Data Cleaning Overview

University of Florida

FLORIDA MUSEUM

Created by Pam Soltis

UF UNIVERSITY of FLORIDA

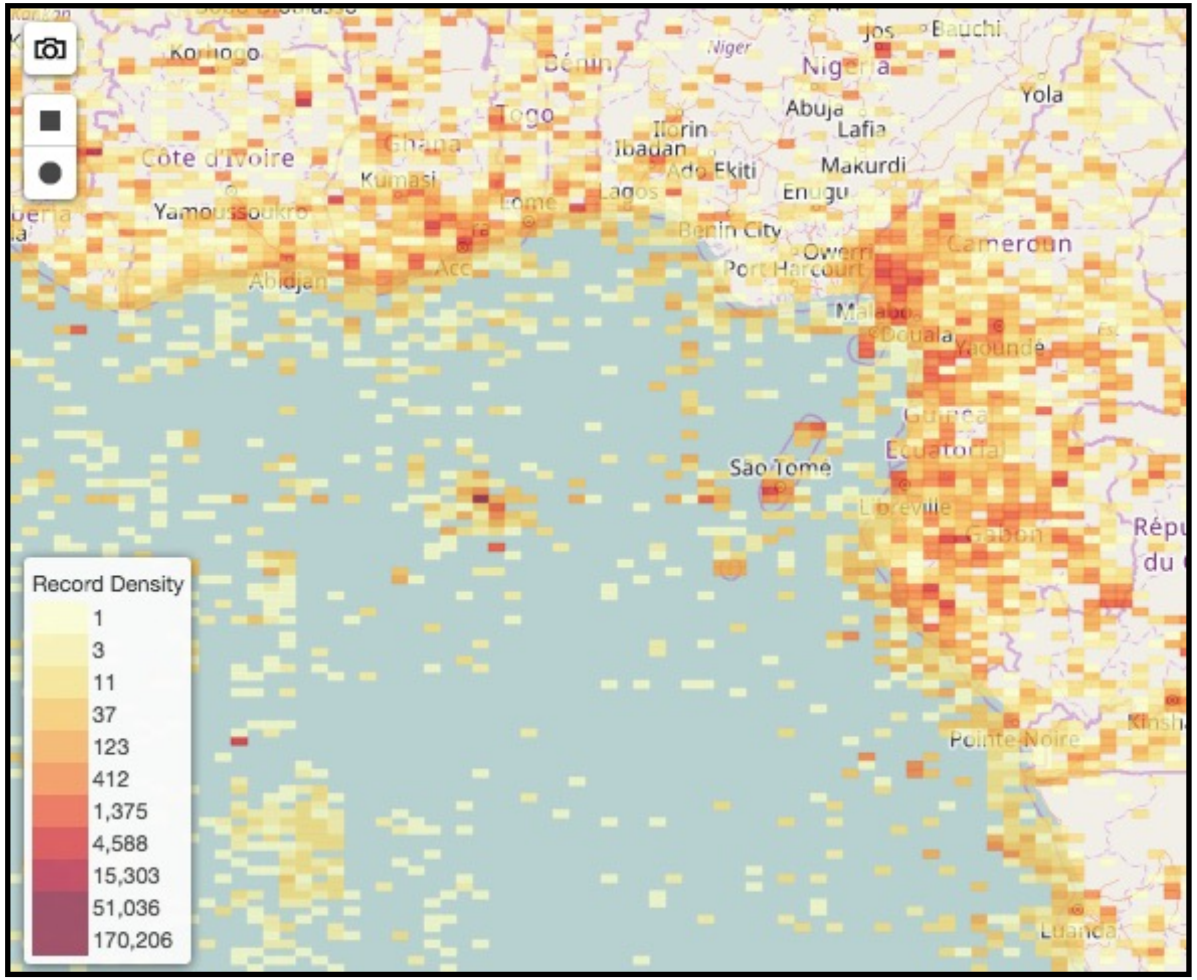
iDigBio Integrated Digitized Biocollections

