



INSTRUCCIONES PARA USO (ES)

OCCUR <https://ecoinformatic.shinyapps.io/OCCUR/> es una aplicación que guía paso a paso el proceso de selección de filtros para la limpieza y validación de registros de especies disponibles. Este flujo de trabajo interactivo ayudará al usuario en la selección de datos entre múltiples opciones disponibles según su caso de estudio.

OCCUR app




- Basis of Record
- Taxonomic
- Geographic
- Temporal
- Duplicates
- Final Report
- References
- About

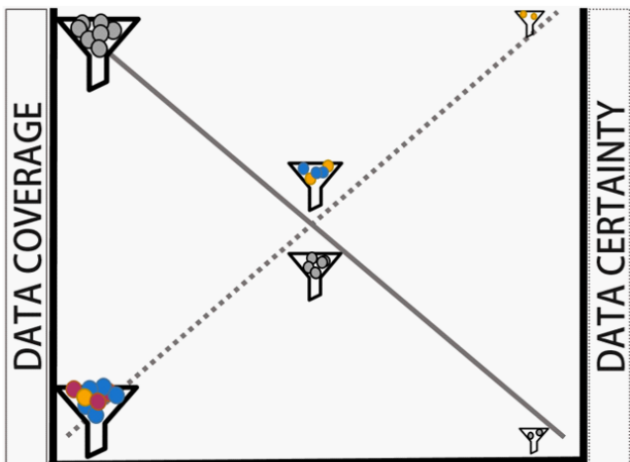


OCCUR app is a "step by step" guide that goes over 5 different modules to curate biodiversity data records. It was created to facilitate the process of filtering, cleaning and validating occurrence species records from data repositories. This interactive workflow will help the user in the selection of data records between all possibilities depending on their study case, considering their pros and cons. Each module will also display how data certainty and data coverage change when selecting different scenarios of the application of filtering and cleaning rules.

INSTRUCTIONS

1. Choose a module of the 5 available in the left panel.
2. Select between filters / steps in left-upper box (there are no previous selections marked).
3. Check the "Trade-off" table that will display with each selection in the right-upper box (left panel).
4. Check the "Methods" table that will display with each selection in the right-upper box (middle panel).
5. Check and copy the "R Code" table that will display with each selection in the right-upper box (right panel).
6. See the bibliography associated in the "References" panel.
7. Check how certainty and data coverage varies with each selection in the left-bottom panel to make your final selection. Values goes from 0 (minimum certainty or data coverage available) to 1 (maximum certainty or data coverage available).
8. Download the final guide to process data and write the methods section based on the selected steps by module in the "Final report" tab.





