SANBI-GBIF training workshop in Data Management and Cleaning:

Overview of issues with biodiversity data

Vernon Visser

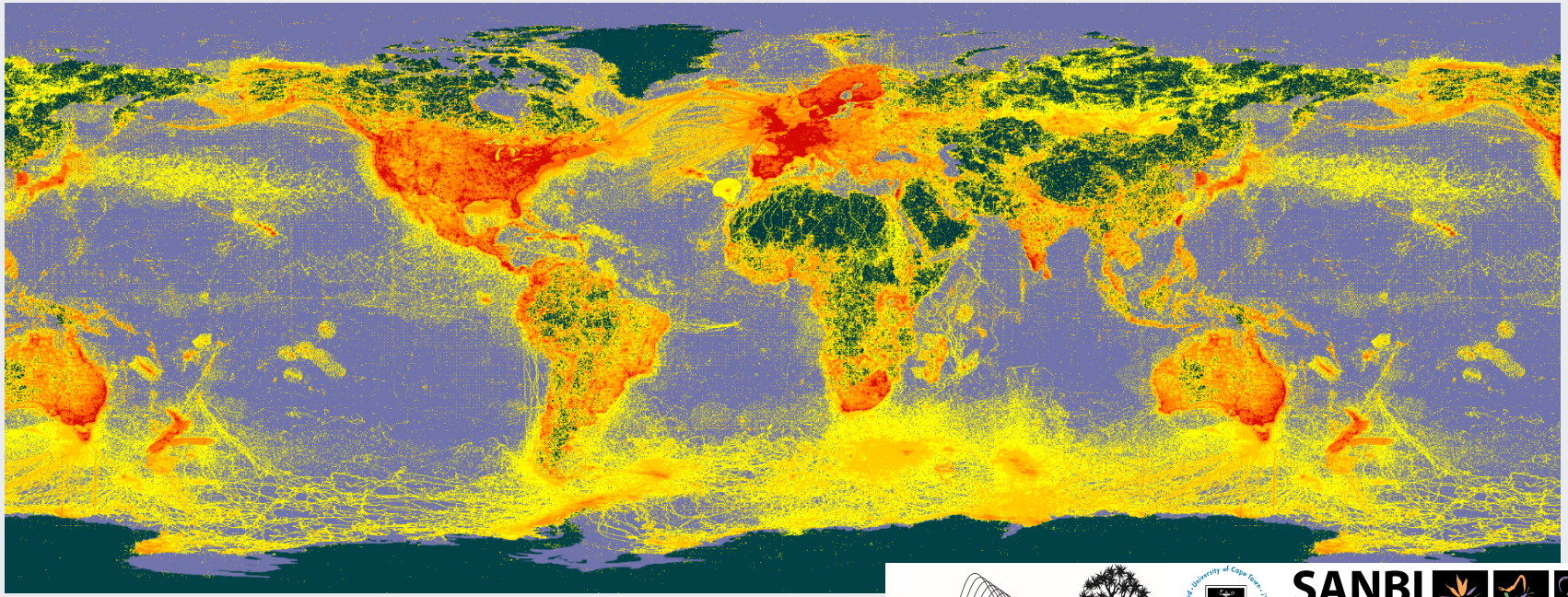
SEEC - Statistics in Ecology, Environment and Conservation
SANBI
Biodiversity for Life
South African National Biodiversity Institute

Issues with biodiversity data

In the initial stages of the establishment of GBIF, there was not much focus on data quality. This has subsequently improved dramatically, but many of these data quality issues are still present in the data but are made explicit using data quality "flags". Moreover, there are inherent biases in biodiversity datasets due to the way in which most of these collections have been and are currently being collected.

In the next few slides we will highlight some of the most noticeable biases. The later sections of this course will then address methods for dealing with some of these biases.

What sorts of biases do you notice just by looking at this recent map of GBIF occurrences?