NSF

BiotaPhy

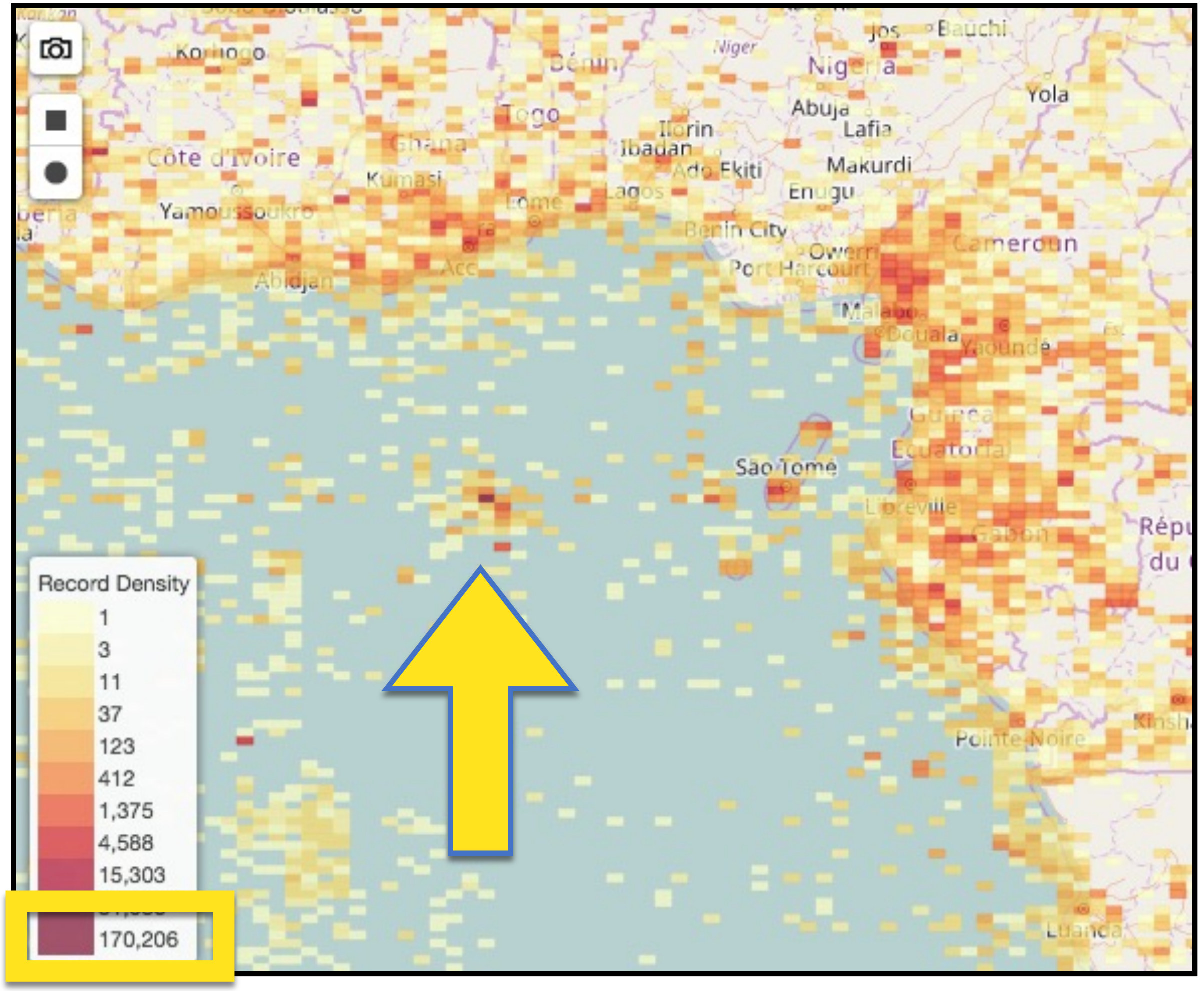
Occurrence Data Cleaning

1. **Resolve taxon names**
2. **Decrease number of columns**
3. **Remove duplicates**
4. Clean localities
   - Round up the latitude/longitude
   - Remove coordinates at 0,0
   - Remove coordinates in cultivated zones, botanical gardens, etc.
   - Remove coordinates outside of the desired range
5. Spatial correction
6. Produce a csv