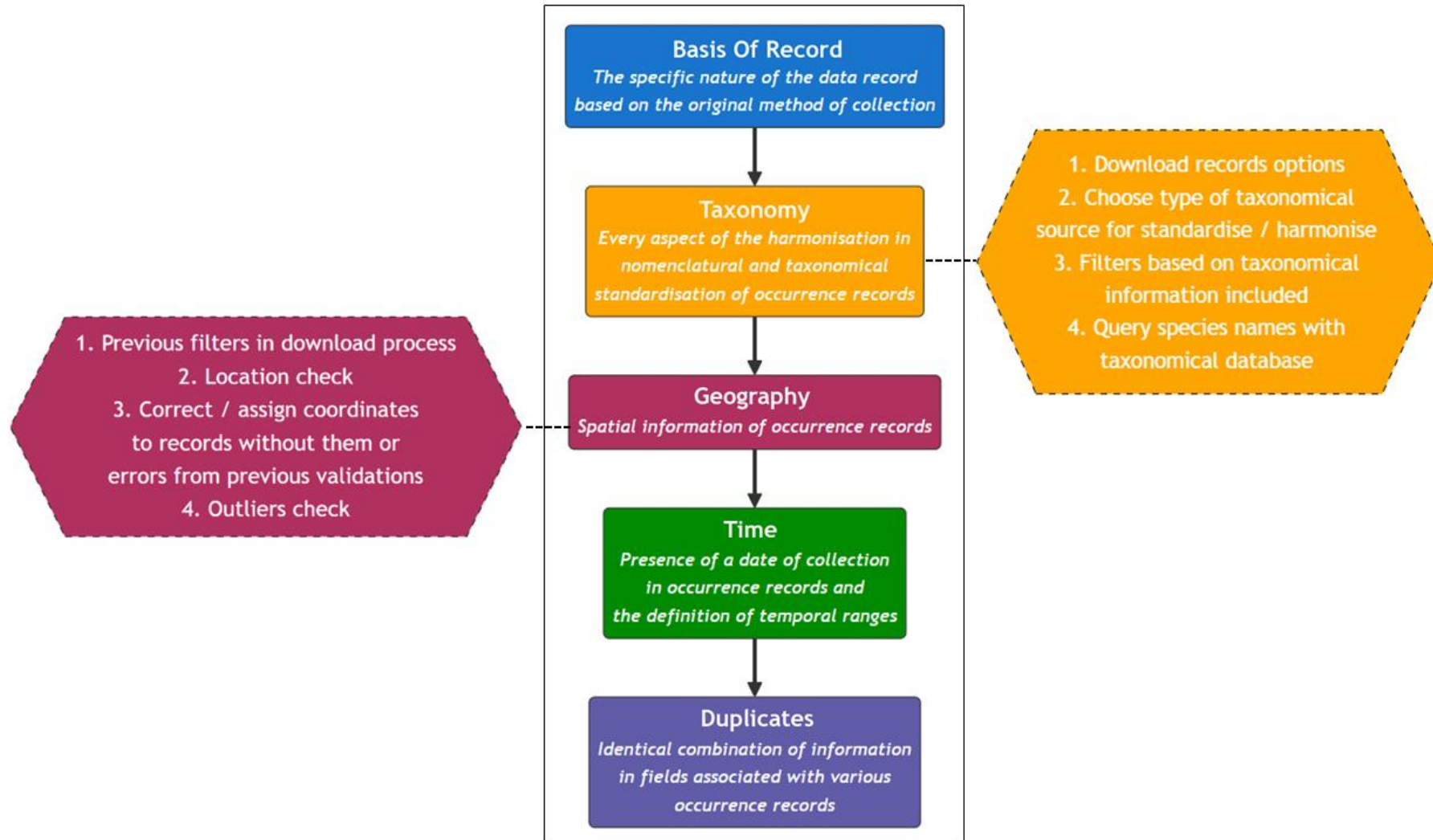
DATA COVERAGE

DATA CERTAINTY

FILTER STRICTNESS


Variation in the number of records available data coverage (continuous line) and data certainty estimates (dotted line) based on precision and accuracy of records information

OCCUR presenta 5 módulos diferentes según la dimensión a tratar en los datos de biodiversidad.



1. Elige un módulo entre los 5 disponibles en el panel izquierdo.

OCCUR app



Basis of Record

Taxonomic

Geographic


Temporal

Duplicates

Final Report

References

About



OCCUR app is a "step by step" guide that goes over 5 different modules to curate biodiversity data records. It was created to facilitate the process of filtering, cleaning and validating occurrence species records from data repositories. This interactive workflow will help the user in the selection of data records between all possibilities depending on their study case, considering their pros and cons. Each module will also display how data certainty and data coverage change when selecting different scenarios of the application of filtering and cleaning rules.

INSTRUCTIONS

1. Choose a module of the 5 available in the left panel.

2. Select between filters / steps in left-upper box (there are no previous selections marked).

3. Check the "Trade-off" table that will display with each selection in the right-upper box (left panel).

4. Check the "Methods" table that will display with each selection in the right-upper box (middle panel).

5. Check and copy the "R Code" table that will display with each selection in the right-upper box (right panel).

6. See the bibliography associated in the "References" panel.

7. Check how certainty and data coverage varies with each selection in the left-bottom panel to make your final selection. Values goes from 0 (minimum certainty or data coverage available) to 1 (maximum certainty or data coverage available).

8. Download the final guide to process data and write the methods section based on the selected steps by module in the "Final report" tab.