Summary of issues with biodiversity data

OCCURRENCES

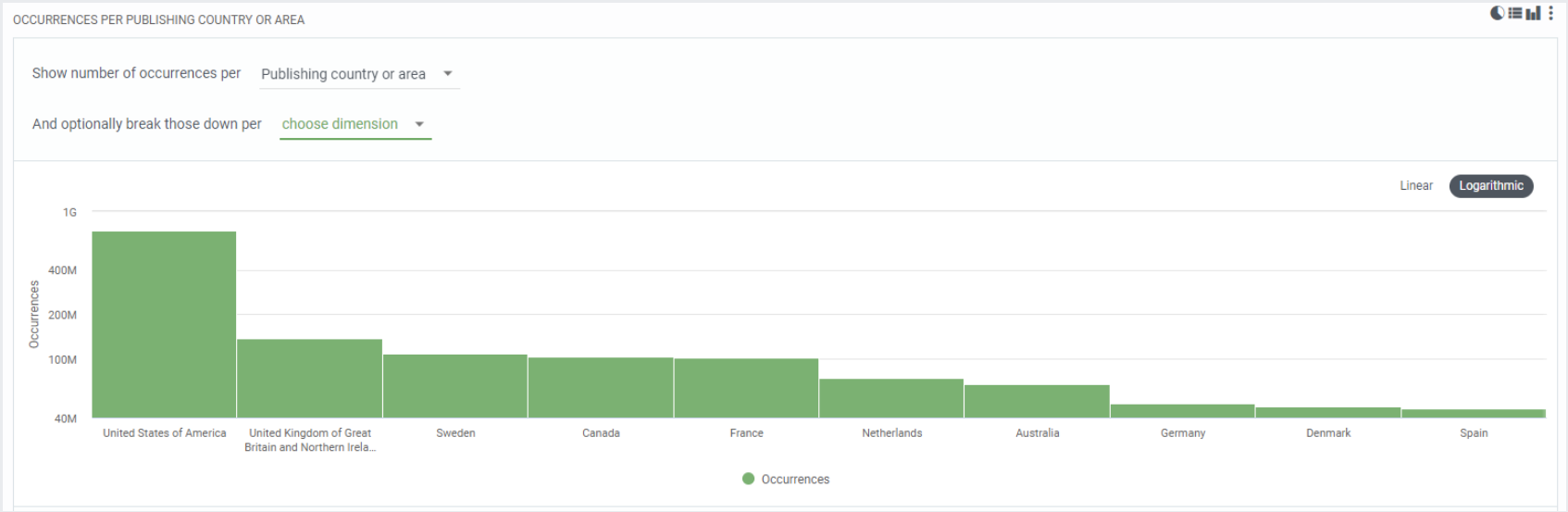
- Sampling biases:
  - Spatial:
    - Countries
    - Roads
    - Population centres
    - Major museums and herbaria
    - Protected areas
    - Altitude / depth
  - Specific projects, surveys or types of data (photos, samples, DNA, tracking, literature)
  - Temporal:
    - Years
    - Seasons
  - Taxonomic
  - Environmental
- Coordinate errors
  - Precision
  - Typos, swaps, nulls

TAXONOMY

BASIS OF RECORD

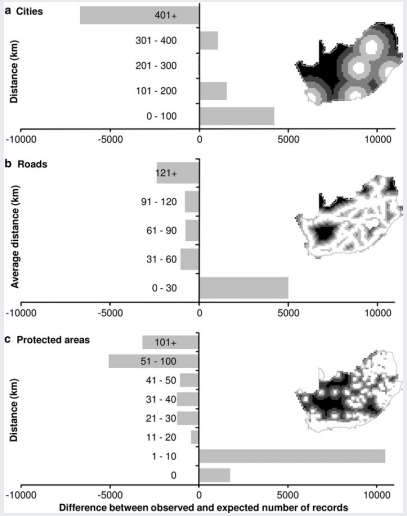
Occurrences - Sampling biases

The most obvious **spatial bias** in most global datasets is the overwhelming number of records in the developed *vs* developing world. Take a look at the statistics of the number of occurrences per country from GBIF below.

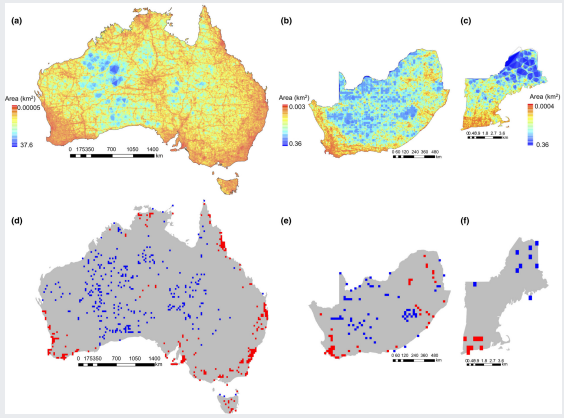


Occurrences - Sampling biases

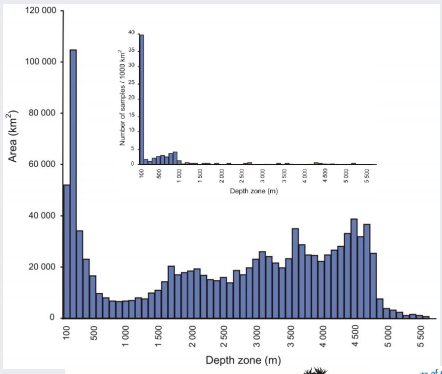
Spatial biases are very common in both older and contemporary biodiversity datasets and largely for similar reasons. The most important cause of sampling biases is accessibility - people tend to make collections in easy to reach localities, be they close to roads or population centres or at lower elevations or shallower depths or close to where people work (major museums and herbaria). We also end to see many records collected in areas where we expect to find lots of biodiversity, e.g. in protected areas or biodiversity hotspots. Here are a few published examples...