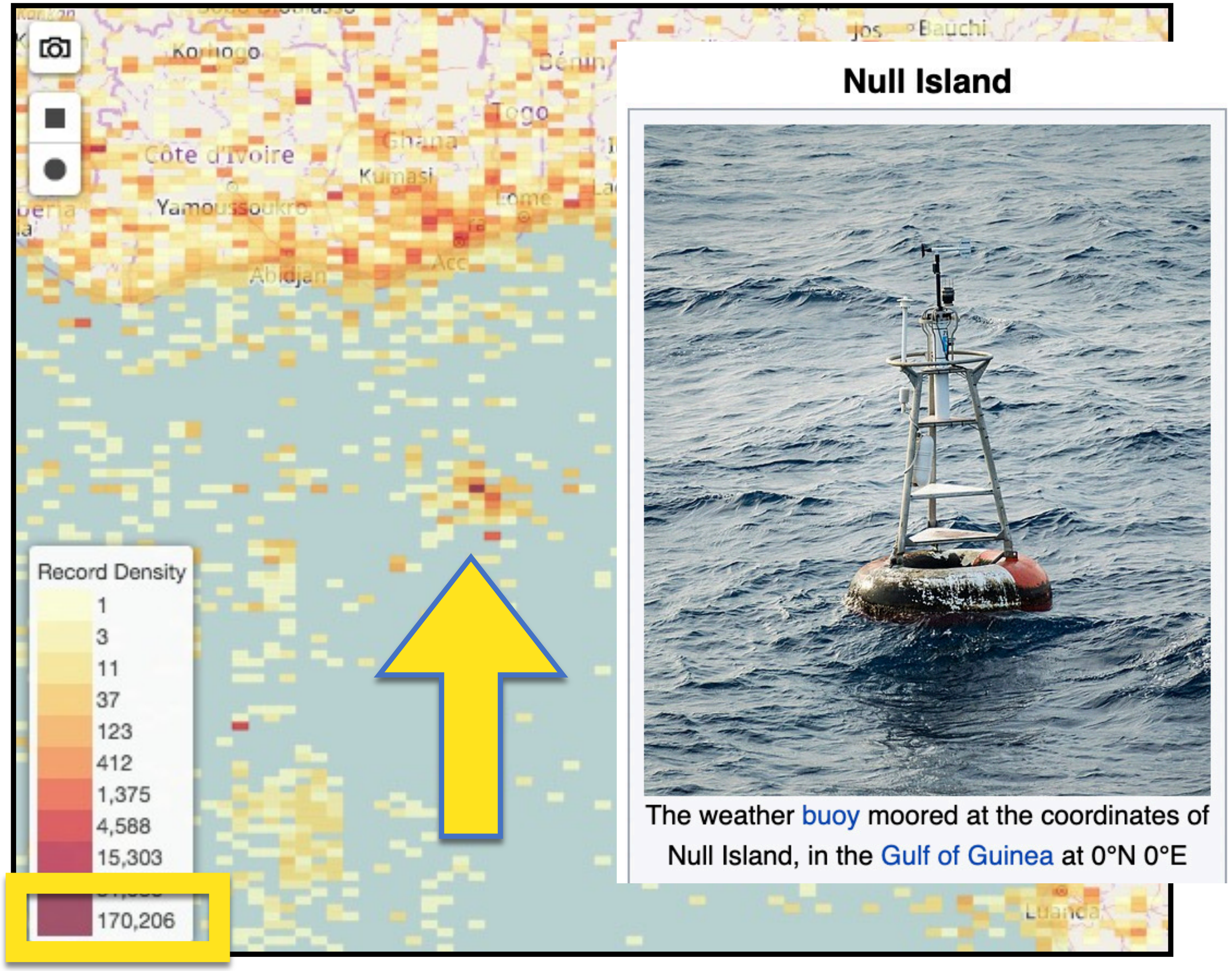
Occurrence Data Cleaning

1. Resolve taxon names
2. **Clean localities**
   - **Round up the latitude/longitude**
   - **Remove coordinates at 0,0**
   - Remove coordinates in cultivated zones, botanical gardens, etc.
   - Remove coordinates outside of the desired range
3. Remove duplicates
4. Spatial correction
5. Produce a csv



Occurrence Data Cleaning

1. Resolve taxon names
2. Decrease number of columns
3. Remove duplicates
4. **Clean localities**
   - **Round up the latitude/longitude**
   - **Remove coordinates at 0,0**
   - Remove coordinates in cultivated zones, botanical gardens, etc.
   - Remove coordinates outside of the desired range
5. Spatial correction
6. Produce a csv