Basis Of Record

Taxonomy

Geography

Time

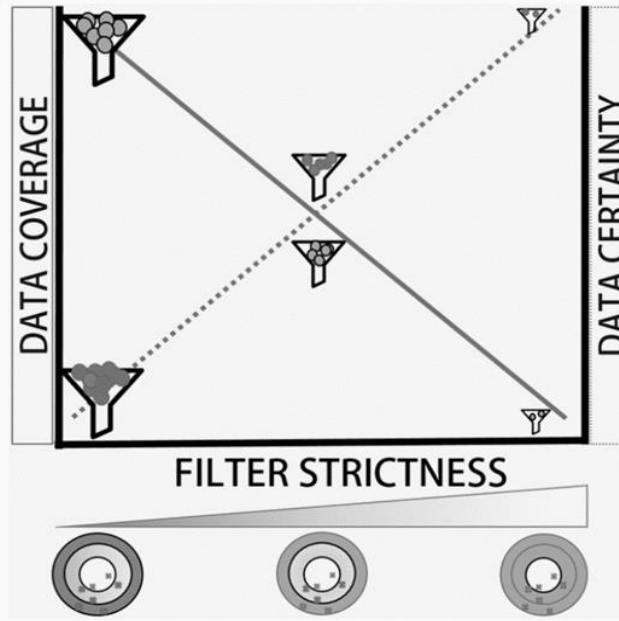
Duplicates




DATA COVERAGE

DATA CERTAINTY

FILTER STRICTNESS






Variation in the number of records available data coverage (continuous line) and data certainty estimates (dotted line) based on precision and accuracy of records information

2. Selecciona entre los filtros / pasos disponibles en el panel superior-izquierdo (ninguna opción está seleccionada previamente).

OCCUR app

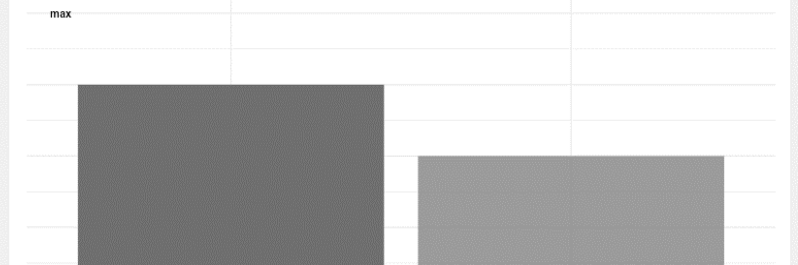


- Basis of Record
- Taxonomic
- Geographic
- Temporal
- Duplicates
- Final Report
- References
- About

In order to validate records geographically, the user may choose what type of data needs before download (this step can also be taken after the download process by filtering the dataset)

Choose between:

- ☐ Do not apply previous filters
- ☐ Only records with coordinates
- ☒ Only records with coordinates filtered by spatial extent (area or administrative units)
- ☐ Only records without known coordinate issues



Trade-off Methods R Code

Pros	Cons
Depending on the selected area, can exclude unreliable records (e.g. zero coordinates, sea points, etc.)	Excludes records from suitable/native regions not considered by bibliography.
Excludes records from introduced distributional areas.	Excludes georeferenciable records by locality information that could be repaired.
Less time of manipulation than download all available records with coordinates.	
More information available including records labelled as 'with coordinates issues'.	

Your selection:

- *Applying previous filter: TRUE
- *Checking coordinates precision:
- *Checking coordinates value:
- *Checking coordinates position:
- *Recovering coordinates:
- *Detecting distributional outliers:
- *Detecting environmental outliers:

3. Valida la información disponible con cada selección en la tabla "Trade-off" en la primera pestaña del panel superior-derecho.

