detection of inaccurate information. We focus on the use of GIS software for fixing problems related to the geographic coordinates of the data, such as changes in the order or the sign of latitude-longitude values, and records placed outside the study region, or outside continents.

LOW QUALITY BIODIVERSITY DATA

Primary biodiversity data have become much more accessible in recent decades. Many institutions (e.g., museums, herbaria, observational data initiatives) have digitized data associated with their work, and increasing numbers now make these data available via the Internet. Even though these are large datasets, most of the records are of low quality.

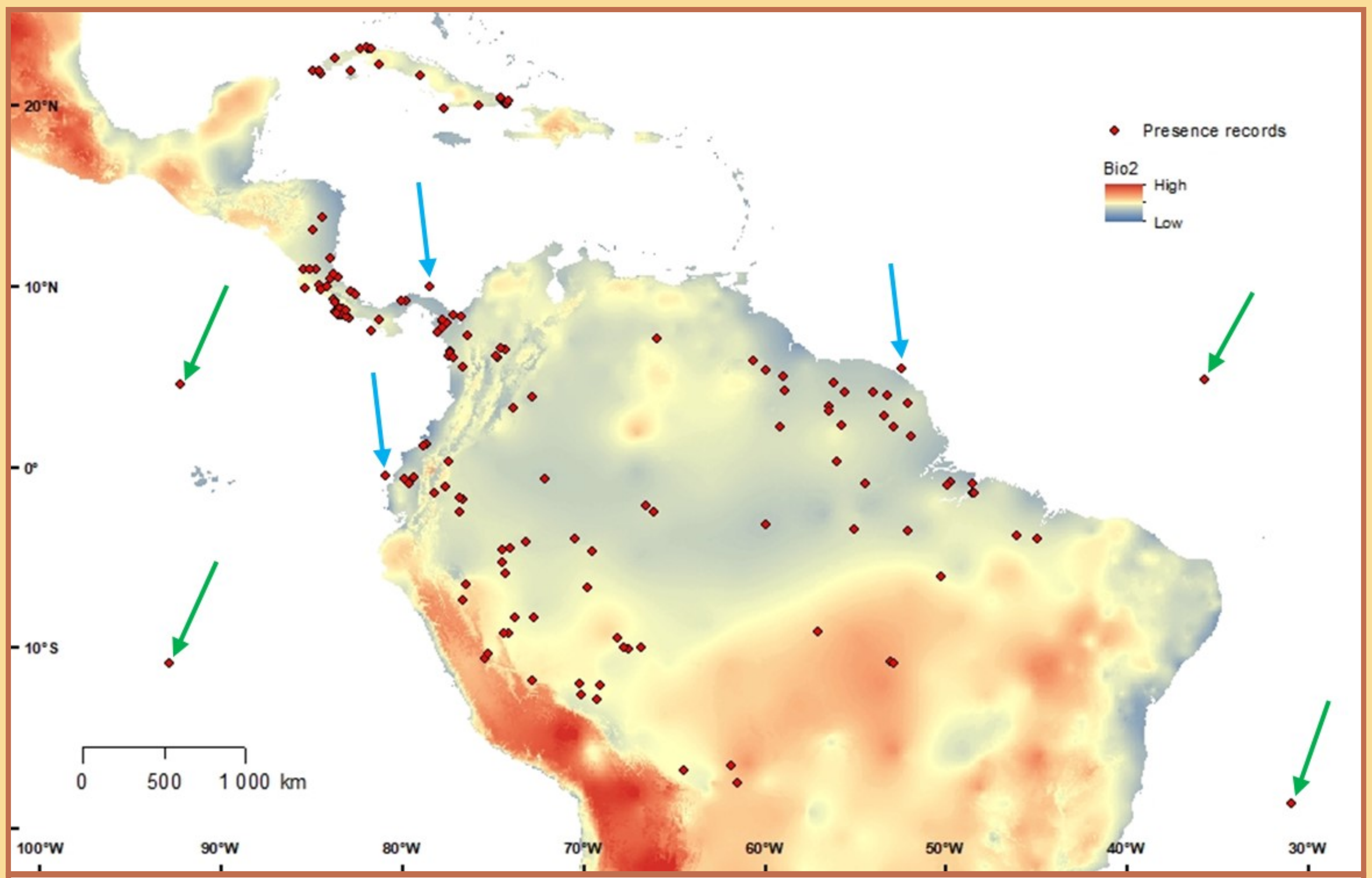


ID	Species	long	lat
1	Cynomys_ludovicianus	-107.58	28.086
2	Cynomys_ludovicianus	-98.08	38.516
3	Cynomys_ludovicianus	0	0
4	Cynomys_arizonensis_	-99.038	38.867
5	Cynomys_ludovicianus	-105	40.195
6	Cynomys_ludovicianus	-100.677	33.885
7	Cynomys_ludovicianus	-99.15	36.439
8	Cynomys_ludovicianus	-104.985	39.739
9	Cynomys_ludovicianus	-101.504	43.174
10	Cynomys_ludovicianus	-108.008	30.245
11	Cynomys_ludovicianus	-104.509	40.36
12	Cynomys_ludovicianus	-100.019	37.759

Occurrence records of a species consist of an ID number, species name and georeference values. They commonly contain diverse errors that should be identified, assessed, and minimized before performing any analysis.

Types of errors found in occurrence data

1. Missing data
2. Repeated observations
3. (0,0) georeference
4. Records with no decimal precision
5. Errors in scientific names
6. Changes in latitude-longitude
7. Records outside the continents
8. Records outside the region of interest
9. Environmental outliers