Botts et al. (2011) Biodivers Conserv; Griffiths et al. (2008) PLoS ONE



Daru et al. (2017) New Phytologist

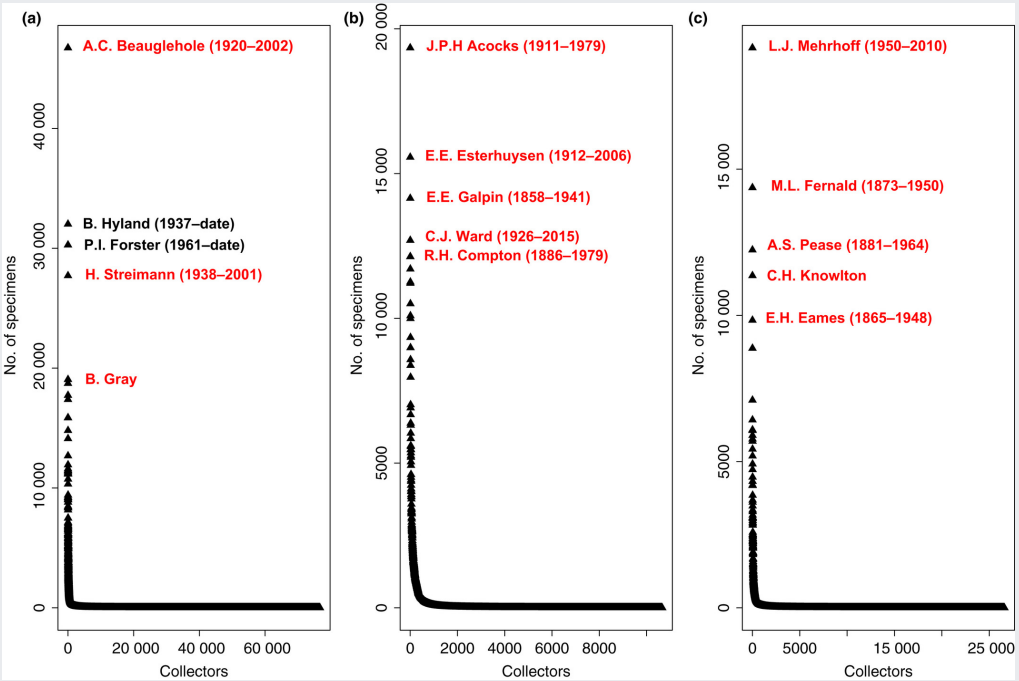


SEEC - Statistics in Ecology, Environment and Conservation

SANBI Biodiversity for Life
South African National Biodiversity Institute

Occurrences - Sampling biases

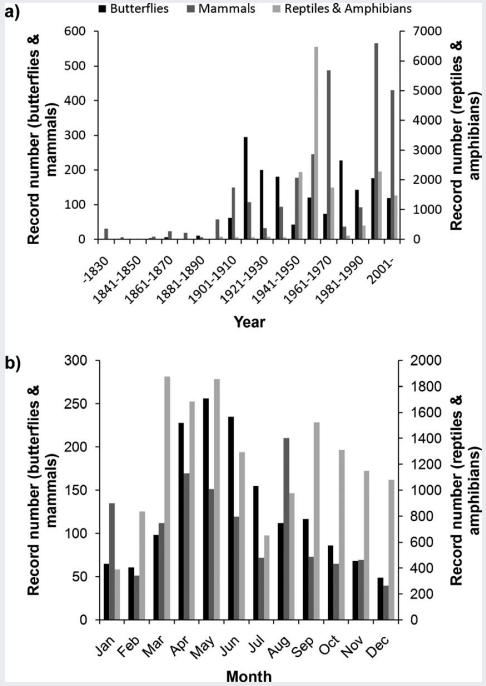
Specific projects, surveys or types of data (photos, samples, DNA, tracking, literature) can be over-represented in biodiversity datasets. One often finds that there are a few "very eager" collectors of particular taxa. Depending on the taxon, one might also find that records are biased because of particular collection methods that are easier or cheaper to do, e.g., citizen science is leading to an explosion of photographic records, but herbarium or museum samples are being made increasingly rarely.