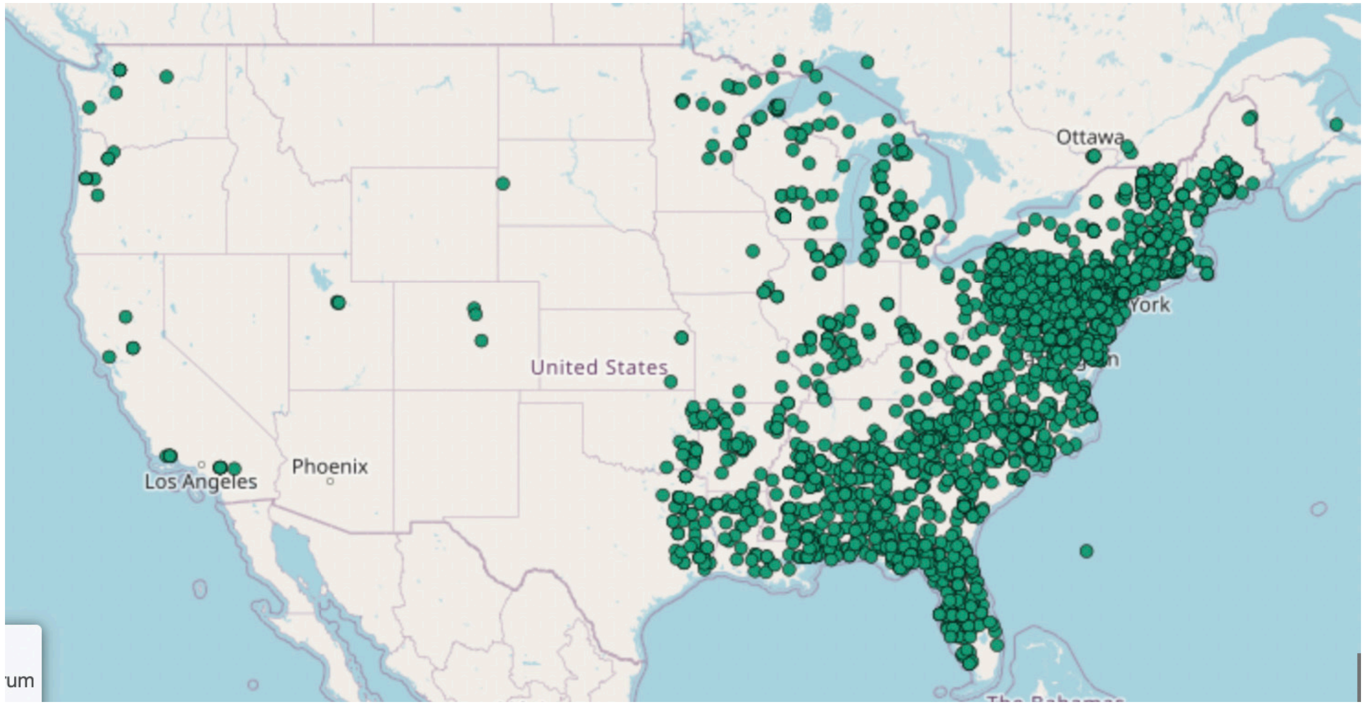


Null Island

The weather buoy moored at the coordinates of Null Island, in the Gulf of Guinea at 0°N 0°E