OCCUR app

- Basis of Record
- Taxonomic
- 1. Download records options
 - 2. Choose type of taxonomical source for standardize / harmonize
 - 3. Filters based on taxonomical information included
 - 4. Query species names with taxonomical database
- Geographic
- Temporal
- Duplicates
- Final Report
- References
- About

In order to validate records geographically, the user may choose what type of data needs before download (this step can also be taken after the download process by filtering the dataset)

Choose between:

- ☐ Do not apply previous filters
- ☐ Only records with coordinates
- ☒ Only records with coordinates filtered by spatial extent (area or administrative units)
- ☐ Only records without known coordinate issues

SOURCE

Trade-off
Methods
R Code


Pros	Cons
Depending on the selected area, can exclude unreliable records (e.g. zero coordinates, sea points, etc.)	Excludes records from suitable/native regions not considered by bibliography.
Excludes records from introduced distributional areas.	Excludes georeferenciable records by locality information that could be repaired.
Less time of manipulation than download all available records with coordinates.	
More information available including records labelled as 'with coordinates issues'.	

Your selection:

- *Applying previous filter: TRUE
- *Checking coordinates value:
- *Recovering coordinates:
- *Detecting environmental outliers:
- *Checking coordinates precision:
- *Checking coordinates position:
- *Detecting distributional outliers:

4. Valida la información disponible en la tabla "Methods" que se muestra en la segunda pestaña del panel superior-derecho.

OCCUR app



Basis of Record

Taxonomic

Geographic

» 1. Previous filters in download process

» 2. Location check

» 3. Correct / assign coordinates to records without them or errors from previous validations

» 4. Outliers check

Temporal

Duplicates

Final Report

References

About

Use distributional information

Use environmental information

Choose between:

☐ a. Calculate environmental centroids for the species and validate outliers.

☒ b. Calculate environmental space for each species, check whether records overlap with it, and delete outliers.

☐ c. Overlap environmental information by geographical position and filter occurrences by threshold.

☐ d. Do not apply a filter for environmental outliers

max

min

CertaintyData coverage

SOURCE

Trade-offMethodsR code

Use distributional information

Use environmental information

Point in polygon function using range maps shapefiles.

CoordinateCleaner::cc_lucn

CoordinateCleaner::cc_outl [17]

Bracatus' R package [23]

Your selection:

*Applying previous filter: TRUE

*Checking coordinates precision:

*Checking coordinates value:

*Checking coordinates position:

*Recovering coordinates:

*Detecting distributional outliers: TRUE

*Detecting environmental outliers: TRUE

5. Revisa la bibliografía asociada en el panel "References". Para acceder al documento original haz click en 'See ref' y se abrirá el enlace en tu buscador.

OCCUR app
SOURCE

Basis of Record

Taxonomic <

Geographic <

>> 1. Previous filters in download process

OCCUR app

Home

Basis of Record

Taxonomic <

Geographic <

>> 1. Previous filters in download process

>> 2. Location check

>> 3. Correct / assign coordinates to records without them or errors from previous validations

>> 4. Outliers check

Temporal

Duplicates

Final Report

References

About

☐ Use distributional information
 ☒ Use environmental information

Choose between:

- ☐ a. Calculate environmental centroids for the species and validate outliers.
- ☒ b. Calculate environmental space for each species, check whether records overlap with it, and delete outliers.
- ☐ c. Overlap environmental information by geographical position and filter occurrences by threshold.
- ☐ d. Do not apply a filter for environmental outliers

Trade-off
Methods
R code

Use distributional information
Use environmental information

Point in polygon function using range maps shapefiles.

```
CoordinateCleaner::cc_lucn
CoordinateCleaner::c_outl [17]
Bracatus' R package [23]
```

[1] Jin, J. & Yang, J. (2020). BDCleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases. *Global Ecology and Conservation*, 21, e00852, ISSN 2351-9894. See ref

[2] Speed JDM, Bendiksby M, Finstad AG, Hassel K, Kolstad AL, et al. (2018) Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. *PLOS ONE* 13(4): e0196417. See ref

[3] Shirey, V., Belitz, M.W., Barve, V. and Guralnick, R. (2021), A complete inventory of North American butterfly occurrence data: narrowing data gaps, but increasing bias. *Ecography*, 44: 537-547 See ref

[4] Tiago, P., Cela-Hasse, A., Marques, T.A. et al. (2017). Spatial distribution of citizen science casuistic observations for different taxonomic groups. *Sci Rep* 7, 12832 See ref

