



1 Taller GBIF.ES: Mejora de la calidad de datos de biodiversidad



GBIF

Global Biodiversity
Information Facility

Cristina Ronquillo

Ayudante investigación

cristinaronquillo@mncn.csic.es

mncn 25  museo
nacional de
ciencias
naturales



CSIC

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Objetivo

Conocer los principales aspectos necesarios para evaluar, filtrar, corregir y validar conjuntos de datos de biodiversidad.

Objetivos

- ✓ ¿Qué tengo mirar?
- ✓ ¿Cómo comprobar los datos?
- ✓ ¿Qué tengo que descartar?
- ✓ ¿Qué tengo que corregir y cómo puedo hacerlo?

Objetivos

- ✓ ¿Qué tengo mirar?
- ✓ ¿Cómo comprobar los datos?
- ✓ ¿Qué tengo que descartar?
- ✓ ¿Qué tengo que corregir y cómo puedo hacerlo?

Ausencia de información / Información incompleta / Información incorrecta

Objetivos

✓ ¿Para qué?



 **GBIF** INTEGRATED PUBLISHING TOOLKIT (IPT)
free and open access to biodiversity data

email password login **ENGLISH**

Home

About

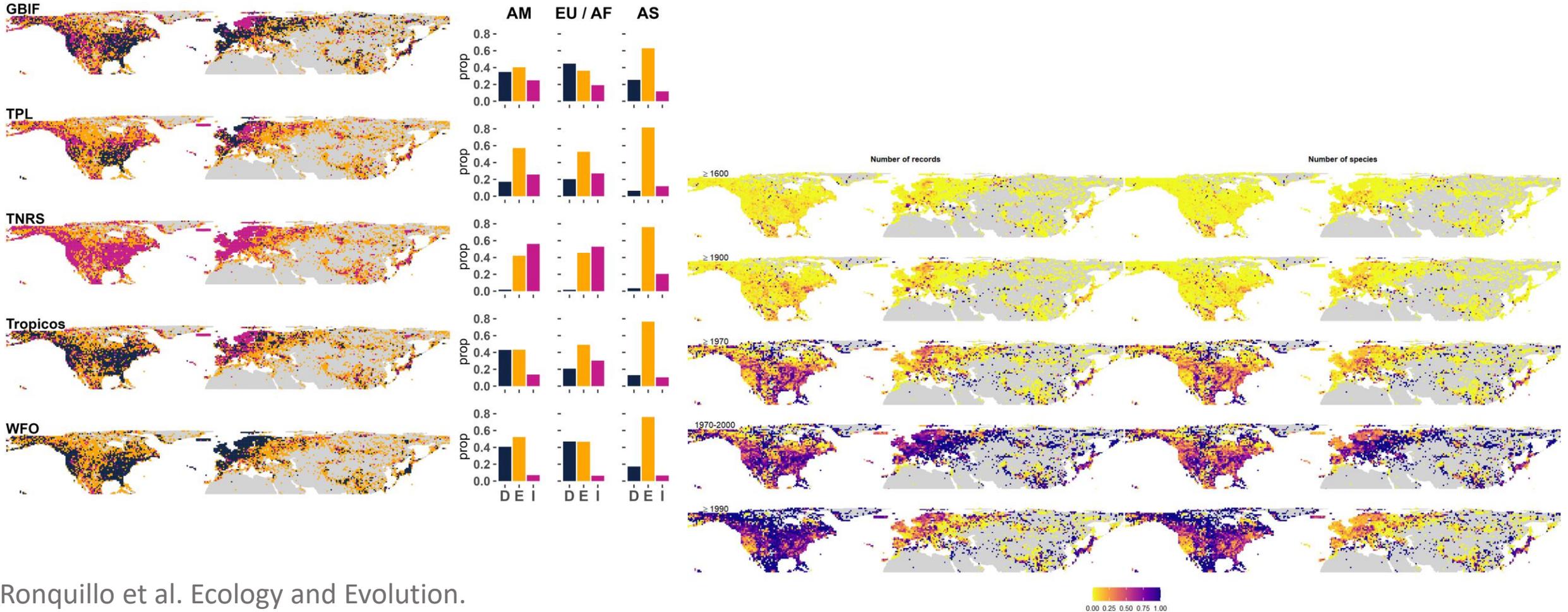
Hosted resources available through this IPT

Filter:

Logo	Name	Organization	Type	Subtype	Records	Last modified	Last publication	Next publication
--	Algae (S)	GBIF-Sweden	Occurrence	Specimen	15,953	2021-04-01	2016-01-05	--
--	Artportalen (Swedish Species Observation System)	ArtDatabanken	Occurrence	Observation	80,765,776	2021-04-09	2021-04-09	2021-04-16 15:00:18
--	Axel W. Erikssons African Bird Collection at Vänersborg Museum	GBIF-Sweden	Occurrence	Specimen	1,000	2021-03-30	2020-12-08	--
--	Beetles (LSM)	GBIF-Sweden	Occurrence	--	13,450	2021-03-18	2021-03-18	--
--	Bird Collection of Helsingborg Museums	GBIF-Sweden	Occurrence	Specimen	2,530	2017-09-18	2017-08-16	--

Objetivos

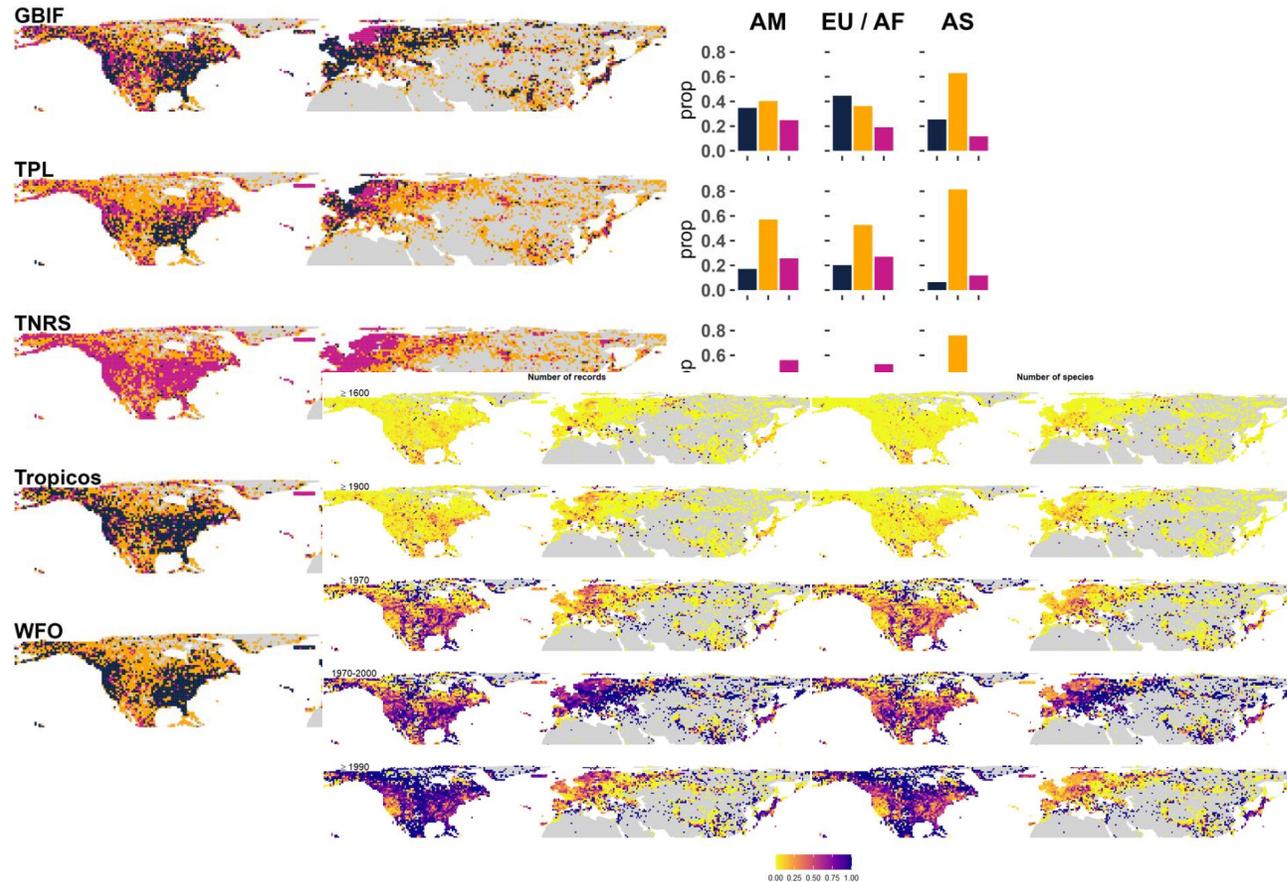
✓ ¿Para qué?



Ronquillo et al. Ecology and Evolution. 2023;13:e10786.

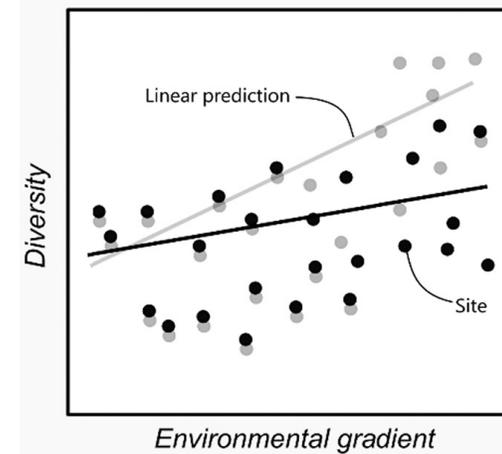
Objetivos

✓ ¿Para qué?