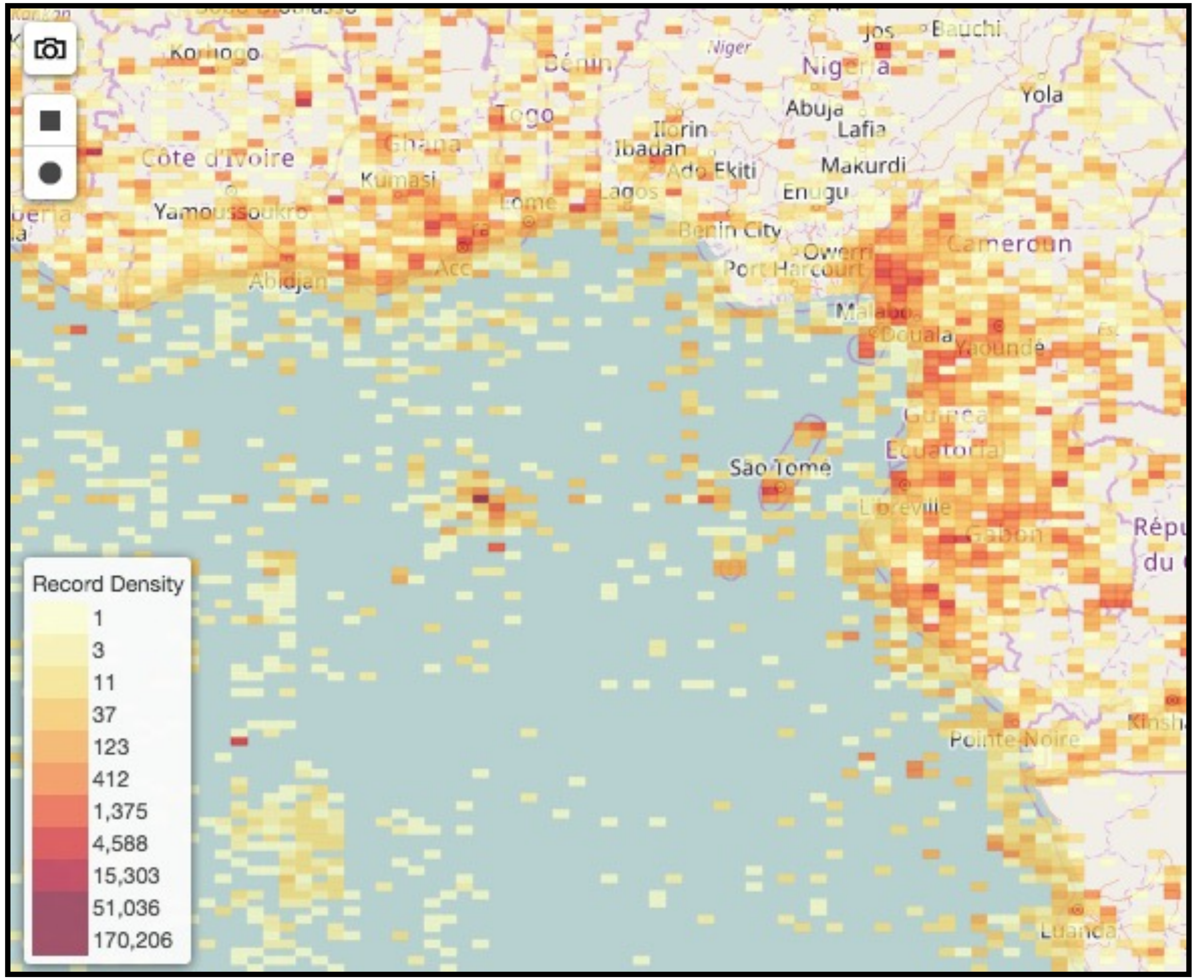
Occurrence Data Cleaning

1. Resolve taxon names
2. Decrease number of columns
3. Remove duplicates
4. **Clean localities**
   - Round up the latitude/longitude
   - Remove coordinates at 0,0
   - **Remove coordinates in cultivated zones, botanical gardens, etc.**
   - **Remove coordinates outside of the desired range**
5. Spatial correction
6. Produce a csv



Occurrence Data Cleaning

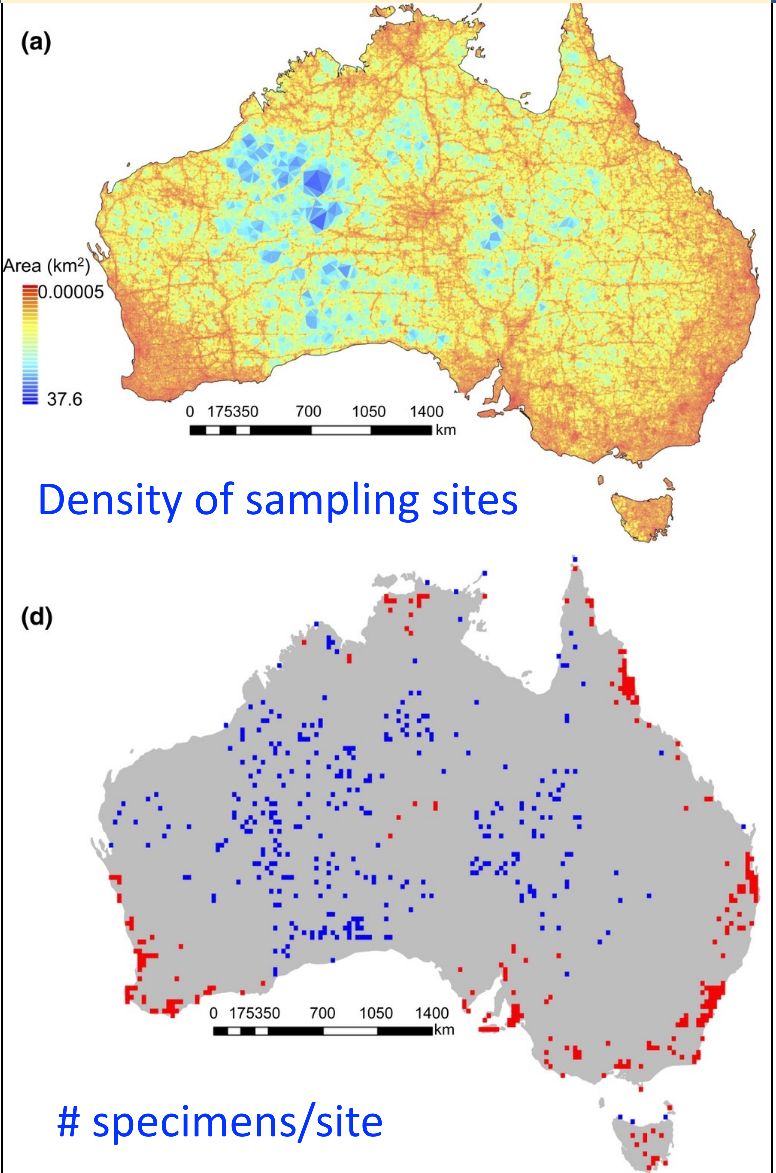
1. Resolve taxon names
2. Decrease number of columns
3. Remove duplicates
4. Clean localities
    - Round up the latitude/longitude
    - Remove coordinates at 0,0
    - Remove coordinates in cultivated zones, botanical gardens, etc.
    - Remove coordinates outside of the desired range
5. **Spatial correction**
6. Produce a csv



Spatial Correction

- Collection efforts can lead to clustering of points
  - Infrastructure (roads, herbaria, etc.)
  - Taxon bias
  - Temporal bias
- Filtering is a procedure to reduce the clustering of species records