[5] Feng, X., Park, D.S., Walker, C. et al. (2019). A checklist for maximizing reproducibility of ecological niche models. *Nat Ecol Evol* 3, 1382-1395 See ref

[6] Tessarolo, G., Ladle, R., Rangeli, T. & Hortal, J. (2017). Temporal degradation of data limits biodiversity research. *Ecology and Evolution*; 7:6863-6870. See ref

[7] Stropp, J., Ladle, R., Malhado, A., Hortal, J., Gauffri, A., Temperley, W., Olav Skoien, J. & Mayaux, P. (2016). Mapping ignorance: 300 years of collecting flowering plants in Africa: 300 Years of collecting flowering plants in Africa. *Global Ecology and Biogeography* 25 (9): 1085-1098 See ref

[8] Meyer, C., Weigelt, P. & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* 19 (8): 992-1006 See ref

[9] Menegotto, A. & Rangeli, T.F. (2018). Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness. *Nature Communications* 9, 4713. See ref

[10] Feeley, K.J. & Silman, M.R. (2010). Modelling the responses of Andean and Amazonian plant species to climate change: the effects of georeferencing errors and the importance of data filtering. *Journal of Biogeography*, 37: 733-740. See ref

[11] Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. (2017). Taxonomic bias in biodiversity data and societal preferences. *Scientific Reports* 7, 9132 See ref

[12] Grenie, M., Berté, E., Carvajal-Quintero, J., Dadlow, G. M., Sagouis, A. & Winter, M. (2022). Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. *Methods in Ecology and Evolution*, 00, 1-14. See ref

[13] Vandepitte, L., Bosch, S., Tyberghein, L., Waumans, F., Vanhoorne, B., Hernandez, F., De Clerck, O. & Mees, J. (2015). Fishing for data and sorting the catch: assessing the data quality, completeness and fitness for use of data in marine biogeographic databases. *Database Vol. 2014*: article ID bau125; See ref

[14] Chapman, A.D. (2005). Principles and methods of data cleaning - Primary species and species occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. See ref

[15] Serra-Diaz, J.M., Enquist, B.J., Maitner, B. et al. Big data of tree species distributions: how big and how good?. *For. Ecosyst.* 4, 30 (2017). See ref

[16] Meiri, S. (2018). The smartphone fallacy - when spatial data are reported at spatial scales finer than the organisms themselves. *Frontiers of Biogeography*, 10(1-2). Retrieved from https://escholarship.org/uc/item/2n3349jg See ref

[17] Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., (...) Antoniou, A. (2021). CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*. -7. R package version 2.0-20, URL: https://github.com/ropensci/CoordinateCleaner. See ref

[18] Ribeiro, B.R., Velasco, S.J., Guidoni-Martins, K., Tessarolo, G., Jardim, L., Bachman, S.P. & Loyola, J. (2022). bdc: A toolkit for standardizing, integrating and cleaning biodiversity data. *Methods in Ecology and Evolution*, 00, 1-8. See ref

[19] Robertson, M.P., Visser, V. and Hui, C. (2016). Biogeom: An R package for assessing and improving data quality of occurrence record datasets. *Ecography*, 39: 394-401. See ref

[20] Tessarolo, G., Ladle, R., Lobo, J.M., Rangeli, T. & Hortal, J. (2021). Using maps of biogeographical ignorance to reveal the uncertainty in distributional data hidden in species distribution models. *Ecography*, 44, 1743-1755. See ref


[21] de Lima, R. A. F., Sanchez-Tapia, A., Mortara, S. R., ter Steege, H., & de Siqueira, M. F. (2021). plantR: An R package and workflow for managing species records from biological collections. *Methods in Ecology and Evolution*, 00, 1-8. See ref

[22] Park, D. S., Xie, Y., Thammavong, H. T., Tulaiha, R., & Feng, X. (2022). Artificial Hotspot Occurrence Inventory (AHOI). *Journal of Biogeography*, 00, 1-9 See ref

[23] Arle E, Zizka A, Keil P, et al. bRacatus: A method to estimate the accuracy and biogeographical status of georeferenced biological data. *Methods Ecol Evol.* 2021;12: 1609-1619 See ref

6. Muchas opciones disponibles en OCCUR muestran un ejemplo simple de código R disponible para copiar en el portapapeles e integrar en los scripts de los usuarios. La tabla está disponible en la tercera pestaña del panel superior-derecho.

