Cleaning species occurrence records

Ian Ondo

2023-05-30

Contents

Package notes:
We need to install the following set of R packages to successfully run the codes in this vignette:

- **CoordinateCleaner** (Zizka et al. 2019)
- **rWCVP** (Brown et al. 2023)
- **dplyr** (Wickham et al. 2023)

Introduction

Cleaning and filtering out data that look spurious or do not fit the purpose of a study is probably the most important step in any data science project. It is generally a good idea to start your data curation by general filters that would exclude observations that do not fit the scope of your study, such as records out of their expected spatial and temporal range. Then, use more specific filters to refine your selection.

In this vignette, we are going to use the World Checklist of Vascular Plants (WCVP) (Govaerts et al. 2021) distribution data and the R package **CoordinateCleaner**, to exclude occurrence locations outside of species known range and common georeferencing errors.

Cleaning records outside of species range

Let's assume that

```
clean0 <- TDWG::rangeCleaner(mySpecies_occurrences,
                    what='occurrences',
                    working_dir=OUTPUT_DIR,
                    status='both',
                    initial_level=2,
                    do.parallel=FALSE,
                    use_name_matching=FALSE,
                    full_data = TRUE,
                    verbose=FALSE)
```

Cleaning unfit and mis-georefenced records

We use metadata information kept from our primary downloads to exclude locations:

- collected before 1945
- with coordinates out of their expected range (i.e. with Lat [-90,90] and Lon [-180,180])
- with coordinates uncertainty <= 20km
- with no decimal place (i.e. rounded)

Then, **CoordinateCleaner** to exclude locations:

- with zero coordinates
- with identical latitude/longitude
- within 10km buffer distance from centroids of countries, provinces, capitals or biodiversity institutions (museums, zoos, botanical gardens, universities)
- within 1km buffer distance from the headquarters of GBIF

```
#-------------------------------------------------------
##= 1. Remove records beyond a certain date
#-------------------------------------------------------
clean <- clean0 %>%
  dplyr::filter(is.na(year) | year >= 1945)

if(nrow(clean)==0) stop('No more records available in the dataset')

#-------------------------------------------------------
##= 2. Remove cultivated specimens
#-------------------------------------------------------
# clean <- clean %>%
#
#   dplyr::filter(is.na(is_cultivated_observation) | is_cultivated_observation == "Yes"
#                 & is.na(basisOfRecord) | basisOfRecord!= "Cultivated habitat")
#
# if(nrow(clean)==0) return(NULL)

#-------------------------------------------------------
##= 3. Remove (un/mis-)georeferenced specimens
#-------------------------------------------------------
clean <- clean %>%
  dplyr::filter(!is.na(decimalLongitude)) %>%
  dplyr::filter(!is.na(decimalLatitude)) %>%

  # Occurences with impossible lat-long coordinates
  dplyr::filter(decimalLatitude < 90 | decimalLatitude > -90 |
                  decimalLongitude < 180 | decimalLongitude > -180)

if(nrow(clean)==0) stop('No more records available in the dataset')

#-------------------------------------------------------
##= 4. Remove coordinates with large uncertainty in m
#-------------------------------------------------------
clean <- clean %>%
  dplyr::filter(coordinateUncertaintyInMeters <= 20000 |
                  is.na(coordinateUncertaintyInMeters))

if(nrow(clean)==0) stop('No more records available in the dataset')
```

```r
#-------------------------------------------------------------
##= 5. Remove coordinates with no decimal place (rounded)
#-------------------------------------------------------------
clean <- clean %>%
  dplyr::filter((decimalLongitude*10) %% 10 != 0) %>%
  dplyr::filter((decimalLatitude*10) %% 10 != 0)

if(nrow(clean)==0) stop('No more records available in the dataset')

#-------------------------------------------------------------
##= 6. Clean records whose individual counts are 0s
#-------------------------------------------------------------
clean <- clean %>%
  dplyr::filter(individualCount > 0 | is.na(individualCount)) %>%
  dplyr::filter(individualCount < 99 | is.na(individualCount))

if(nrow(clean)==0) stop('No more records available information in the dataset')

#=============================================================
##
##=  Start CoordinateCleaner cleaning process
##
#=============================================================
# Rename lon/lat fields
clean <- clean %>%
  dplyr::rename(decimallongitude = 'decimalLongitude', decimallatitude = 'decimalLatitude')

#-------------------------------------------------------------
##= 7. Clean records with 0 long and lat
#-------------------------------------------------------------
clean <- clean %>%
  CoordinateCleaner::cc_zero(buffer = 0.5, verbose=FALSE)

if(nrow(clean)==0) stop('No more records available in the dataset')

#-------------------------------------------------------------
##= 8. Clean records with equal latitude and longitude
#      coordinates, either exact or absolute
#-------------------------------------------------------------
clean <- clean %>%
  CoordinateCleaner::cc_equ(test = "absolute", verbose=FALSE)

if(nrow(clean)==0) stop('No more records available in the dataset')

#-------------------------------------------------------------
##= 9. Clean records outside reported country
#-------------------------------------------------------------
# clean <- clean %>%
#
#   CoordinateCleaner::cc_coun(iso3='countryCode', ref=NULL, verbose=FALSE)
#
# if(nrow(clean)==0) stop('No more records available in the dataset')
```

```
#-------------------------------------------------------
##= 10. Clean records with country and province centroids
#-------------------------------------------------------
clean <- clean %>%
  CoordinateCleaner::cc_cen(buffer = 10000,
                            geod=TRUE,
                            test='both',
                            species='species',
                            ref=NULL,
                            verify=FALSE,
                            verbose=FALSE)

if(nrow(clean)==0) stop('No more records available in the dataset')

#-------------------------------------------------------
##= 11. Clean records in country capitals
#-------------------------------------------------------
clean <- clean %>%
  CoordinateCleaner::cc_cap(buffer = 10000,
                            geod=TRUE,
                            species='species',
                            ref=NULL,
                            verify=FALSE,
                            verbose=FALSE)

if(nrow(clean)==0) stop('No more records available in the dataset')

#-------------------------------------------------------
##= 12. Clean records with institutional coordinates
#-------------------------------------------------------
clean <- clean %>%
  CoordinateCleaner::cc_inst(buffer = 10000,
                             geod=TRUE,
                             species='species',
                             verify=FALSE,
                             verify_mltpl = 10,
                             verbose=FALSE)

if(nrow(clean)==0) stop('No more records available in the dataset')

#-------------------------------------------------------
##= 13. Clean records with GBIF HQ coordinates
#-------------------------------------------------------
clean <- clean %>%
  CoordinateCleaner::cc_gbif(buffer = 1000,
                             geod=TRUE,
                             species='species',
                             verify=FALSE,
                             verbose=FALSE)

if(nrow(clean)==0) stop('No more records available in the dataset')

#=======================================================
```

```r
##
##=  End CoordinateCleaner cleaning process
##
#=============================================================
# Rename lon/lat fields
clean <- clean %>%
  dplyr::rename(decimalLongitude = 'decimallongitude', decimalLatitude = 'decimallatitude')
```

References

Brown, Matilda J. M., Barnaby E. Walker, Nicholas Black, Rafaël Govaerts, Ian Ondo, Robert Turner, and Eimear Nic Lughadha. 2023. "rWCVP: A Companion r Package to the World Checklist of Vascular Plants." *New Phytologist*.

Govaerts, R., E. Nic Lughadha, N. Black, R. Turner, and A. Paton. 2021. "The World Checklist of Vascular Plants, a Continuously Updated Resource for Exploring Global Plant Diversity." Journal Article. *Sci Data* 8 (1): 215. https://doi.org/10.1038/s41597-021-00997-6.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zizka, Alexander, Daniele Silvestro, Tobias Andermann, Josue Azevedo, Camila Duarte Ritter, Daniel Edler, Harith Farooq, et al. 2019. "CoordinateCleaner: Standardized Cleaning of Occurrence Records from Biological Collection Databases." *Methods in Ecology and Evolution*, no. 10: –7. https://doi.org/10.1111/2041-210X.13152.