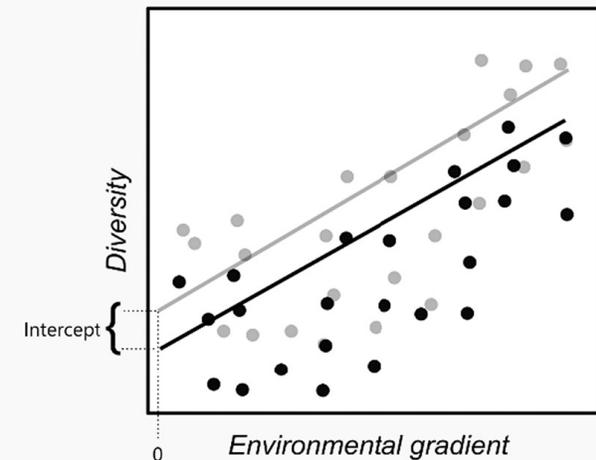


○ All occurrence data
● Only reliable occurrence data

A) Bias (change in the model slope)



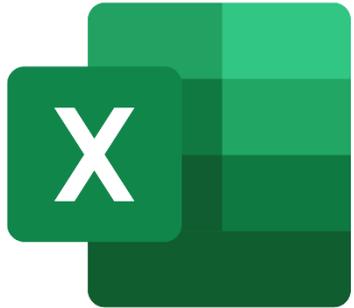
B) Noise (change in the model intercept)



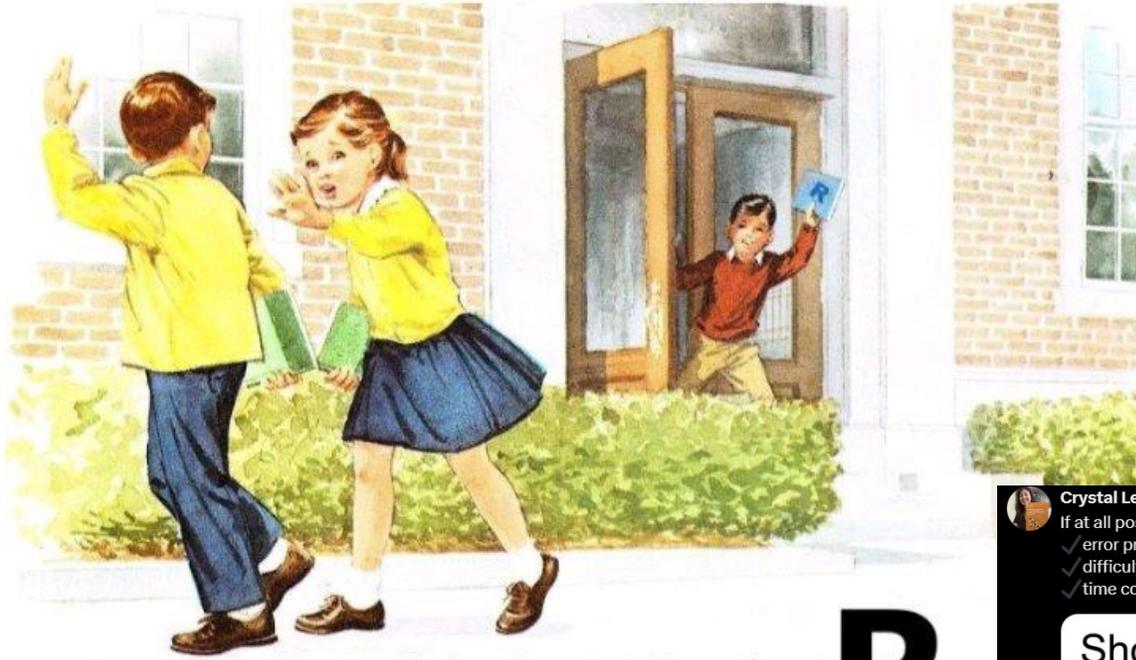
Ronquillo et al. Ecology and Evolution. 2023;13:e10786.

Rodrigues et al. 2022. Ecological Informatics <https://doi.org/10.1016/j.ecoinf.2022.101625>

Herramientas:

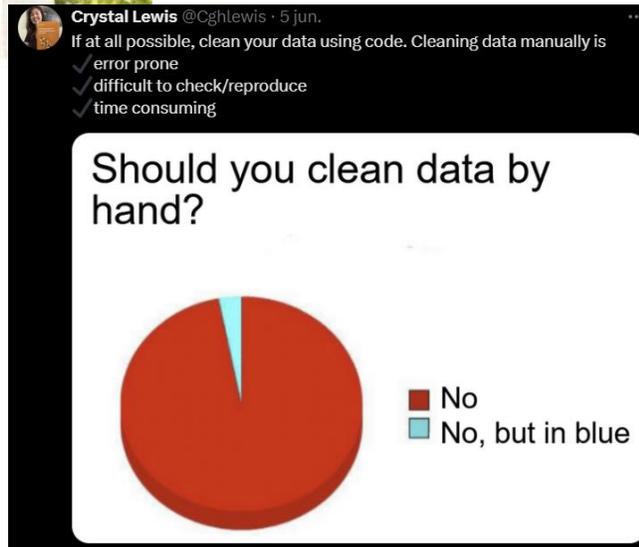


Herramientas:



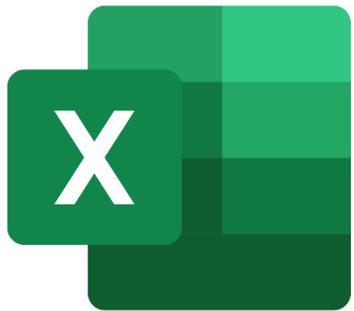
Run, or he's going to tell us about again!

R



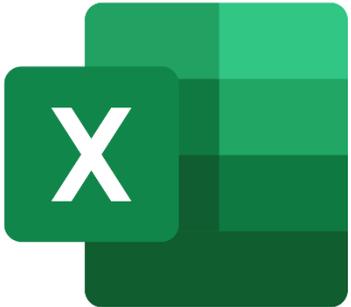
R Studio®

Herramientas:



Studio[®]

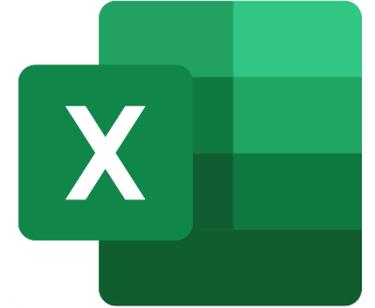
Herramientas:



¿Sé usar Excel?

Herramientas:

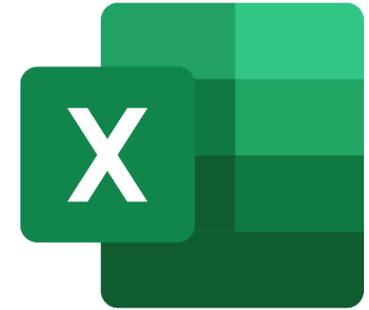
¿Sé usar Excel?



- Abrir un archivo de texto (CSV, TSV)

Herramientas:

¿Sé usar Excel?



- ❑ Abrir un archivo de texto (CSV, TSV)
- ❑ Separadores de campo: tabuladores = '\t'

comas ','

punto y coma ';'

A	B	C	D	E	F	G	H	I	J	K
family	genus	species	infraspecificEpithet	taxonRank	scientificName	verbatimScientificName	verbatimScientificName	Authorship	identifiedBy	
Pinaceae	Pinus	Pinus sylvestris	SPECIES	Pinus sylvestris L.	Pinus sylvestris L.	L.	F.M. Vázquez			
Cupressaceae	Juniperus	Juniperus communis	hemisphaerica	SUBSPECIES	Juniperus communis subsp. hemisphaerica (Jacq. & C.Presl) Nyman	Juniperus communis subsp. hemisphaerica (Jacq. & C.Presl) Nyman				
Cupressaceae	Juniperus	Juniperus communis	alpina	SUBSPECIES	Juniperus communis subsp. alpina (Suter) Celak.	Juniperus communis subsp. alpina (Suter) Celak.				
Pinaceae	Pinus	Pinus sylvestris	SPECIES	Pinus sylvestris L.	Pinus sylvestris L.	L.				
Pinaceae	Pinus	Pinus pinea	SPECIES	Pinus pinea L.	Pinus pinea L.	L.				
Cupressaceae	Juniperus	Juniperus oxycedrus	SPECIES	Juniperus oxycedrus L.	Juniperus oxycedrus L.	L.				
Pinaceae	Pinus	Pinus halepensis	SPECIES	Pinus halepensis Mill.	Pinus halepensis Mill.	Mill.				
Cupressaceae	Cupressus	Cupressus sempervirens	SPECIES	Cupressus sempervirens L.	Cupressus sempervirens L.	L.				
Pinaceae	Pinus	Pinus pinaster	SPECIES	Pinus pinaster Aiton	Pinus pinaster Aiton	Aiton				
Cupressaceae	Juniperus	Juniperus communis	SPECIES	Juniperus communis L.	Juniperus communis L.	L.				
Cupressaceae	Juniperus	Juniperus thurifera	SPECIES	Juniperus thurifera L.	Juniperus thurifera L.	L.				

Herramientas:

¿Sé usar Excel?



Abrir un archivo de texto (CSV, TSV)

Separadores de campo:

A	B	C	D	E	F		
family	genus	species	infraspecificEpithet	taxonRank	scientificName	verbatimScientifi	
Pinaceae	Pinus	Pinus sylvestris	SPECIES	Pinus sylvestris L.	Pinus sylvestris L.	L.	F.M.
Cupressaceae	Juniperus	Juniperus communis	hemisphaerica	SUBSPECIES	Juniperus con		
Cupressaceae	Juniperus	Juniperus communis	alpina	SUBSPECIES	Juniperus communis s		
Pinaceae	Pinus	Pinus sylvestris	SPECIES	Pinus sylvestris L.	Pinus sylvestris L.	L.	
Pinaceae	Pinus	Pinus pinea	SPECIES	Pinus pinea L.	Pinus pinea L.	L.	
Cupressaceae	Juniperus	Juniperus oxycedrus	SPECIES	Juniperus oxycedrus L.	Juniperu		
Pinaceae	Pinus	Pinus halepensis	SPECIES	Pinus halepensis Mill.	Pinus halepensis Mill.		
Cupressaceae	Cupressus	Cupressus sempervirens	SPECIES	Cupressus sempervirens L.			
Pinaceae	Pinus	Pinus pinaster	SPECIES	Pinus pinaster Aiton	Pinus pinaster Aiton	Aitc	
Cupressaceae	Juniperus	Juniperus communis	SPECIES	Juniperus communis L.	Juniperu		
Cupressaceae	Juniperus	Juniperus thurifera	SPECIES	Juniperus thurifera L.	Juniperus t		

Asistente para importar texto - paso 2 de 3

Esta pantalla le permite establecer los separadores contenidos en los datos. Se puede ver cómo cambia el texto en la vista previa.