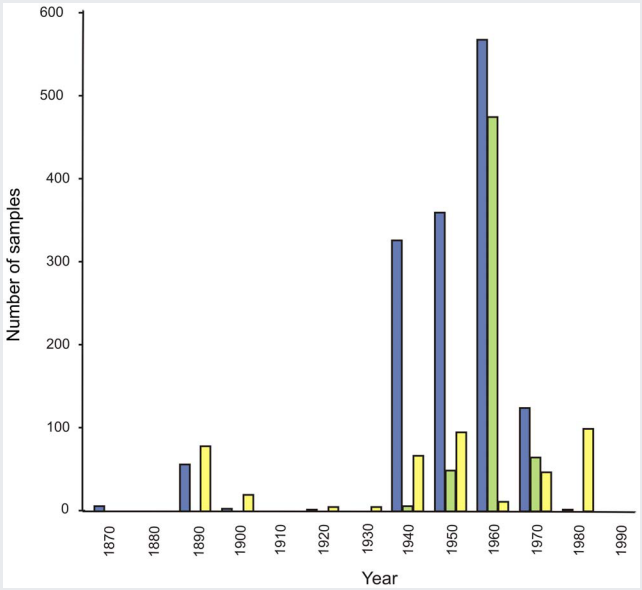
Daru et al. (2017) New Phytologist

Occurrences - Sampling biases

Temporal biases are another common form of bias in biodiversity datasets. For many taxonomic groups there was a "boom" in data collection in the early to mid 1900s (see examples below). More recently though there has been an explosion in the uptake of citizen science data collections. GBIF does actually source data from some of these platforms, the most notable are probably iNaturalist for terrestrial records and OBIS for marine records. This has resulted in many recent records being accumulated, but with largely unique issues, that we will discuss later on.



Newbold et al. (2010) Progress in Physical Geography



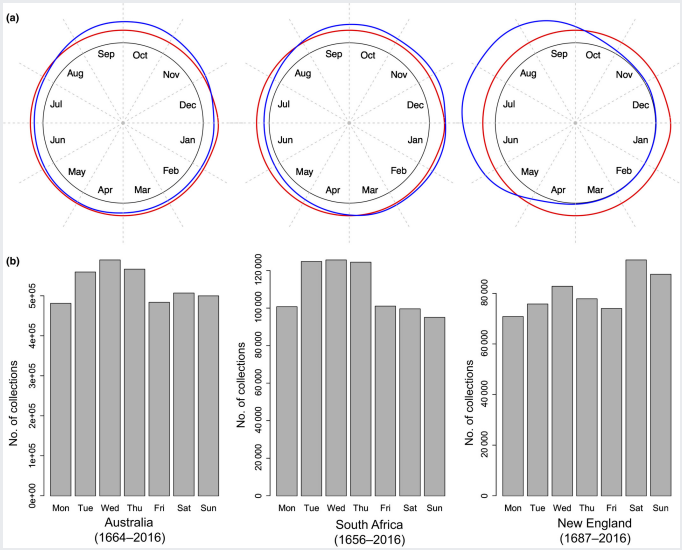
Griffiths et al. (2008) PLoS ONE

SEEC - Statistics in Ecology, Environment and Conservation

SANBI
Biodiversity for Life
South African National Biodiversity Institute

Occurrences - Sampling biases

Another form of **temporal bias** that you may have picked up from the Newbold et al. (2010) figure on the previous slide is a seasonal bias. Depending on the climate of the region in which you are working, you will probably find that there is far more collecting effort during the "growing season" of that region.