Sampling bias for 857,245 sites



Density of sampling sites

# specimens/site

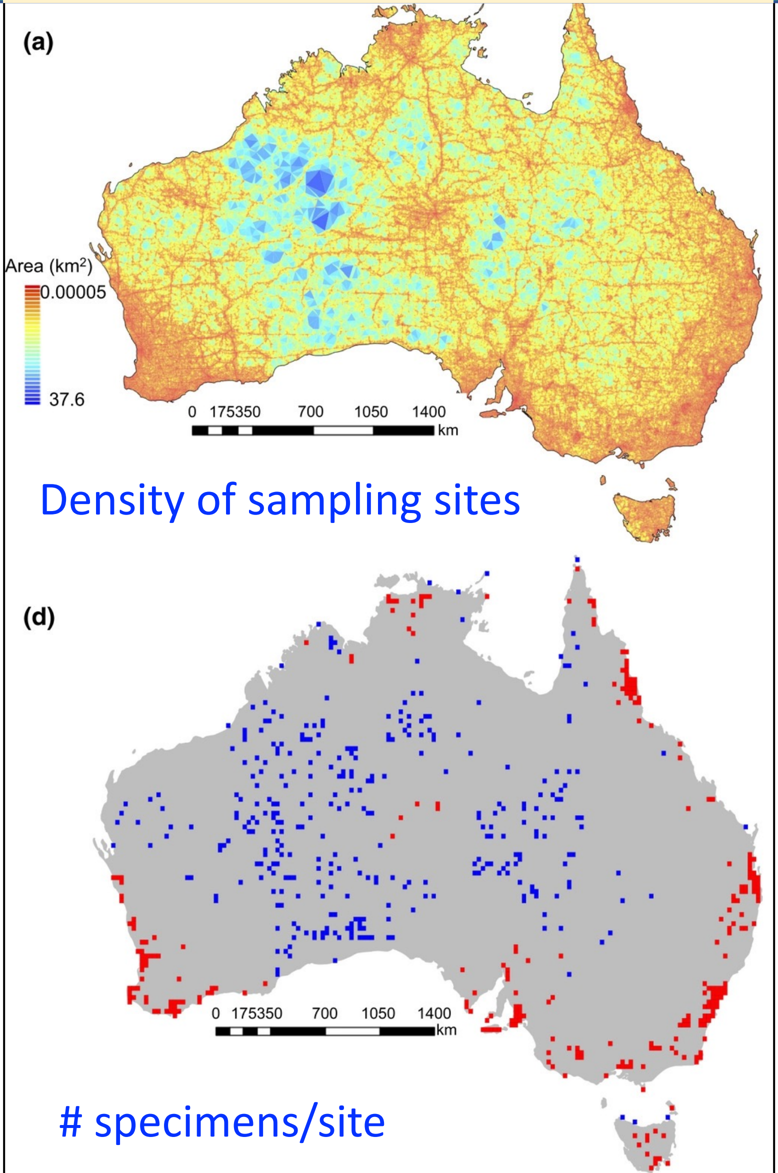
Daru et al. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. New Phytologist.

Spatial Correction

- Collection efforts can lead to clustering of points
  - Infrastructure (roads, herbaria, etc.)
  - Taxon bias
  - Temporal bias
- Filtering is a procedure to reduce the clustering of species records
- After filtering, there may still be spatial autocorrelation
  - This can be accounted for by data partitioning

Daru et al. 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. New Phytologist.