Separadores

- Tabulación
- punto y coma
- Coma
- Espacio
- Otro:

Considerar separadores consecutivos como uno solo

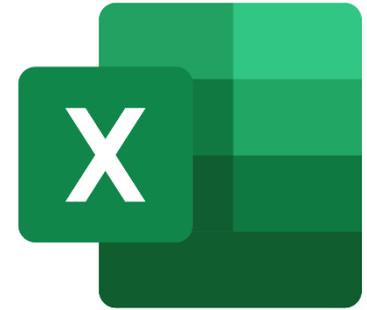
Calificador de texto:

Vista previa de los datos

gbifID	datasetKey	occurrenceID	kingdom	phylum	cl
910454707	837acfc2-f762-11e1-a439-00145eb45e9a	HSS:HSS:40827	Plantae	Tracheophyta	Pi
895210902	834a4794-f762-11e1-a439-00145eb45e9a	8EA68B28-5282-417B-9A69-5ADBFD63BD77	Plantae	Tracheophyta	Pi
857365043	59bf2c83-1e3c-40c8-9437-39ce3d3d462c	URJC:BG URJC:88 - 1	Plantae	Tracheophyta	Pi
728786709	fab4c599-802a-4bfc-8a59-fc7515001bfa	MAGRAMA:IFN3:462510	Plantae	Tracheophyta	Pi
728739861	fab4c599-802a-4bfc-8a59-fc7515001bfa	MAGRAMA:IFN3:420344	Plantae	Tracheophyta	Pi

Herramientas:

¿Sé usar Excel?



- Abrir un archivo de texto (CSV, TSV)
- Separadores de campo
- Símbolo de decimales

decimalLatitude	decimalLongitude	coordinateUncertainty
403,638	-42,837	25.0
403,644	-42,248	25.0
403,645	-42,131	25.0
403,646	-42,013	25.0
403,661	-3.92	
403,661	-3.92	
403,662	-40,364	25.0
403,667	-39,893	25.0
403,675	-4.31	
403,697	-35,182	25.0
403,697	-35,182	25.0
403,704	-4,071,549	5197.0
403,722	-4,331	25.0
403,724	-43,192	25.0
403,725	-43,074	25.0
403,726	-42,957	25.0
403,726	-42,957	25.0
403,729	-42,721	25.0
403,737	-42,014	25.0
403,737	-4,328,028	16.0
403,797	-32,356	25.0
403,806	-42,854	
403,812	-43,312	25.0
403,812	-43,312	25.0
403,815	-43,076	25.0
403,816	-42,958	25.0
403,818	-42,841	25.0
403,822	-3,598,478	300.0

Herramientas:

¿Sé usar Excel?



- ❑ Abrir un archivo
- ❑ Separadores
- ❑ Símbolo de c

Asistente para importar texto - paso 3 de 3

Esta pantalla permite seleccionar cada columna y establecer el formato de los datos.

Formato de los datos en columnas

- General
- Texto
- Fecha: DMA
- No importar columna (saltar)

'General' convierte los valores numéricos en números, los valores de fechas en fechas y todos los demás valores en texto.

Avanzadas...

Configuración avanzada de importación de text...

Valores predeterminados para reconocer datos numéricos

Separador decimal: |

Separador de miles: ,

Nota: Los números se mostrarán usando las opciones de número especificadas en el panel de control Configuración regional.

Restablecer Signo menos detrás de los números negativos

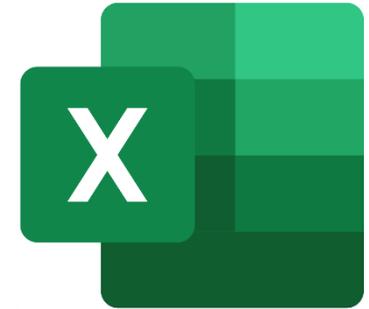
Aceptar Cancelar

Vista previa de los datos

General	General
gbifID	datasetKey
910454707	837acfc2-f762-11e1-a439-00145eb49
895210902	834a4794-f762-11e1-a439-00145eb49
857365043	59bf2c83-1e3c-40c8-9437-39ce3d3d4
728786709	fab4c599-802a-4bfc-8a59-fc7515001
728739861	fab4c599-802a-4bfc-8a59-fc7515001

Herramientas:

¿Sé usar Excel?

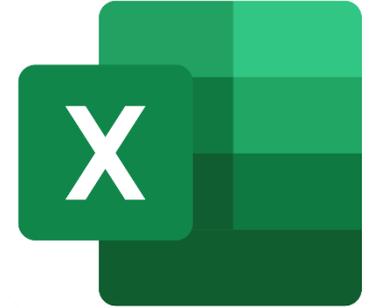


- ❑ Abrir un archivo de texto (CSV, TSV)
- ❑ Separadores de campo
- ❑ Símbolo de decimales
- ❑ Encoding
(Windows, ASCII, UTF-8)

1	verbatimScientificName	verbatimScientific	countryCo	locality	si
2	Pinus halepensis Mill.	Mill.	ES	Colmenar de Oreja	M
3	Pinus halepensis Mill.	Mill.	ES	Villamanrique de Tajo, carretera a Belmonte, finca La Encomienda	M
4	Juniperus oxycedrus L.	L.	ES	Pelahustán	T
5	Pinus halepensis Mill.	Mill.	ES	Morata de Tajuá	N
6	Cupressus sempervirens L.		ES		
7	Juniperus communis L.	L.	ES	Cenicientos	M
8	Juniperus communis L.	L.	ES	Cenicientos	M
9	Juniperus oxycedrus L.	L.	ES	Villa del Prado	M
10	Pinus pinaster Aiton	Aiton	ES	Cenicientos	M
11	Juniperus oxycedrus L.	L.	ES	Peña de Cenicientos	N
12	Juniperus oxycedrus L.	L.	ES	Villa del Prado	N
13	Pinus halepensis Mill.	Mill.	ES	Arganda del Rey	N
14	Juniperus oxycedrus L.	L.	ES	Villa del Prado	M
15	Pinus pinaster Aiton	Aiton	ES		
16	Juniperus communis L.	L.	ES	San Martín de Valdeiglesias	M
17	Juniperus communis L.	L.	ES	Navas del Rey	N
18	Juniperus oxycedrus L.	L.	ES	Villamantilla	N
19	Juniperus thurifera L.	L.	ES	Villaviciosa de Odón	N
20	Pinus pinea L.	L.	ES	Villaviciosa de Odón	N
21	Pinus pinea L.	L.	ES	Villaviciosa de Odón	M
22	Pinus pinea L.	L.	ES	San Martín de Valdeiglesias	M
23	Pinus pinea L.	L.	ES	San Martín de Valdeiglesias	M
24	Juniperus oxycedrus L.	L.	ES	Colmenar del Arroyo	N
25	Pinus pinea L.	L.	ES	San Martín de Valdeiglesias	N

Herramientas:

¿Sé usar Excel?



- Abrir un archivo de
- Separadores de car
- Símbolo de decima
- Encoding
(Windows, ASCII)

Asistente para importar texto - paso 1 de 3

El asistente estima que sus datos son Delimitados.
Si esto es correcto, elija Siguiente, o bien elija el tipo de datos que mejor los describa.

Tipo de los datos originales

Elija el tipo de archivo que describa los datos con mayor precisión:

Delimitados - Caracteres como comas o tabulaciones separan campos.
 De ancho fijo - Los campos están alineados en columnas con espacios entre uno y otro.

Comenzar a importar en la fila: Origen del archivo: Windows (ANSI)

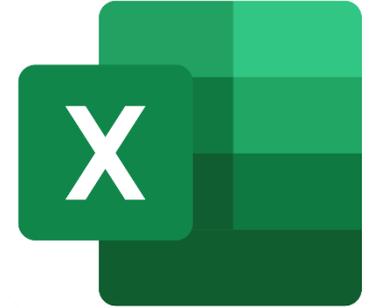
Mis datos tienen encabezados.

Vista previa del archivo C:\Users\Joaquin Hortal\Desktop\NICED_SCENIC\Cursos_seminarios\GBIF_DATOS\practica1\occurrences_CAM_Pinopsida.cs

1	gbifIDdatasetKeyoccurrenceIDkingdomphylumclassorderfamilygenusspeciesinfraspecificEpithettaxon
2	910454707837acfc2-f762-11e1-a439-00145eb45e9aHSS:HSS:40827PlantaeTracheophytaPinopsidaPinalesPinac
3	895210902834a4794-f762-11e1-a439-00145eb45e9a8EA68B28-5282-417B-9A69-5ADBFD63BD77PlantaeTracheophyta
4	85736504359bf2c83-1e3c-40c8-9437-39ce3d3d462cURJC:BG URJC:88 - 1PlantaeTracheophytaPinopsidaPinales
5	728786709fab4c599-802a-4bfc-8a59-fc7515001bfaMAGRAMA:IFN3:462510PlantaeTracheophytaPinopsidaPinales
6	728739861fab4c599-802a-4bfc-8a59-fc7515001bfaMAGRAMA:IFN3:420344PlantaeTracheophytaPinopsidaPinales

Herramientas:

¿Sé usar Excel?



- AVISO:** Si guardáis un archivo sin atender a su formato la siguiente vez que lo abráis o compartáis con otro ordenador pueden haber lágrimas
- SIEMPRE** mantened una copia con la versión original de los datos
- SIEMPRE** usar y mantener identificadores únicos (gbifID)

Trabajaremos a dos niveles:



Aprender a validar los datos generados en el proceso de investigación propio o aquellos recolectados por múltiples investigadores en un proyecto y detectar los errores más habituales.

Establecer el tratamiento básico que debemos dar a los datos que **descargamos de repositorios** públicos para poder utilizarlos en nuestra investigación.

Trabajaremos a dos niveles:



Aprender a validar los datos generados en el **proceso de investigación propio** o aquellos recolectados por múltiples investigadores en un proyecto y detectar los errores más habituales.

Establecer el tratamiento básico que debemos dar a los datos que **descargamos de repositorios** públicos para poder utilizarlos en nuestra investigación.

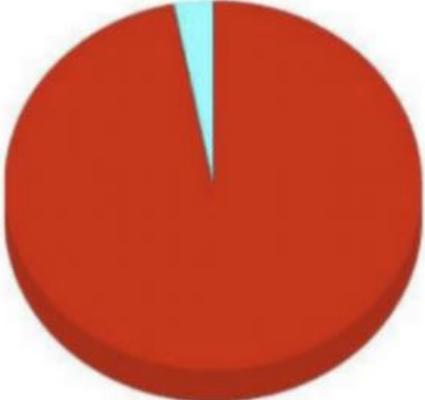
Trabajaremos a dos niveles:

 **Crystal Lewis** @Cghlewis · 5 jun.

If at all possible, clean your data using code. Cleaning data manually is

- ✓ error prone
- ✓ difficult to check/reproduce
- ✓ time consuming

Should you clean data by hand?