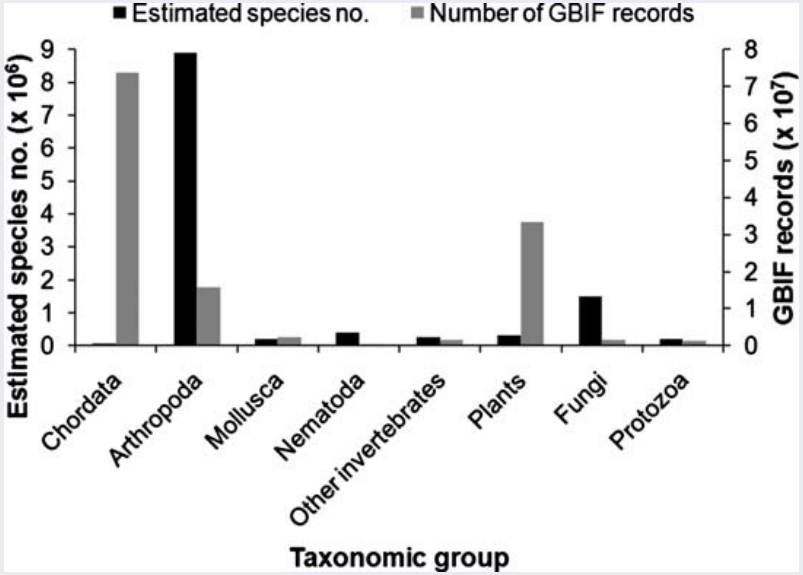


Daru et al. (2017) New Phytologist

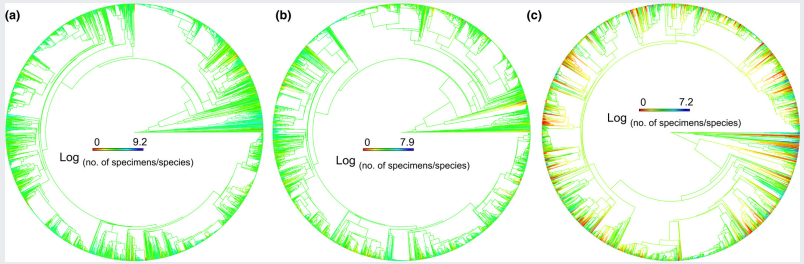
Reasons that temporal biases are worth taking into consideration are (1) that you may be using these biodiversity data together with other environmental data (often spatial layers) that represent specific time periods that only partly overlap with the biodiversity data, or (2) you may want to compare biodiversity in some way between multiple time periods, but there is uneven sampling across these periods.

Occurrences - Taxonomic or phylogenetic biases

Most biodiversity datasets have some form of taxonomic or phylogenetic bias (see examples below). This is often the result of "cute and cuddly" charismatic species being favoured in biodiversity data collections, but may also result from certain sampling methods favouring particular taxa or well-funded projects focusing on a particular group of species.



Newbold et al. (2010) Progress in Physical Geography



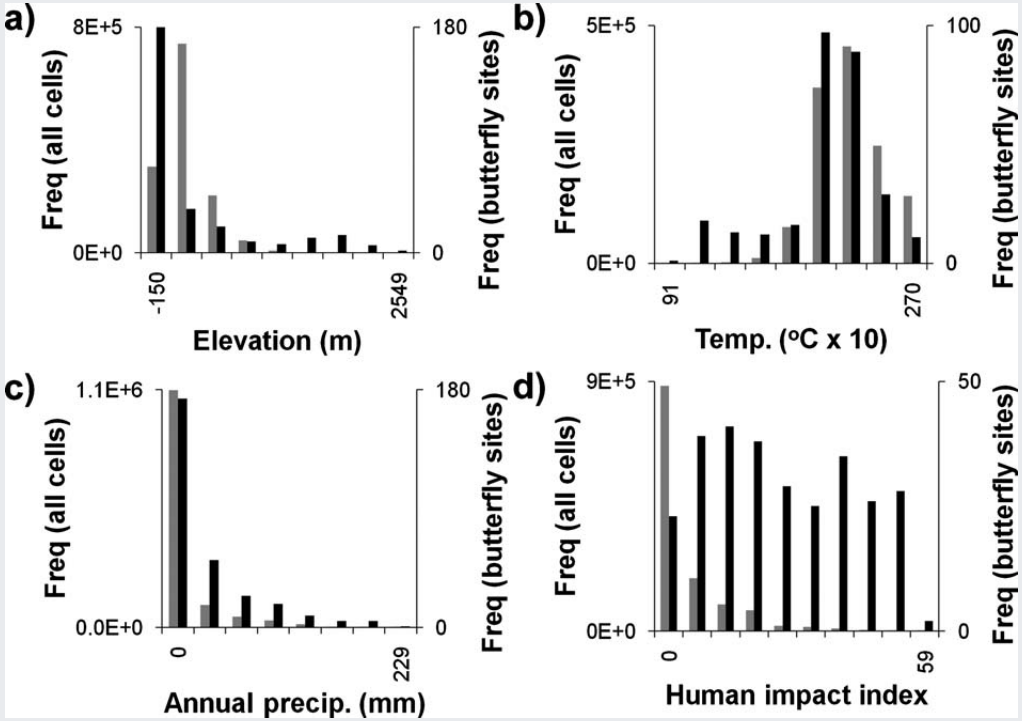
Daru et al. (2017) New Phytologist

SEEC - Statistics in Ecology, Environment and Conservation

SANBI Biodiversity for Life
South African National Biodiversity Institute

Occurrences - Environmental biases

Biodiversity datasets often exhibit environmental biases. These may reflect natural biogeographical patterns of diversity in relation to environmental factors, or can be the result of sampling issues (e.g., hot places are just not very pleasant places to collect samples).