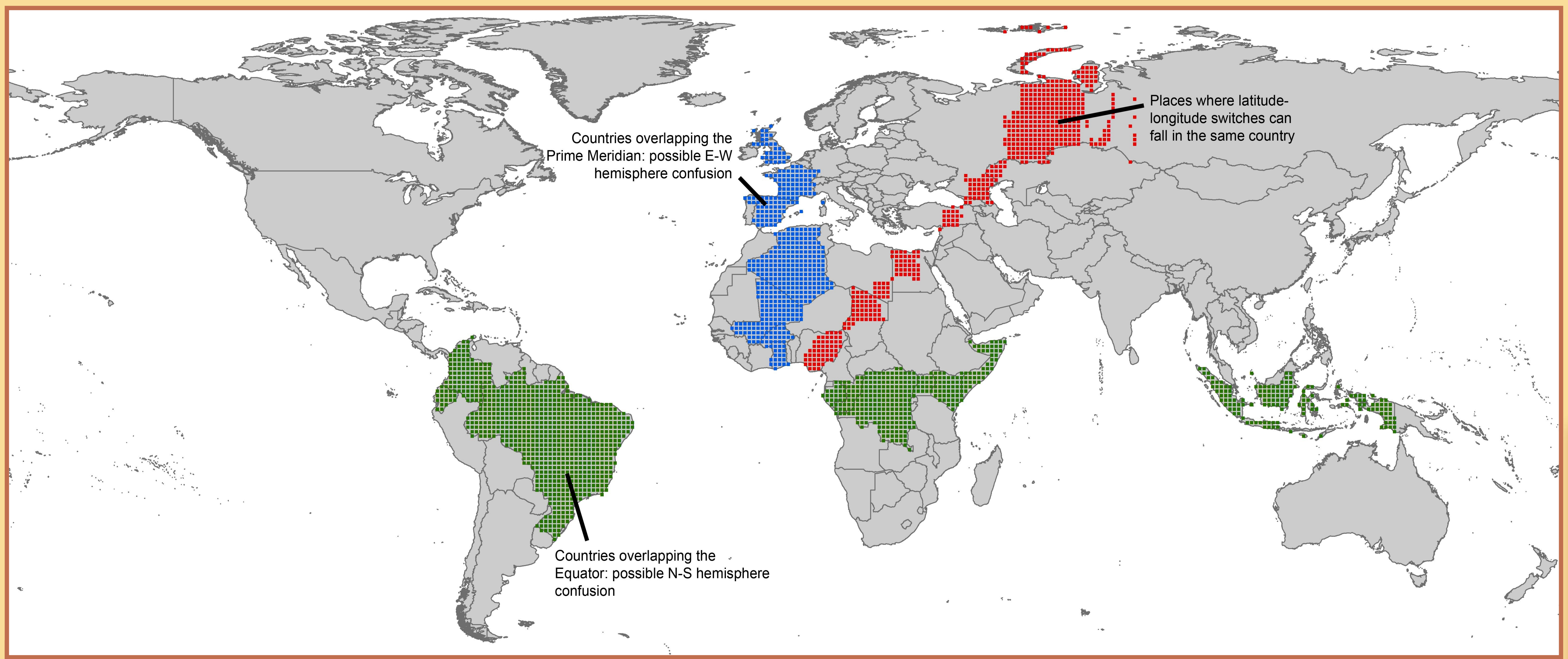
RECORDS OUTSIDE THE CONTINENTS

Sometimes records fall in the ocean (for terrestrial species) and these are easy to detect when plotted together with a map of the study region and a shapefile with the same extent as the environmental predictors to be used for the study. See the figure above where occurrence records for one species (red dots) are represented on top of an environmental predictor (BIO2: mean diurnal range). Occurrences pointed with a green arrow should be erased because they lay too far from the continental area, and those pointed with a blue arrow could be moved to the closer pixel of the environmental predictor.

It is important to notice that often, because of the spatial resolution of environmental layers, some occurrence records will fall outside the extent, even if they were correctly recorded as being on land.

IMPORTANCE OF DATA CLEANING IN BIODIVERSITY STUDIES

- Misidentified records of a species can lead to erroneous interpretation such as thinking that the records provide evidence for an expansion of the distribution or a signal of a biological invasion.
- Duplicated records, geographical outliers and environmental outliers can easily bias the estimation of a species distribution or niche, they might contain unsuitable sites for the species.


Countries overlapping the Prime Meridian: possible E-W hemisphere confusion
Places where latitude-longitude switches can fall in the same country
Countries overlapping the Equator: possible N-S hemisphere confusion

CHANGES IN LATITUDE AND LONGITUDE

Case 1: Latitude and longitude values are switched

1. Plot the occurrences on top of a map of the world or the region of interest.
2. Identify records that appear far from the region of interest, if numerous, they will even create a mirror image of the study area.
3. Once detected, exchange the values of longitude and latitude accordingly, and plot again to corroborate. If occurrences still fall in incongruent places, it is possible that the errors were not of this type, and then, they require a different solution.

Case 2: Changes in sign of latitude-longitude

Follow steps 1 and 2 from Case 1. The mirror effect for this kind of errors is usually seen across the Prime Meridian for longitude values, and across the Equator for latitude values. Once detected, exchange the signs accordingly and corroborate.

This kind of errors can be difficult to detect when working with broadly distributed species, so, we may need extra information about the records in order to fix them. For example, we can use information about the country or state where the species was recorded, if it is available.

The figure shows countries in which it is harder to identify reversed coordinates, both by means of switched latitude and longitude, or coordinates with the sign changed (i.e., + to -, or vice versa).