Response	Percentage
No	~95%
No, but in blue	~5%

■ No
■ No, but in blue



Establecer el tratamiento básico que debemos dar a los datos que **descargamos de repositorios** públicos para poder utilizarlos en nuestra investigación.

¿Qué es un registro biológico?

¿Qué es un registro biológico?

*“Información de que un determinado taxon (**qué**) aparece en una localización específica (**dónde**) en un momento dado (**cuándo**) y recogida por alguien (**quién**)”*



¿Qué es un registro biológico?

‘Información de que un determinado taxon (‘qué’) aparece en una localización específica (‘dónde’) en un momento dado (‘cuándo’) y recogida por alguien (‘quién’)’



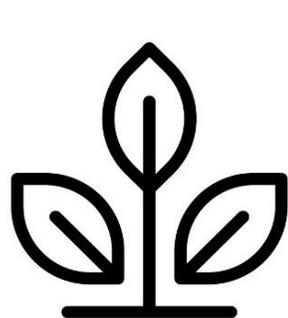
¿Qué es un registro biológico?

‘Información de que un determinado taxon (‘qué’) aparece en una localización específica (‘dónde’) en un momento dado (‘cuándo’) y recogida por alguien (‘quién’)’



¿Qué es un registro biológico?

‘Información de que un determinado taxon (‘qué’) aparece en una localización específica (‘dónde’) en un momento dado (‘cuándo’) y recogida por alguien (‘quién’)’



Bases de datos

BOEN

OBIS
OCEAN BIOGEOGRAPHIC
INFORMATION SYSTEM

TRY
Plant Trait Database

 **Atlas of Living
Australia**
ala.org.au


GLOBAL
Building a Global Consortium of Bryophytes and Lichens

 **GBIF**

Bases de datos

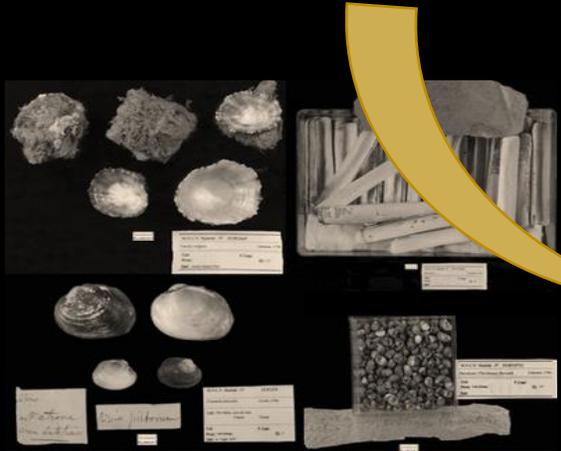


Ciencia Ciudadana

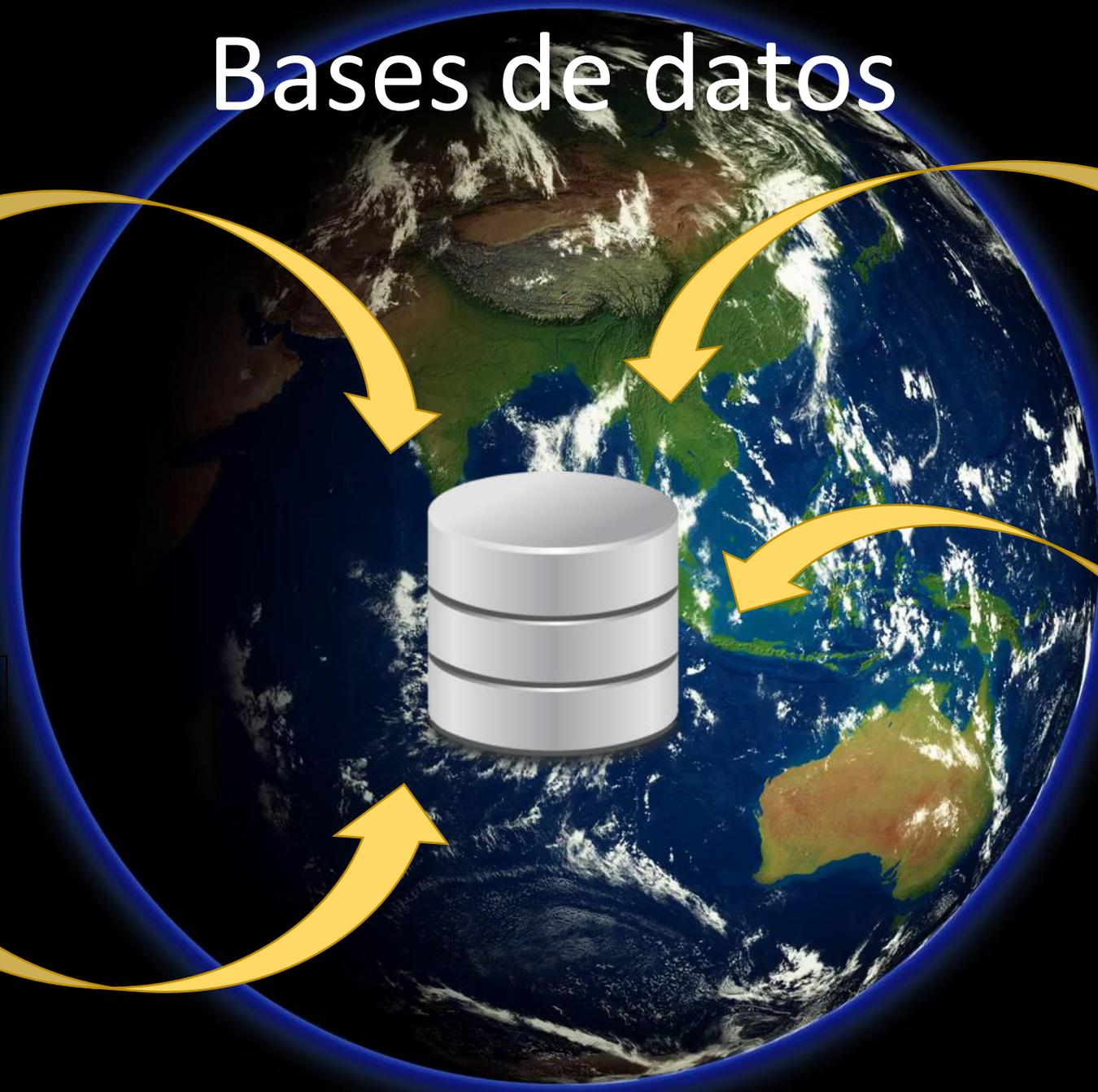
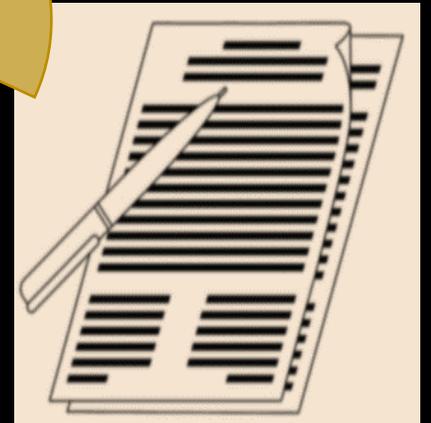


Investigación

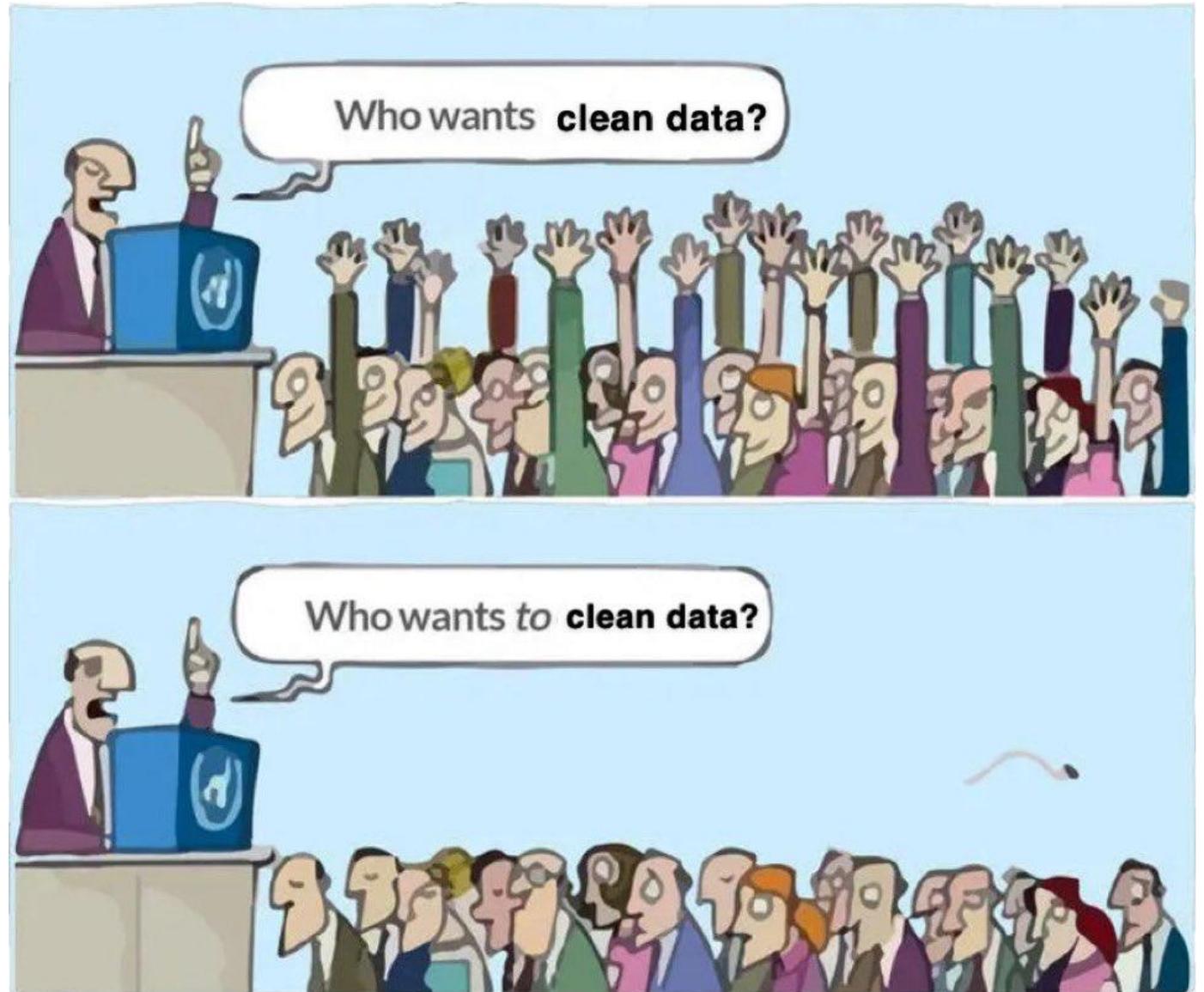
Colecciones



Literatura



¿Por dónde empezamos?