Sampling bias for 857,245 sites

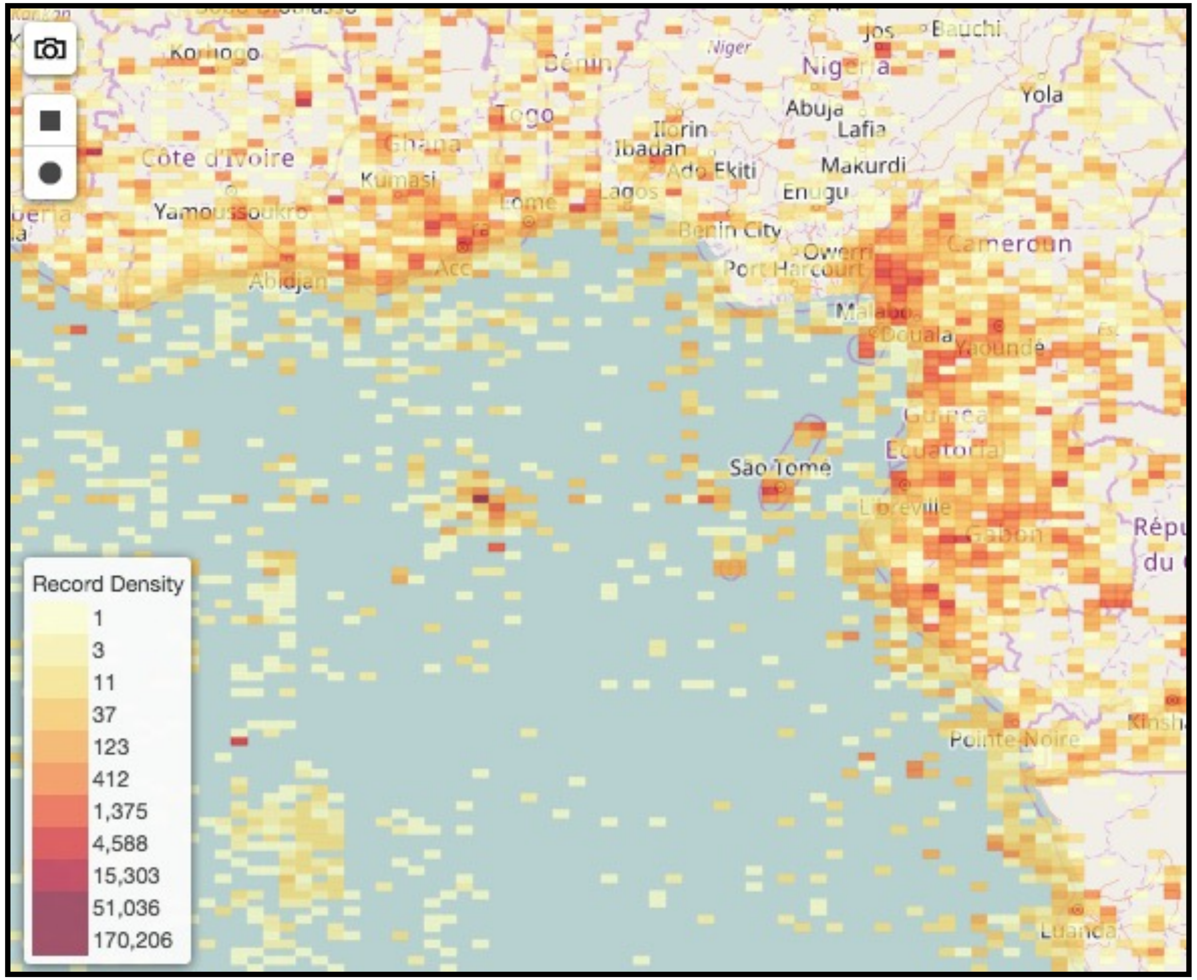


Density of sampling sites

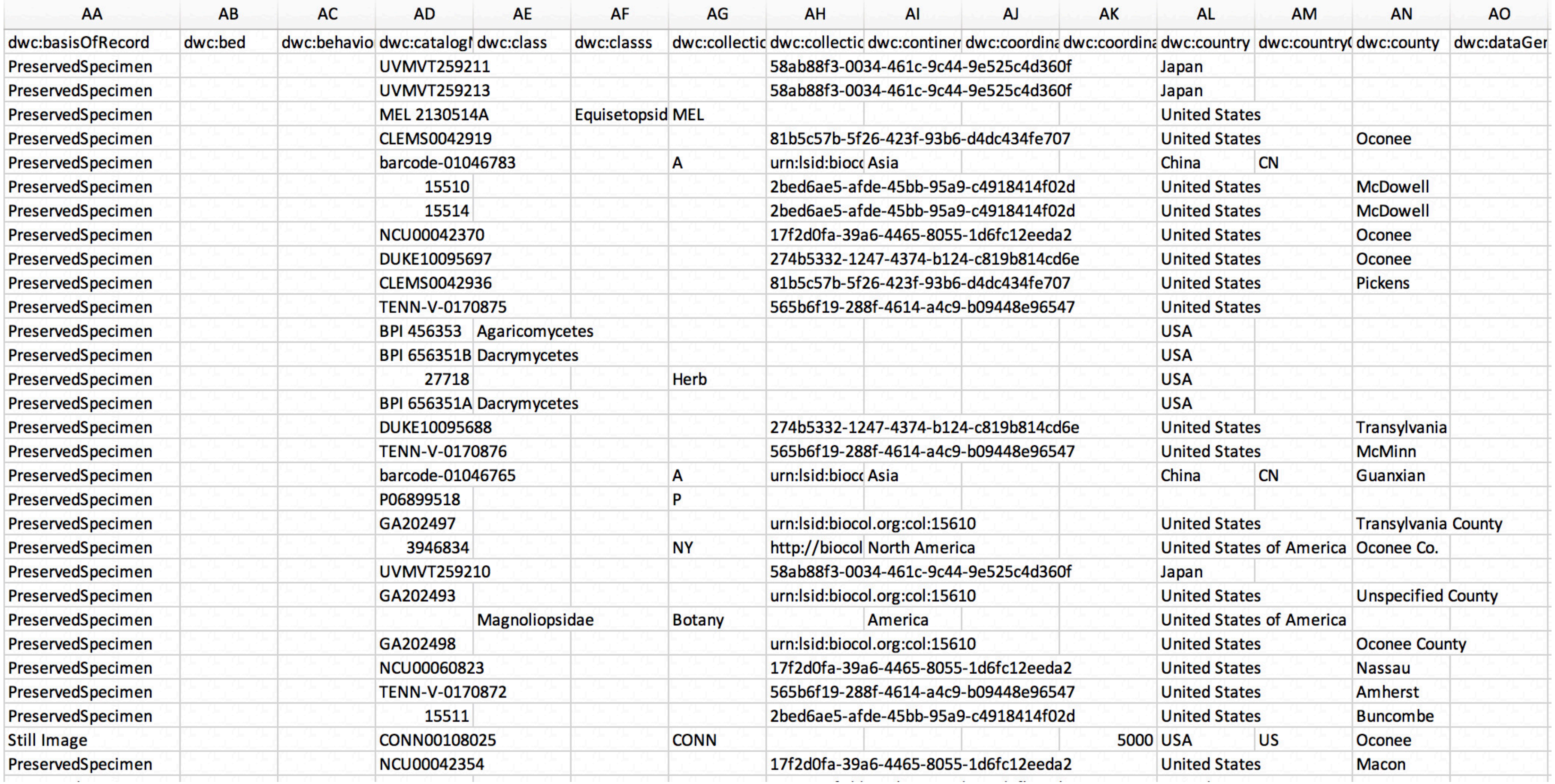
# specimens/site

Occurrence Data Cleaning

1. Resolve taxon names
2. Decrease number of columns
3. Remove duplicates
4. Clean localities
   - Round up the latitude/longitude
   - Remove coordinates at 0,0
   - Remove coordinates in cultivated zones, botanical gardens, etc.
   - Remove coordinates outside of the desired range
5. Spatial correction
6. **Produce a csv**



Producing a CSV File



AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO
dwc:basisOfRecord	dwc:bed	dwc:behavio	dwc:catalogN	dwc:class	dwc:classs	dwc:collectic	dwc:collectic	dwc:continer	dwc:coordina	dwc:coordina	dwc:country	dwc:country(	dwc:county	dwc:dataGer
PreservedSpecimen			UVMVT259211				58ab88f3-0034-461c-9c44-9e525c4d360f				Japan			
PreservedSpecimen			UVMVT259213				58ab88f3-0034-461c-9c44-9e525c4d360f				Japan			
PreservedSpecimen			MEL 2130514A		Equisetopsid	MEL					United States			
PreservedSpecimen			CLEMS0042919				81b5c57b-5f26-423f-93b6-d4dc434fe707				United States		Oconee	
PreservedSpecimen			barcode-01046783			A	urn:lsid:biocc	Asia			China	CN		
PreservedSpecimen			15510				2bed6ae5-afde-45bb-95a9-c4918414f02d				United States		McDowell	
PreservedSpecimen			15514				2bed6ae5-afde-45bb-95a9-c4918414f02d				United States		McDowell	
PreservedSpecimen			NCU00042370				17f2d0fa-39a6-4465-8055-1d6fc12eeda2				United States		Oconee	
PreservedSpecimen			DUKE10095697				274b5332-1247-4374-b124-c819b814cd6e				United States		Oconee	
PreservedSpecimen			CLEMS0042936				81b5c57b-5f26-423f-93b6-d4dc434fe707				United States		Pickens	