Newbold et al. (2010) Progress in Physical Geography. Black bars show the sampling effort while grey bars show the availability of that environment.

Coordinate errors - Precision

Precision is a measure of how consistent repeated measurements are to each other. It is not the same as accuracy, which reflects the amount of error between measurements and a true value. See the figure to the right for a visual depiction of this difference.

One cannot assume that occurrence coordinates represent the exact location where a specimen was collected. There are numerous reasons why this might be the case:

- Coordinates are derived from non-point localities, e.g., transects, grids, polygons...
- GPS inaccuracies
- Vague locality descriptions
- Coordinates derived from paper maps
- Positional inaccuracies in OpenStreetMap, Google Maps and Google Earth
- Coordinate precision, i.e., numbers of decimal places
- Unknown datum (coordinate reference system, e.g. WGS84)

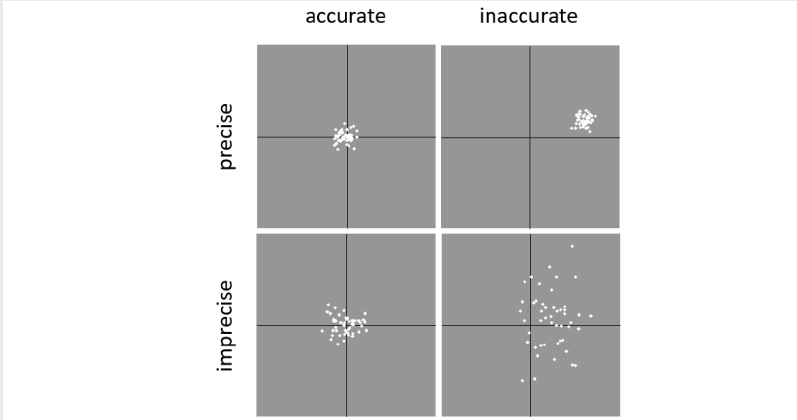
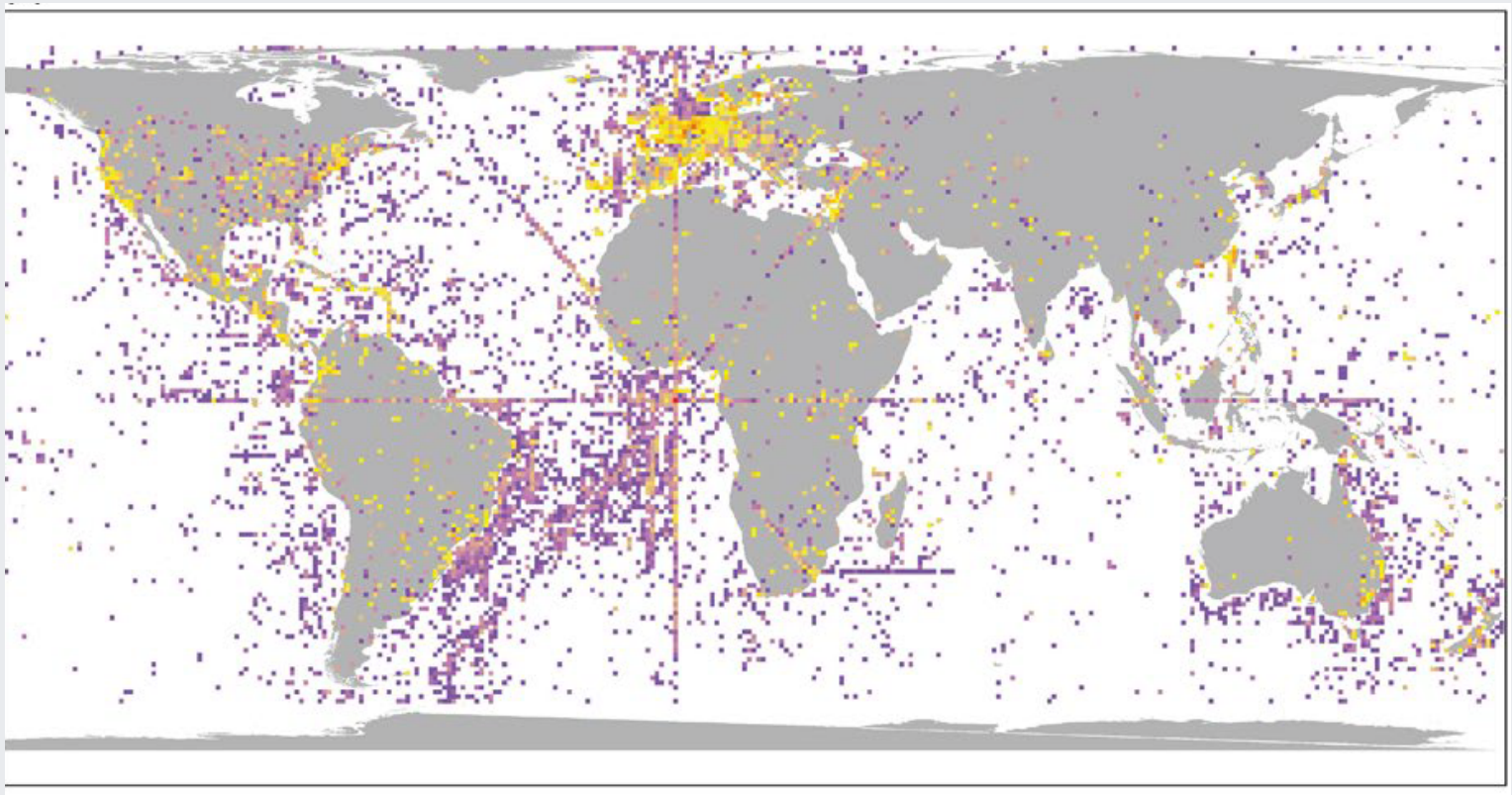