Trabajando con registros de presencia de especies

Planteamiento de trabajo



Necesito hacer un modelo de distribución en la península Ibérica para 10 especies de musgos

Necesito analizar la distribución latitudinal de todos los musgos de la región templada del hemisferio norte



Necesito analizar los 'shortfalls' en Brazil usando todas las especies de termitas del mundo



Necesito hacer modelos predictivos para saber el nicho de las especies de los 'drylands' en esta checklist



Necesito analizar la distribución de 2 especies de *Quercus* en la región mediterránea



Trabajando con registros de presencia de especies

Planteamiento de trabajo



Necesito hacer un modelo de distribución en la península Ibérica para 10 especies de musgos

Necesito analizar la distribución latitudinal de todos los musgos de la región tropical del hemisferio...

Necesito analizar los 'shortfalls' en Brazil usando todos los registros de termitas del...



'Que los datos sean buenos'



Necesito hacer modelos predictivos para saber el nicho de las especies de los 'drylands' en esta checklist

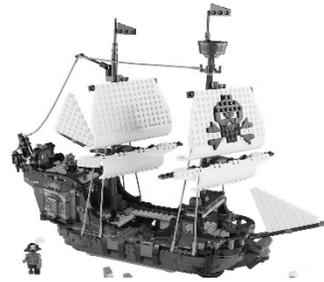
Necesito analizar la distribución de 2 especies de *Quercus* en la región mediterránea



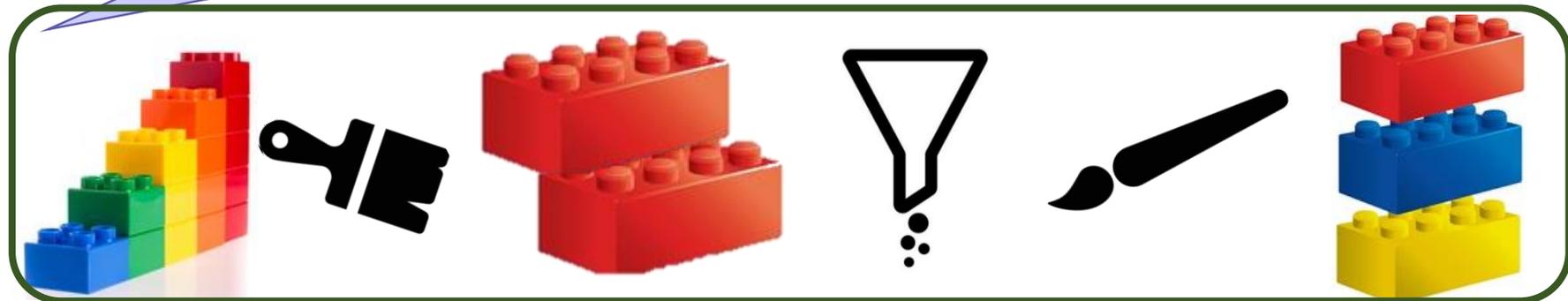
Trabajando con registros de presencia de especies

Planteamiento de trabajo

Objetivo



¿Usando alguna taxonomía concreta?
¿En qué periodo de tiempo?
¿Sólo observaciones o **todo**?
¿Con año de colecta y mes o da igual?
¿Sólo en la región nativa o la introducida también?...



Trabajando con registros de presencia de especies

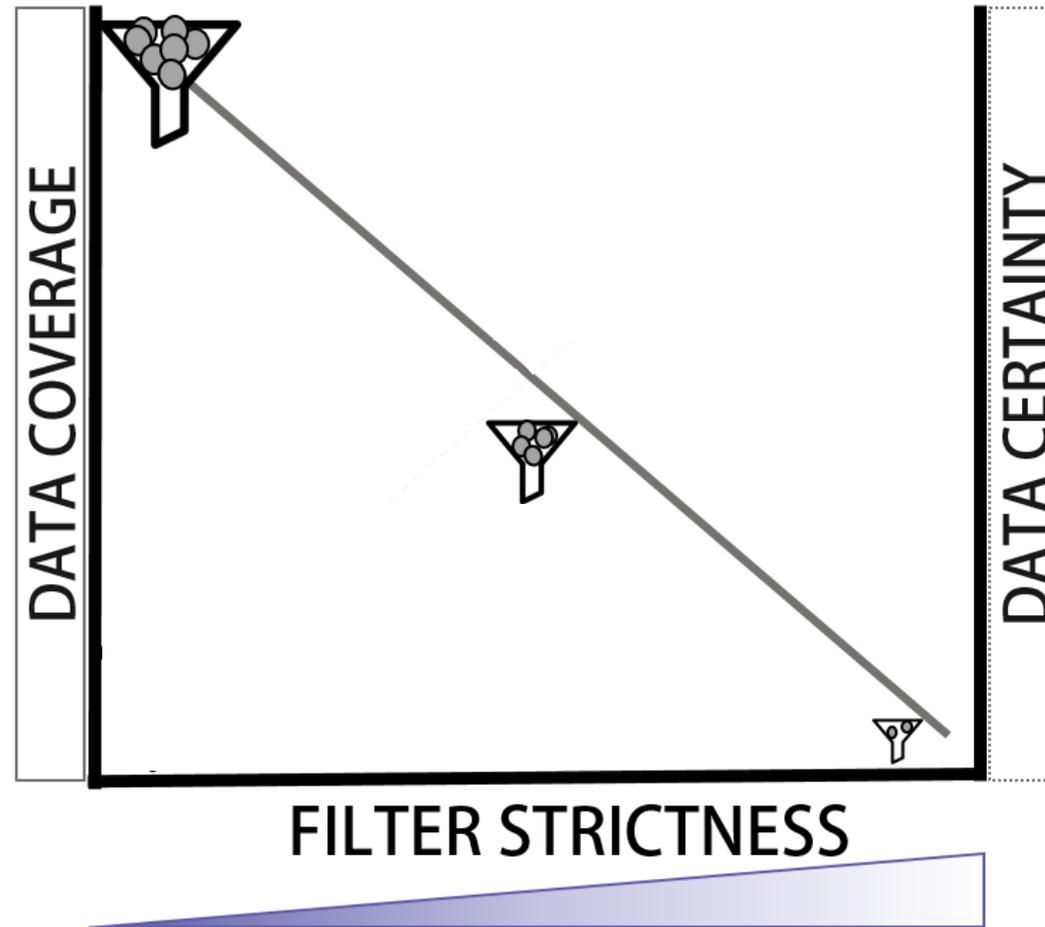
Planteamiento de trabajo

- ¿Usando alguna taxonomía concreta?
- ¿En qué periodo de tiempo?
- ¿Sólo observaciones o **todo**?
- ¿Con año de colecta y mes o da igual?
- ¿Sólo en la región nativa o la introducida también?...



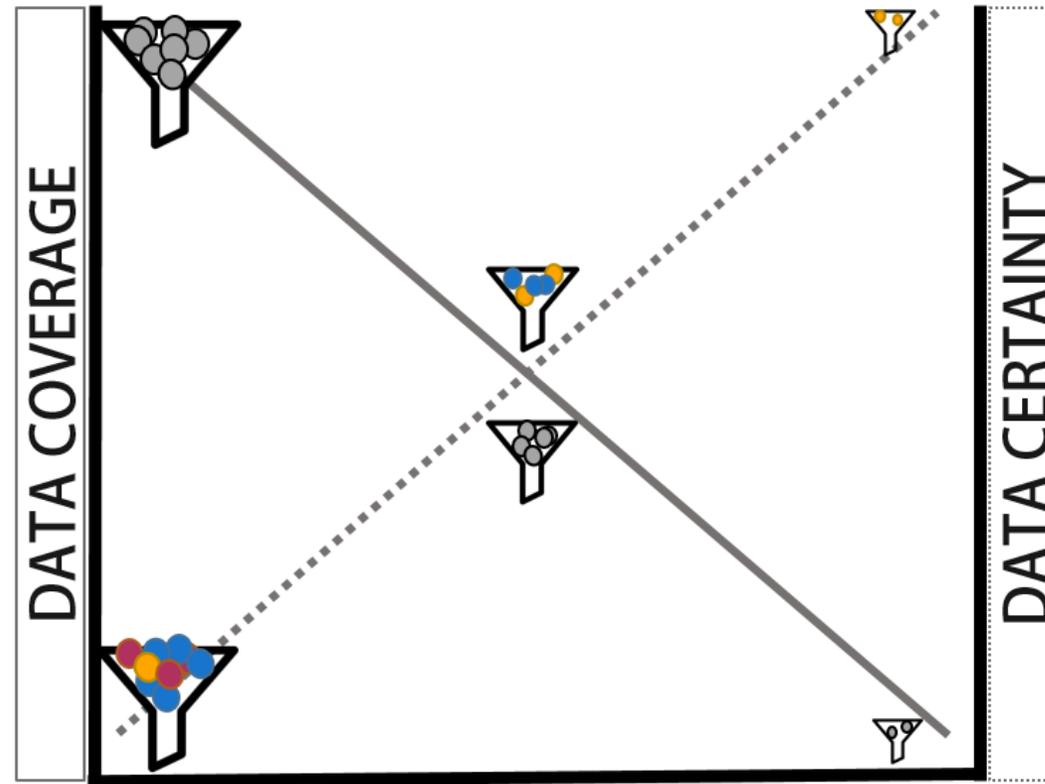
Trabajando con registros de presencia de especies