OCCUR app



Basis of Record

Taxonomic

Geographic

>> 1. Previous filters in download process

>> 2. Location check

>> 3. Correct / assign coordinates to records without them or errors from previous validations

>> 4. Outliers check

Temporal

Duplicates

Final Report

References

About

Use distributional information

Use environmental information

Choose between:

☐ a. Calculate environmental centroids for the species and validate outliers.

☒ b. Calculate environmental space for each species, check whether records overlap with it, and delete outliers.

☐ c. Overlap environmental information by geographical position and filter occurrences by threshold.

☐ d. Do not apply a filter for environmental outliers

max

min

Certainty

Data coverage

SOURCE

Trade-off

Methods

R code

Copy

library(sf) # Point in polygon analysis

Upload dataframe with occurrences (occData) and shapefile of administrative units (countriesSHP)

datapoints <- st_as_sf(x = occData, coords = c('decimalLongitude', 'decimalLatitude'), crs = '+proj=longlat +datum=WGS84 +no_defs')

occData <- st_join(datapoints, countriesSHP)

Your selection:

*Applying previous filter: TRUE

*Checking coordinates precision:


*Checking coordinates value:

*Checking coordinates position:

*Recovering coordinates:


*Detecting distributional outliers: TRUE

*Detecting environmental outliers: TRUE



7. Valida como varía la certidumbre y disponibilidad de los datos con cada selección en el panel inferior-izquierdo. Los valores mostrados van desde el mínimo al máximo disponible según la bibliografía indicada.

OCCUR app



Basis of Record

Taxonomic

Geographic

>> 1. Previous filters in download process

>> 2. Location check

>> 3. Correct / assign coordinates to records without them or errors from previous validations

>> 4. Outliers check

Temporal

Duplicates

Final Report

References

About

Use distributional information

Use environmental information

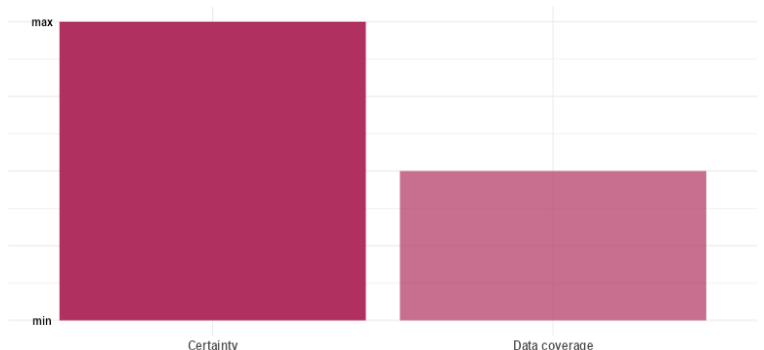
Choose between:

☐ a. Calculate environmental centroids for the species and validate outliers.

☒ b. Calculate environmental space for each species, check whether records overlap with it, and delete outliers.

☐ c. Overlap environmental information by geographical position and filter occurrences by threshold.

☐ d. Do not apply a filter for environmental outliers



Category	Value
Certainty	max
Data coverage	Medium

Trade-off

Methods

R code

Copy

library(sf) # Point in polygon analysis

Upload dataframe with occurrences (occData) and shapefile of administrative units (countriesSHP)

datapoints <- st_as_sf(x = occData, coords = c('decimalLongitude', 'decimalLatitude'), crs = '+proj=longlat +datum=WGS84 +no_defs')

occData <- st_join(datapoints, countriesSHP)

Your selection:

*Applying previous filter: TRUE

*Checking coordinates precision:

*Checking coordinates value:

*Checking coordinates position:


*Recovering coordinates:

*Detecting distributional outliers: TRUE

*Detecting environmental outliers: TRUE

8. Valida las opciones marcadas en cada módulo mirando el panel inferior-derecho.

OCCUR app



Basis of Record

Taxonomic

Geographic

>> 1. Previous filters in download process

>> 2. Location check

>> 3. Correct / assign coordinates to records without them or errors from previous validations