Figure 1. Accuracy versus Precision. Data may be accurate and precise, accurate and imprecise, precise but inaccurate, or both imprecise and inaccurate. Reproduced with permission from Arturo Ariño (2020).

Chapman et al. (2020) Georeferencing Best Practices

Coordinate errors - Typos, swaps, nulls

Many times coordinate errors are just due to pure mistakes. If you take a look at the map of flowering plant records below, you will notice there are strange diagonal and horizontal lines radiating from a point that is actually 0 deg lat and lon. Can you take a guess as to what errors these lines represent?

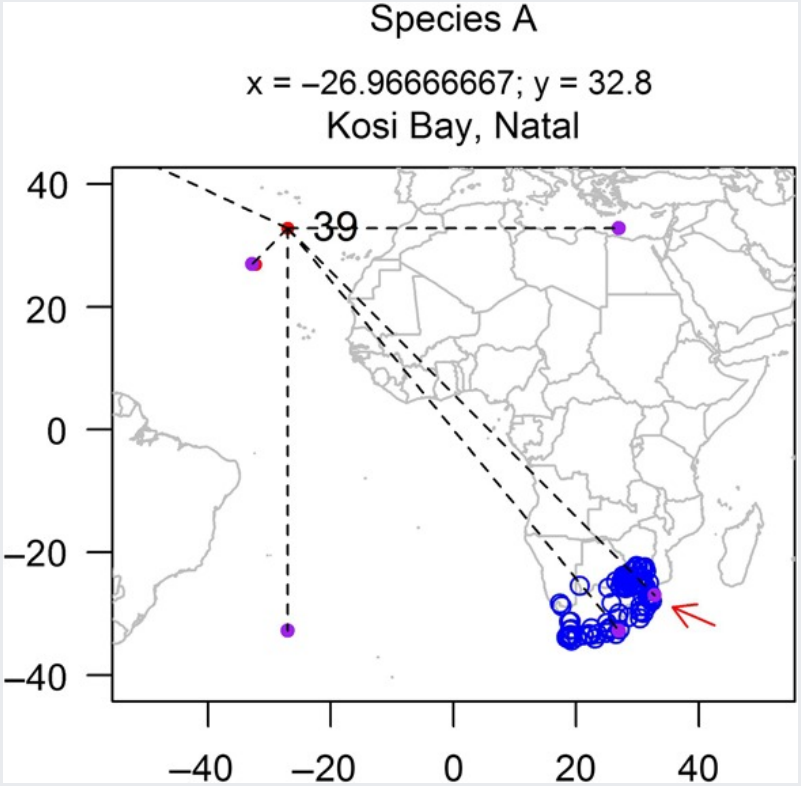


Zizka et al. (2019) Methods Ecol Evol

Coordinate errors - Typos, swaps, nulls

Common mistakes in the recording of coordinates are:

- Swapped latitude and/or longitude
- Missing "-" signs before both or one coordinate
- Zeros to replace missing coordinates (both or just one of the lat and lon)
- Swapped degrees and minutes or minutes and seconds before conversion to decimal degrees