Número de registros



Trabajando con registros de presencia de especies

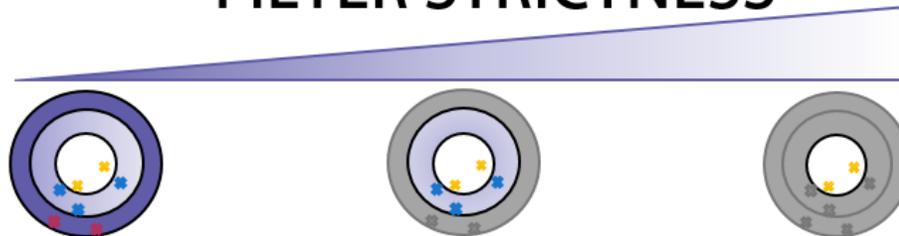
Número de registros



Precisión y exactitud



FILTER STRICTNESS



Trabajando con registros de presencia de especies



Trabajando con registros de presencia de especies



ELIGE TU PROPIA AVENTURA

Tú eres el protagonista de esta historia, personaliza tu experiencia profesional

ELIGE TU PROPIA AVENTURA

PROFESIONALES - ORIENTACIÓN



ETPOEP GRANADA

Trabajando con registros de presencia de especies

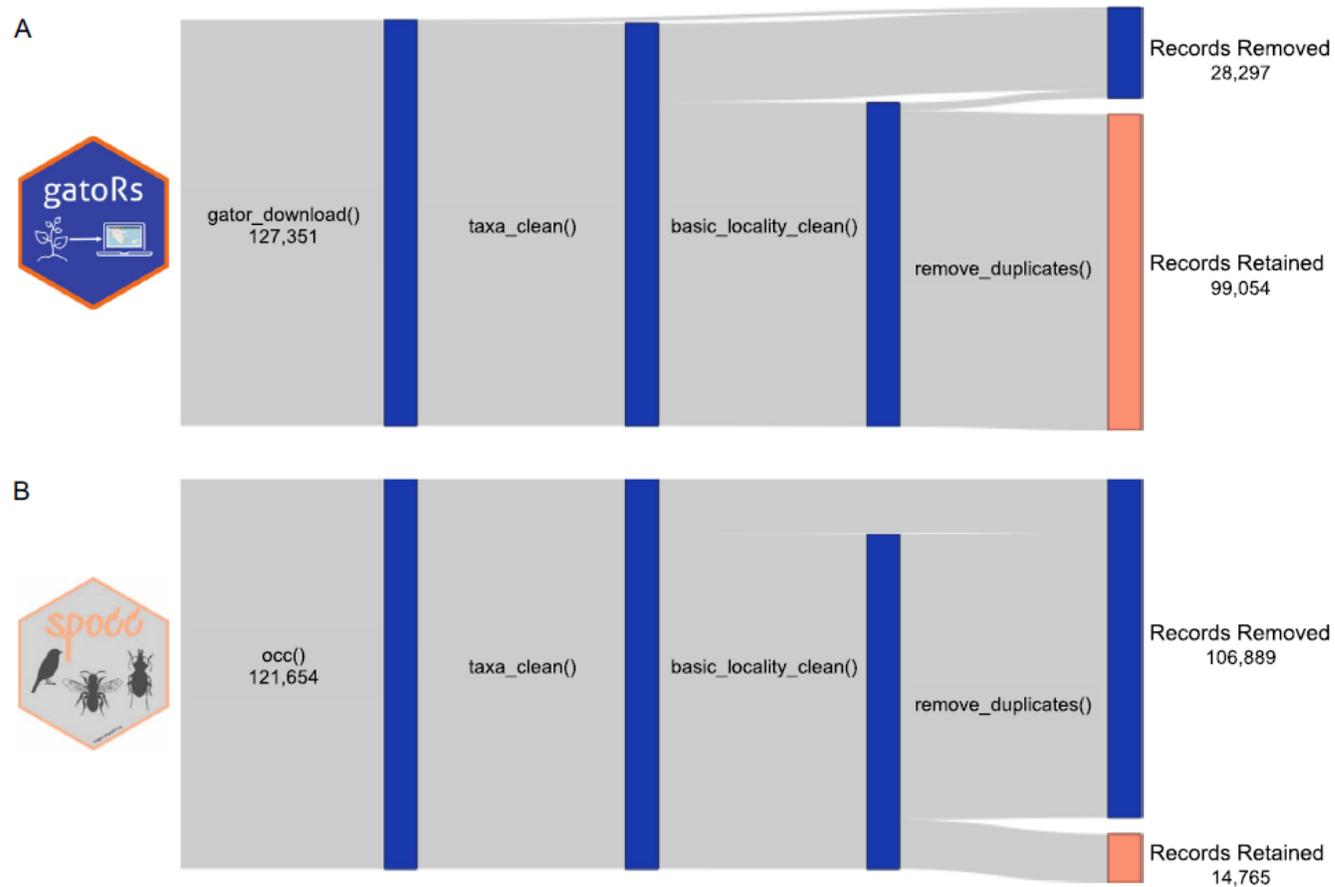


FIGURE 3 Sankey diagrams showing the sum of records returned for all 25 species after each cleaning step when using (A) `gators_download()` from `gatoRs` and (B) `occ()` from `spocc` with the limit set to 100,000. This Sankey diagram was generated using the `networkD3` R package (Allaire et al., 2017) and was inspired by Panter et al. (2020) (see their Figure 3). The number of records after each processing step can be found in Appendix S6. The `spocc` logo was sourced from <https://github.com/ropensci/spocc/blob/master/man/figures/logo.png>.



OCCUR Shiny application: A user-friendly guide for curating species occurrence records



Contents lists available at [ScienceDirect](http://www.elsevier.com/locate/jecol)

Journal homepage: <http://www.elsevier.com/locate/jecol>

Original Research Article

BDcleaner: A workflow for cleaning taxonomic and geographic errors in occurrence data archived in biodiversity databases

Jing Jin^a, Jun Yang^{a,b,*}

^a Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing 100084, China

^b Tsinghua Center for Global Change Studies, Beijing, 100084, China

SpeciesGeoCoder: Fast Categorization of Species Occurrences for Analyses of Biodiversity, Biogeography, Ecology, and Evolution

Mats Tegel^a, Alexander Zizka^a, Maria Fernanda Caló^a, Raad Scharn^a, Daniele Silvestro^a, Alexandre Antonelli^a Author Notes

Systematic Biology, Volume 66, Issue 2, March 2017, Pages 145–151, <https://doi.org/10.1093/sysbio/ytw064>

Published: 02 August 2016 Article history

PDF Split View Cite Permissions Share

Ecography 39, 394–401, 2016
doi:10.1111/ecog.02118
© 2016 The Authors. Ecography © 2016 Nordic Society Oikos
Subject Editor: Brady Sandt, Editor-in-Chief: Miguel Ángel Asís, Accepted: 18 January 2016

Biogeogr: an R package for assessing and improving data quality of occurrence record datasets

Mark P. Robertson, Vernon Visser

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

APPLICATION

bdc: A toolkit for standardizing, integrating and cleaning biodiversity data

Bruno R. Ribeiro^a, Santiago José Elias Velazco^a, Karlo Guidoni-Martins^a, Geiziane Tassarolo^a, Lucas Jardim^a, Steven P. Bachman^a, Rafael Loyola^a

First published: 13 April 2022 | <https://doi.org/10.1111/2041-210X.13868>

Handling Editor: Samantha Price

COORDINATECLEANER: Standardized cleaning of occurrence records from biological collection databases

Alexander Zizka^{1,2,3} | Daniele Silvestro^{1,2,4} | Tobias Andermann^{1,2}



Table 3 cont.

Process	Method	Pros	Cons	DC	C
1. Check and filter based on identifiable coordinates	• Filter and download records with identifiable coordinates (geographic coordinates)	• Exclude filter records with the highest precision in geographic coordinates	• May affect records with low precision in geographic coordinates	-0.25	-0.25
2. Check and filter based on identifiable coordinates	• Use a list of known geographic coordinates (e.g., grid coordinates, lat/lon, UTM, etc.) to filter records	• Exclude records with coordinates that do not match the known coordinates	• May affect records with coordinates that are not in the known list	-0.25	-
3. Check and filter based on identifiable coordinates	• Use a list of known geographic coordinates (e.g., grid coordinates, lat/lon, UTM, etc.) to filter records	• Exclude records with coordinates that do not match the known coordinates	• May affect records with coordinates that are not in the known list	-0.25	-0.25
4. Check and filter based on identifiable coordinates	• Use a list of known geographic coordinates (e.g., grid coordinates, lat/lon, UTM, etc.) to filter records	• Exclude records with coordinates that do not match the known coordinates	• May affect records with coordinates that are not in the known list	-0.25	-

Table 3. Processes of records' filter based on geographical information, pros / cons and estimated values of data coverage (DC) and certainty of data (C) associated to them from 0 (min) to 1 (max).

Process	Pros	Cons	DC	C
Download records without known coordinates	• Exclude cleaning process	• Exclude georeferenced records by locally information that could be repaired	0.25	1
Download records with coordinates filtered by spatial extent (e.g., administrative units)	• Exclude records with artificial assigned coordinates [22]	• Exclude records from suitable/native regions not considered by biogeography	0.8	0.75
Download records with coordinates	• Exclude records with coordinates	• Exclude georeferenced records by locally information that could be repaired	0.75	0.5
Do not apply previous filter	• Exclude all the available information	• Needs an exhaustive process of filtering, cleaning and repairing the data	1	0.25

Table 2. Processes of records' filter based on taxonomical information, pros / cons and estimated values of data coverage (DC) and certainty of data (C) associated to them from 0 (min) to 1 (max).