>> 4. Outliers check

Temporal

Duplicates

Final Report

References

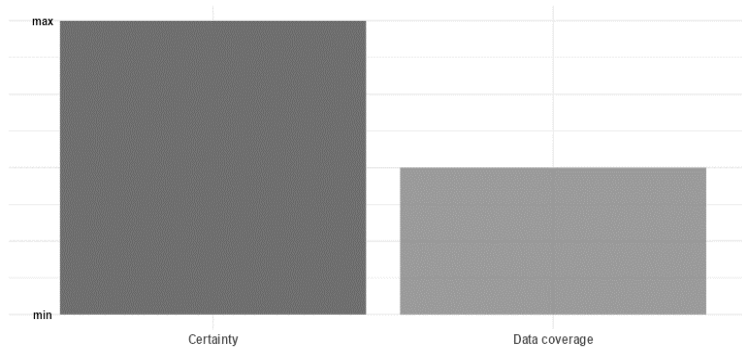
About

Use distributional information

Use environmental information

Choose between:

- ☐ a. Calculate environmental centroids for the species and validate outliers.
- ☒ b. Calculate environmental space for each species, check whether records overlap with it, and delete outliers.
- ☐ c. Overlap environmental information by geographical position and filter occurrences by threshold.
- ☐ d. Do not apply a filter for environmental outliers



Category	Value
Certainty	max
Data coverage	~0.75

Trade-off

Methods

R code

Copy

```
library(sf) # Point in polygon analysis

# Upload dataframe with occurrences (occData) and shapefile of administrative units (countriesSHP)

datapoints <- st_as_sf(x = occData, coords = c('decimalLongitude', 'decimalLatitude'), crs = '+proj=longlat +datum=WGS84 +no_defs')

occData <- st_join(datapoints, countriesSHP)
```

Your selection:

*Applying previous filter: TRUE

*Checking coordinates precision:


*Checking coordinates value:

*Checking coordinates position:

*Recovering coordinates:

*Detecting distributional outliers: TRUE

*Detecting environmental outliers: TRUE




9. Pulsa el botón ‘Download’ disponible en el panel ‘Final report’ para descargar un archivo txt con tu guía final. Incluye los pasos seleccionados para procesar los datos y escribir tu sección de métodos en base a la información de cada módulo.

OCCUR app

☰

SOURCE



👁 Basis of Record

🌿 Taxonomic <

📍 Geographic <

» 1. Previous filters in download process

» 2. Location check

» 3. Correct / assign coordinates to records without them or errors from previous validations

» 4. Outliers check

🕒 Temporal

📄 Duplicates

📄 Final Report

📖 References

📖 About

|----- FINAL REPORT -----|

Based on the steps selected in OCCUR App, the summary of methods chose by the user to filter and clean biodiversity records is:

Basis of Record: The user will filter records to select Observations

*Taxonomical check sums up following the steps:

1. Download option NOT PROVIDED

2. The taxonomical source for standardization / harmonization will be:
Type AUTOMATIC;
Spatial coverage GLOBAL;
Taxonomical coverage GENERAL;
using Matching Type EXACT

3. Selecting only records identified at a proper taxonomic rank

4. Selecting records with or without authorship information in their scientific name

5. Including scientific names classified with taxonomical status: Accepted

*Geographical check sums up the following the steps:

1. Previous filters in download process: Only records without known coordinate issues

2. Location check:

- Check coordinates precision: Use number of decimal digits of coordinates as a measure of their precision
- Validate records based on whether coordinates values are out of a reliable range
- Validate coordinates position

3. Correct / assign coordinates to records without them or errors from previous validations: Do not correct coordinate values

4. Distributional Outliers check: Do not apply

5. Environmental Outliers check: Do not apply

*Temporal information filter as: Data with no temporal range using Date of collection

Finally the identification and deletion of *duplicate records* will be done as the combination of same species Cell Year Recorder

The user agrees to use this final report as a guide to process their data and rewrite this text to avoid conflicts due to plagiarism.

Download