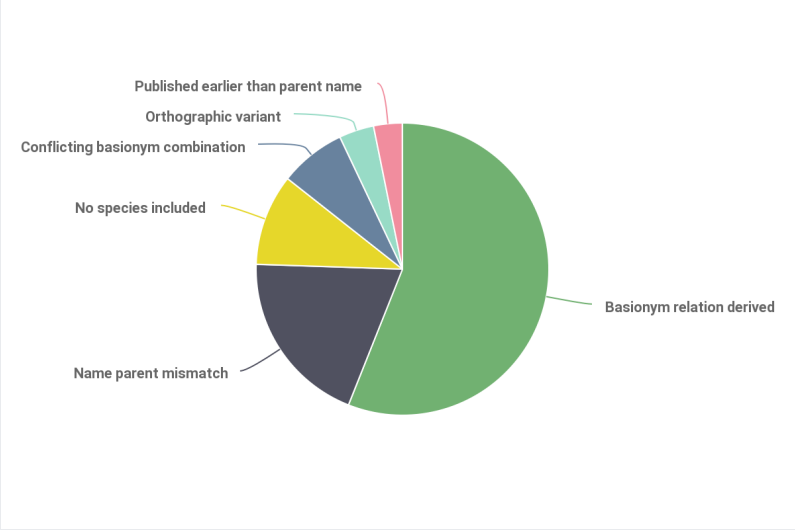
Robertson et al. (2016) Ecography. This figure represents a function in the R biogeo package that highlights possible alternative positions for coordinates based on common mistakes.

Taxonomy

As you are probably already aware, scientific nomenclature groups organisms at various levels of taxonomic hierarchy (family, tribe, genus...). Names are applied to groups of organisms "to promote **stability** and **universability**" in scientific names "and to ensure the name of each taxon is **unique and distinct**".

GBIF relies on an enormous database of taxon names (~3.7M accepted names, ~2.8M synonyms) called the `GBIF Backbone Taxonomy` to standardise names of records that are brought into the database from the many data providers GBIF uses. The GBIF Backbone Taxonomy is actually downloadable and has been assembled from 100 different taxonomic databases.

Even this massive database though is not perfect and when you download data from GBIF you will be alerted in a column corresponding to the relevant taxonomic hierarchical level whether a match has been found or not.



Taxonomy is obviously very important if you are working with datasets from sources other than GBIF and it is critical that you use a methodical workflow to ensure you are working with accepted names.

Basis of record

The last issue I would like to draw your attention to is the basis of a record, i.e. what type of observation the record represents. GBIF (which follows the Darwin Core data standard) recognises the following types of observations:

- Fossils
- Human observation
- Living specimen
- Machine observation
- Material citation (specimens referenced or cited in scholarly publications)
- Material sample (based on samples taken from other specimens or the environment)
- Observation (an occurrence record describing an observation, see below)
- Occurrence (existence of an organism at a particular place at a particular time)
- Preserved specimen (occurrence representing e.g. a herbarium or museum specimen)

This sort of information can be useful depending on your data use case. For example, if you want to determine species richness for an area, excluding fossils is necessary.

Above we have only discussed the GBIF-recognised types for basis of record. However, there have been calls for (and attempts at) additional fields, including whether a specimen is **cultivated or not** and whether it is **native or alien**. We will revisit this issue later on in the course.

Summary

We have seen that there are many issues to be aware of when using data from biodiversity datasets. These may or may not be problematic depending on your use case, but it is good to be aware of these factors because they may influence your results.

In the following sections of the course we will investigate a number of methods to address **some** of these issues. The point of this course is to introduce you to some of the possibilities that are out there so that **you will be comfortable navigating biodiversity datasets in R in the future**.

Table from the GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network

Issue	# records	% of issued	% of total
{No issue}	182276324	--	96.20%
Latitude probably neglected	102702	1.43%	0.05%
Longitude probably neglected	249780	3.47%	0.13%
Latitude and longitude probably transposed	582850	8.10%	0.31%
Coordinates supplied as 0.0 , 0.0	2421605	33.66%	1.28%
Supplied coordinates out of range	206559	2.87%	0.11%
Coordinates fall outside specified country	3915635	54.42%	2.07%
Supplied altitude out of range	277768	3.86%	0.15%
Altitude value suspect	3314	0.05%	<0.01%
Minimum and maximum altitude reversed	13871	0.19%	0.01%
Supplied depth out of range	69	<0.01%	<0.01%
Minimum and maximum depth reversed	26297	0.37%	0.01%
Total issued records	7194999	--	--
Total records	189471323	--	--

Table 1. Overview of error types and quantities issued by the GBIF filter as of December, 2009. The most common combinations are as follows: 1. Latitude and longitude probably transposed + latitude negated = 48964 records. 2. Coordinates fall outside specified country + Latitude and longitude probably transposed = 482831 records. Note: since a record may be affected by more than one geospatial issue, the sum of all issued records is lower than the sum of the differently issued records.