Process	Method	Pros	Cons	DC	C
Download of records from higher taxonomic level	• Download all the information available to create a higher taxonomic level	• Avoid records without proper taxonomic level	• May affect records with low precision in taxonomic level	-0.1	-0.1
Create a list of species accepted names and synonyms from previous taxonomical knowledge and expert databases	• Use a list of known taxonomic names (e.g., GBIF Backbone Taxonomy, etc.) to filter records	• Exclude records with taxonomic names that do not match the known list	• May affect records with taxonomic names that are not in the known list	-0.1	-0.1
Checklist Type	• Automatic: e.g., Species, Name, Backbone Taxonomy (GBIF), World Bank, GBIF Backbone Taxonomy, etc. • Manual: e.g., GBIF Backbone Taxonomy, etc.	• Increase processing time	• May affect records with low precision in taxonomic level	-	-
Spatial coverage	• Global: e.g., GBIF Backbone Taxonomy, etc. • Regional: e.g., GBIF Backbone Taxonomy, etc.	• Exclude records with coordinates that do not match the known list	• May affect records with coordinates that are not in the known list	-	-
Taxon. coverage	• General: e.g., GBIF Backbone Taxonomy, etc. • Specific: e.g., GBIF Backbone Taxonomy, etc.	• Exclude records with taxonomic names that do not match the known list	• May affect records with taxonomic names that are not in the known list	-	-
Type of matching	• Exact: e.g., GBIF Backbone Taxonomy, etc. • Fuzzy: e.g., GBIF Backbone Taxonomy, etc.	• Exclude records with taxonomic names that do not match the known list	• May affect records with taxonomic names that are not in the known list	-0.1	-0.1

Table 2 cont.

Process	Method	Pros	Cons	DC	C
Is the record identified at a proper taxonomic rank?	• Use a list of known taxonomic names (e.g., GBIF Backbone Taxonomy, etc.) to filter records	• Exclude records with taxonomic names that do not match the known list	• May affect records with taxonomic names that are not in the known list	-0.1	-0.1
Does the scientific name have authorship information?	• Use a list of known taxonomic names (e.g., GBIF Backbone Taxonomy, etc.) to filter records	• Exclude records with taxonomic names that do not match the known list	• May affect records with taxonomic names that are not in the known list	-0.1	-0.1
Check taxonomical status	• Accepted: e.g., GBIF Backbone Taxonomy, etc. • Synonym: e.g., GBIF Backbone Taxonomy, etc. • Unresolved / No match: e.g., GBIF Backbone Taxonomy, etc.	• Exclude records with taxonomic names that do not match the known list	• May affect records with taxonomic names that are not in the known list	-0.1	-0.1

SUPPLEMENTARY MATERIAL

Table 1. Processes of records' filter based on Basis of records information, pros / cons and estimated values of data coverage (DC) and certainty of data (C) associated to them from 0 (min) to 1 (max).

Process	Pros	Cons	DC	C
Do not apply filter	• Exclude records with low precision in geographic coordinates	• Exclude records with low precision in taxonomic level	1	0.25
Select one type	• Preserve Species: e.g., GBIF Backbone Taxonomy, etc. • Observations: e.g., GBIF Backbone Taxonomy, etc.	• Exclude records with low precision in geographic coordinates	0.25	0.5
Select	• Preserve Species: e.g., GBIF Backbone Taxonomy, etc. • Observations: e.g., GBIF Backbone Taxonomy, etc.	• Exclude records with low precision in geographic coordinates	0.5	0.5





SCAN ME

OCCUR app SOURCE

OCCUR

OCCUR

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

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.

Basis Of Record

 →

Taxonomy

 →

Geography

 →

Time

 →

Duplicates

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

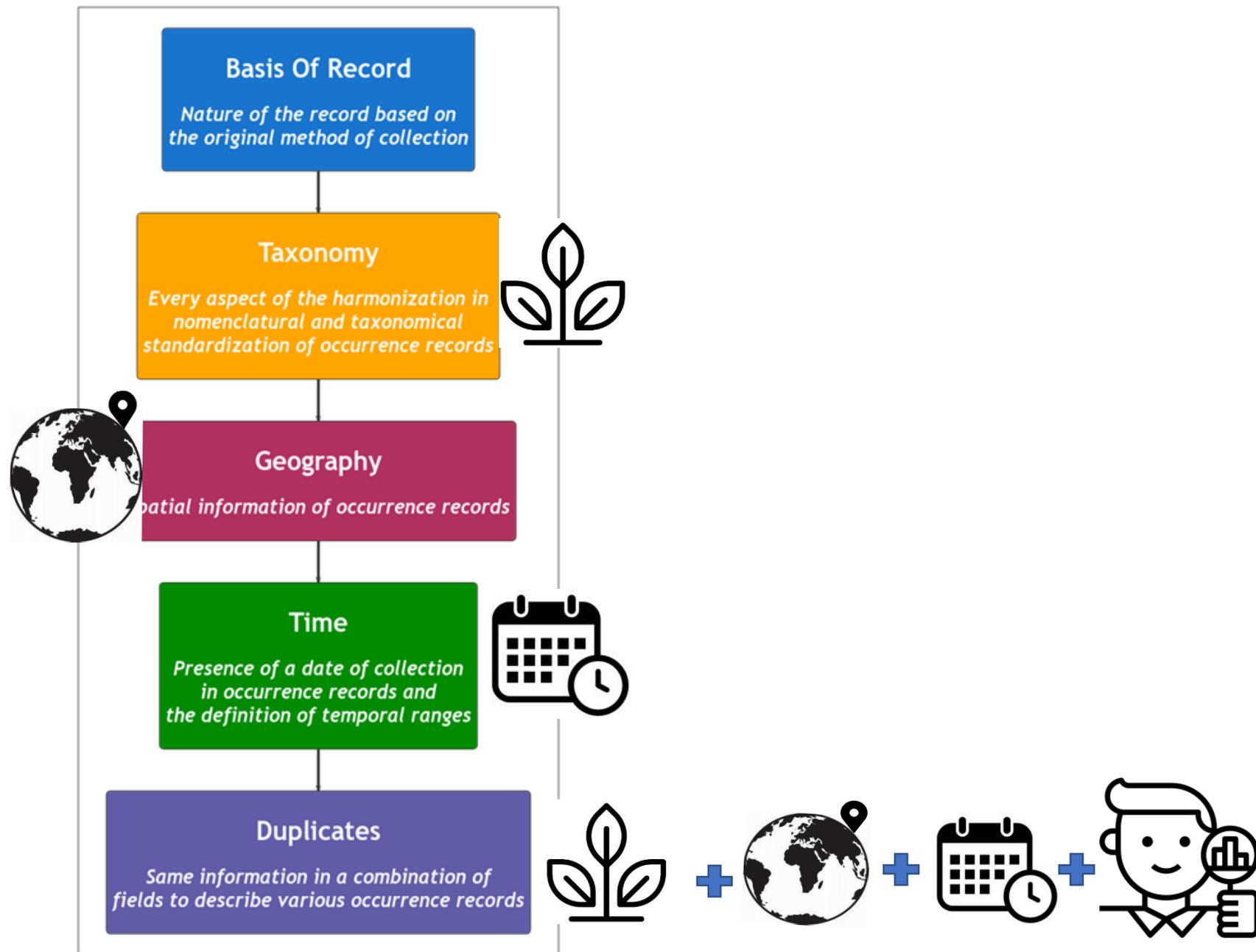
FILTER STRICTNESS

DATA CERTAINTY

https://ecoinformatic.shinyapps.io/OCCUR/_w_c622ec2c/#shiny-tab-dashboard

I Taller GBIF.ES: Mejora de la calidad de datos de biodiversidad

5 módulos de trabajo



5 módulos de trabajo

OCCUR app

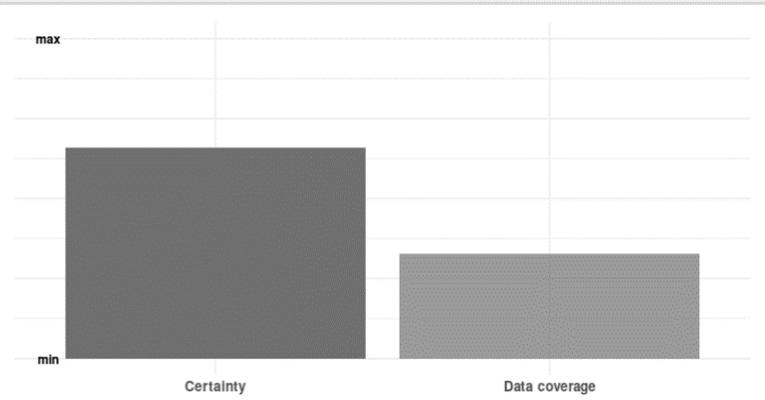


- Basis of Record
- Taxonomic
- Geographic
- 1. Previous filters in download process
- 2. Location check
- 3. Correct / assign coordinates to records without them or errors from previous validations
- 4. Outliers check
- Temporal
- Duplicates
- Final Report
- References
- About

1. Check coordinates' precision 2. Check coordinates' values 3. Check position of coordinates

Validate each option from low to high strictness:

- a. Are coordinates placed in correct habitat (sea / land)?
- b. Are coordinates placed in the country assigned?
- c. Check position of records that are not in the country assigned.
- d. Check records placed in prime meridian or equator countries
- e. Delete or label as potential errors those records whose coordinates are centroids
- f. Skip this step



Trade-off	Methods	R Code
1. Check coordinates precision	2. Check coordin. values	

Pros

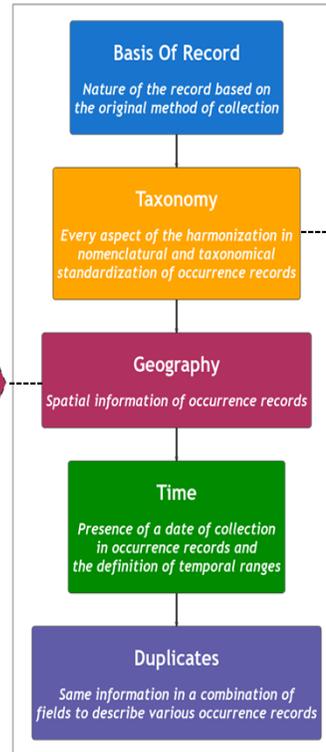
Identifies location errors due to signs of coordi

Excludes unreliable records.