GEOGRAPHIC OUTLIERS

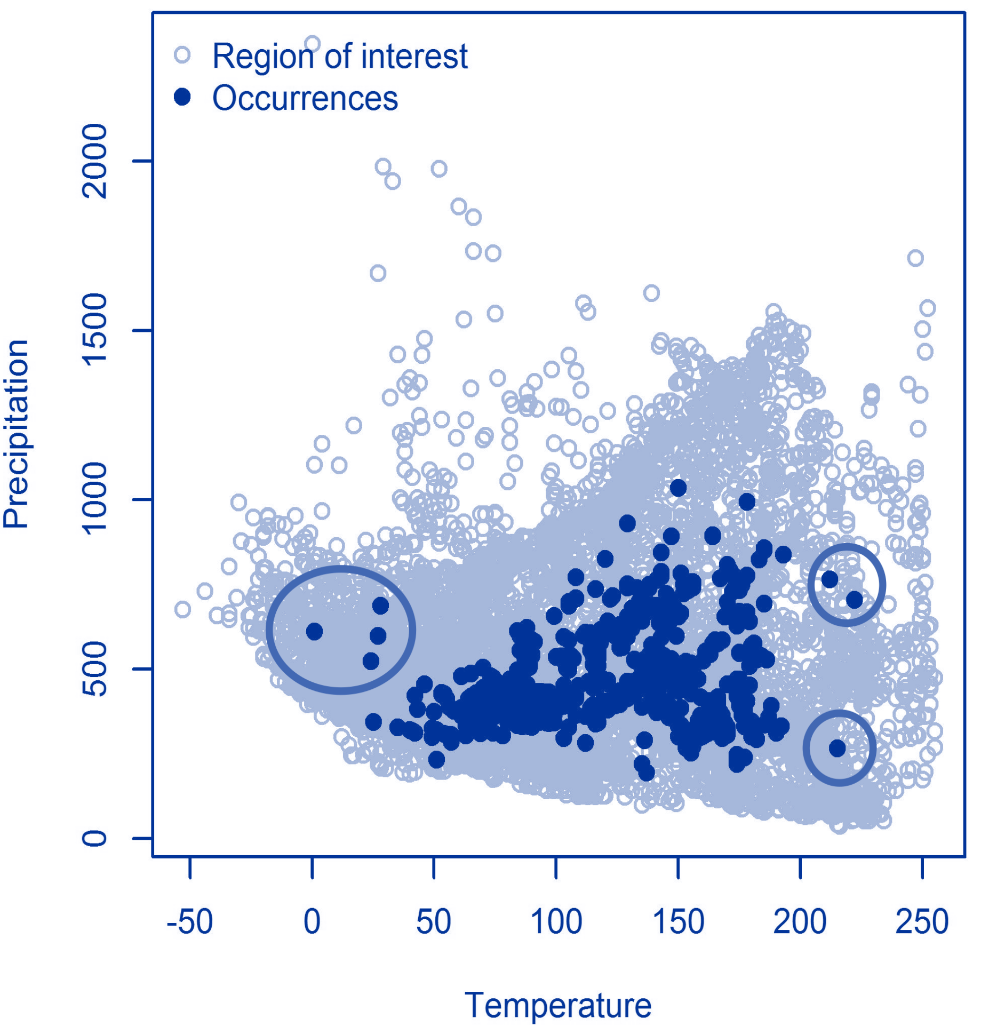
The definition of the region of interest is directly interconnected with the research question. If one is familiar with the species, it is useful to restrict the region of interest to sites to which the species has had access to in a relevant amount of time, and any point outside this region will be considered as a geographic outlier. For example: in a study of the terciopelo snake, *Bothrops asper*, this species has being found from southern Mexico to northern South America, but, if we are interested in assessing snakebite risk by this species in Ecuador, then the region of interest could only include Ecuador and the adjacent countries.

Once the region of interest has been established, it is crucial to double check the coordinates of the occurrences to either erase or fix the latitude and longitude values of suspicious records.

ENVIRONMENTAL OUTLIERS

These errors often come from subtle problems with georeferencig in which a record falls within the known distributional area of the species but in the wrong specific environment. For example, in the mountains, a record that is off by only a few kilometers may fall at a radically different elevation where the climatic features are very different from what the species can tolerate.

1. With GIS software, extract environmental values for each record.
2. Create a large number of random points inside the region of interest and extract its environmental values.
3. Create a scatterplot among small sets of environmental variables and plot the occurrence records on top.
4. Visually assess if any occurrence could be considered as outlier and fix it.



REFERENCE

- Cobos, M.E., Jiménez, L., Nuñez-Penichet, C., Romero-Alvarez, D., Simoes, M. 2018. Sample data and training modules for cleaning biodiversity information. Biodiversity Informatics 13:49–50. https://doi.org/10.17161/bi.v13i0.7600
- Scripts with R code are available in a permanent GitHub repository named ENM_manuals: https://github.com/marlonecobos/ENM_manuals