Your selection:

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

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



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

Trade-off | **Methods**

Checklist Type
 e.g. Taxonomic Name Resolution Service (TNRS); WorldFlora, GBIF backbone
 bdc::bdc_query_names_taxadb [18]

Spatial coverage
 e.g. Flora Iberica

Taxonomical coverage
 e.g. GBIF backbone name parser rgbif; Global Name Resolver web service

Matching Type
 Match taxon names with the exact same spelling [12] (e.g. Taxonstand based on The Plant List)
 bdc::bdc_query_names_taxadb suggest_names = FALSE [18]

Methods

Copy to clipboard



OCCUR

Copy to clipboard

OCCUR

Copy to clipboard

Trade-off **Methods**

Checklist Type

e.g. Taxonomic Name Resolution Service (TNRS); WorldFlora, GBIF backbone
 bdc::bdc_query_names_taxadb [18]

Spatial coverage

e.g. Flora Iberica

Taxonomical coverage

e.g. GBIF backbone name parser rgbf; Global Name Resolver web service

Matching Type

Match taxon names with the exact same spelling [12] (e.g. Taxonstand based on The Plant List)

bdc::bdc_query_names_taxadb suggest_names = FALSE [18]

Methods

R Code

Trade-off **Methods** **R Code** **SOURCE**

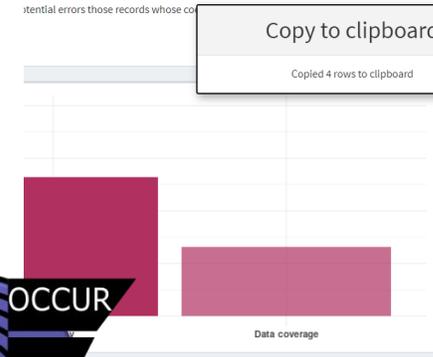
1. Check coordinates precision 2. Check coordinates values 3. Check position of coordinates

Copy

```
library(sf) # Point in polygon analysis
# Create a 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")
countries <- st_join(datapoints, countriesSHP)
```

Copy to clipboard
Copied 4 rows to clipboard

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



OCCUR app

1- 14. See ref
 [13] Vandepitte, L., Bosch, S., Tyberghein, L., Waumans, F., Vanhoorne, B., Hernandez, F., De Clerck, M., & Verbeeck, R. H. G. (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 occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. See ref

[15] Serra-Diaz, J.M., Enquist, B.J., Malmgren, B. et al. (2017). Big data of tree species distributions: how big and how good? Forest Ecosystems, 8, 30-40. See ref

[16] Meiri, S. (2018). The smartphone fallacy - when spatial data are reported at spatial scales finer than the organism. Methods in Ecology and Evolution, 9, 1-14. See ref

[17] Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., (...) Antonelli, A. (2019). Coordinate error in species distribution models: how big and how good? Forest Ecosystems, 8, 30-40. See ref

[18] Ribeiro, B.B., Velazco, S.J., Guidoni-Martins, K., Tassarolo, G., Jardim, L., Bachman, S.P. & Loyola, R. (2022). bdc: a package for cleaning taxonomic and geographic errors in occurrence data. PLOS ONE, 17(12): e0241417. See ref

[19] Robertson, M.P., Visser, V. & Hul, C. (2016). Bioclean: an R package for assessing and improving data quality of occurrence data. PLOS ONE, 11(12): e0163417. See ref

[20] Tassarolo, G., Ladle, R., Lobo, J.M., Rangel, T. & Hortal, J. (2021). Using maps of biogeographical ignorance to assess data quality. PLOS ONE, 16(12): e0241417. 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 for cleaning taxonomic and geographic errors in occurrence data. PLOS ONE, 16(12): e0241417. See ref

[22] Park, D. S., Xie, Y., Thammavong, H. T., Tulaiha, R., & Feng, X. (2022). Artificial Hotspot Occurrence Inventory (AHOI): A method to estimate the accuracy and biogeographical status of occurrence data. PLOS ONE, 17(12): e0241417. See ref

[23] Arle, E., Zizka, A., Keil P. et al. (2021). bRacatus: A method to estimate the accuracy and biogeographical status of occurrence data. PLOS ONE, 16(12): e0241417. See ref

[24] Flannery-Sutherland, J. T., Raja, N. B., Kocals, A. T., & Kiesslring, W. (2022). fossilbrush: An R package for automating the cleaning of fossil occurrence data. PLOS ONE, 17(12): e0241417. See ref

References

Blurred screenshot of a document or code editor.

Trade-off | **Methods**

Checklist Type

e.g. Taxonomic Name Resolution Service (TNRS); WorldFlora, GBIF backbone
 bdc::bdc_query_names_taxadb [18]

Spatial coverage

e.g. Flora Iberica

Taxonomical coverage

e.g. GBIF backbone name parser rgbf; Global Name Resolver web service

Matching Type

Match taxon names with the exact same spelling [12] (e.g. Taxonstand based on The Plant List)

bdc::bdc_query_names_taxadb suggest_names = FALSE [18]

Methods

R Code

Trade-off | **Methods** | R Code

1. Check coordinates precision | 2. Check coordinates values | 3. Check position of coordinates

Copy

library(sf) # Point in polygon analysis

occData <- read.csv('occData.csv') # dataframe with occurrences (occData) and shapefile of administrative units (countriesSHP)

```

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

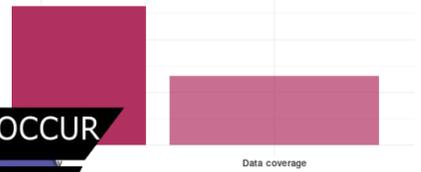
```

Your selection:

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

Copy to clipboard

Copied 4 rows to clipboard



OCCUR app

1-14. See ref

[13] Vandepitte, L., Bosch, S., Tyberghein, L., Waumans, F., Vanhoorne, B., Hernandez, F., De Clerck, M., & Veron, J. E. N. (2014). *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

[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., Malmgren, B. et al. (2017). Big data of tree species distributions: how big and how good? *Forest Ecosystems & Restoration*, 1-14. See ref

[16] Meiri, S. (2018). The smartphone fallacy - when spatial data are reported at spatial scales finer than the organism. *Methods in Ecology and Evolution*, 9, 1-14. See ref

[17] Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., (...) Antonelli, A. (2019). Coordinate cleaning: An R package for cleaning and standardizing geographic coordinates. *Methods in Ecology and Evolution*, 10, 1-14. See ref

[18] Ribeiro, B.B., Velasco, S.J., Guidoni-Martins, K., Tassarolo, G., Jardim, L., Bachman, S.P. & Loyola, R. (2022). bdc: a workflow for cleaning taxonomic and geographic errors in occurrence data. *Global Ecology and Conservation*, 21, e00852. ISSN 2351-9894. See ref

[19] Robertson, M.P., Visser, V. & Hul, C. (2016). Bioclean: an R package for assessing and improving data quality of occurrence data. *Methods in Ecology and Evolution*, 7, 913-922. See ref

[20] Tassarolo, G., Ladle, R., Lobo, J.M., Rangel, T. & Hortal, J. (2021). Using maps of biogeographical ignorance to improve data quality. *Methods in Ecology and Evolution*, 12, 1-14. 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 for cleaning and standardizing plant occurrence data. *Methods in Ecology and Evolution*, 12, 1-14. See ref

[22] Park, D. S., Xie, Y., Thammavong, H. T., Tulaiha, R., & Feng, X. (2022). Artificial Hotspot Occurrence Inventory (AHOI): A method to estimate the accuracy and biogeographical status of occurrence data. *Methods in Ecology and Evolution*, 13, 1-14. See ref

[23] Arle, E., Zizka, A., Keil, P. et al. (2021). bRacatus: A method to estimate the accuracy and biogeographical status of occurrence data. *Methods in Ecology and Evolution*, 12, 1-14. See ref

[24] Flannery-Sutherland, J. T., Raja, N. B., Kocalis, A. T., & Kiessling, W. (2022). fossilbrush: An R package for automating the cleaning of fossil occurrence data. *Methods in Ecology and Evolution*, 13, 1-14. See ref

References

FINAL REPORT

Download

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* filter NOT PROVIDED
- *Taxonomical check sums up following the steps:
 - Download option NOT PROVIDED
 - The taxonomical source for standardization / harmonization will be:
 - Type AUTOMATIC;
 - Spatial coverage REGIONAL;
 - Taxonomical coverage GENERAL;
 - using Matching Type EXACT
 - Selecting records identified at ANY taxonomic rank
 - Selecting records with or without authorship information in their scientific name
 - Including scientific names classified with taxonomical status: NOT PROVIDED
- *Geographical check sums up the following the steps:
 - Previous filters in download process: NOT CONSIDERED
 - Location check:
 - Geographical check sums up the following the steps:
 - Previous filters in download process: NOT CONSIDERED
 - Location check:

Download

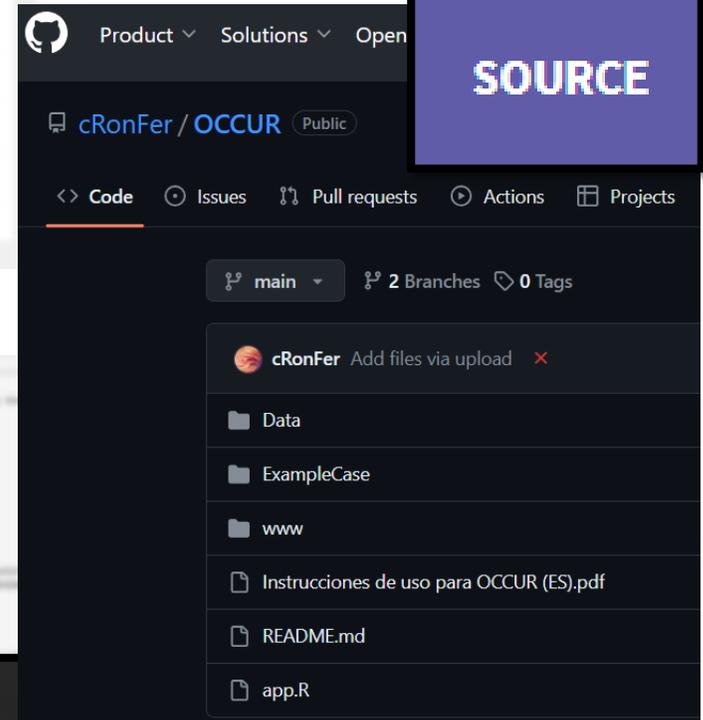
Final Report

✓ Reproducible

✓ Comparable

✓ 'User-friendly'

✓ **Docencia**



Caso de estudio:

Caso de estudio:

- 1. Elige un conjunto de datos de GBIF o usa uno propio*
- 2. Elige alguna validaciones durante estos días y aplícalas a tus datos*
- 3. Anota el número de registros iniciales y el final*
- 4. Crea un mapa donde compares la distribución de registros inicial y tras validarlos*

Caso de estudio:

- Datos de registros de presencia de especies de la clase pinopsida desde 1980
- Descartadas subespecies y variedades o no identificadas a nivel de especie
- Sólo observaciones humanas
- Área Noroeste de la Comunidad de Madrid



Número de registros inicial = 9417



Número de registros final (azul) = 4066