PreservedSpecimen			TENN-V-0170875				565b6f19-288f-4614-a4c9-b09448e96547				United States			
PreservedSpecimen			BPI 456353	Agaricomycetes							USA			
PreservedSpecimen			BPI 656351B	Dacrymycetes							USA			
PreservedSpecimen			27718			Herb					USA			
PreservedSpecimen			BPI 656351A	Dacrymycetes							USA			
PreservedSpecimen			DUKE10095688				274b5332-1247-4374-b124-c819b814cd6e				United States		Transylvania	
PreservedSpecimen			TENN-V-0170876				565b6f19-288f-4614-a4c9-b09448e96547				United States		McMinn	
PreservedSpecimen			barcode-01046765			A	urn:lsid:biocc	Asia			China	CN	Guanxian	
PreservedSpecimen			P06899518			P								
PreservedSpecimen			GA202497				urn:lsid:biocol.org:col:15610				United States		Transylvania County	
PreservedSpecimen			3946834			NY	http://biocol	North America			United States of America		Oconee Co.	
PreservedSpecimen			UVMVT259210				58ab88f3-0034-461c-9c44-9e525c4d360f				Japan			
PreservedSpecimen			GA202493				urn:lsid:biocol.org:col:15610				United States		Unspecified County	
PreservedSpecimen				Magnoliopsidae		Botany		America			United States of America			
PreservedSpecimen			GA202498				urn:lsid:biocol.org:col:15610				United States		Oconee County	
PreservedSpecimen			NCU00060823				17f2d0fa-39a6-4465-8055-1d6fc12eeda2				United States		Nassau	
PreservedSpecimen			TENN-V-0170872				565b6f19-288f-4614-a4c9-b09448e96547				United States		Amherst	
PreservedSpecimen			15511				2bed6ae5-afde-45bb-95a9-c4918414f02d				United States		Buncombe	
Still Image			CONN00108025			CONN				5000	USA	US	Oconee	
PreservedSpecimen			NCU00042354				17f2d0fa-39a6-4465-8055-1d6fc12eeda2				United States		Macon	

Geographic And Taxonomic Occurrence R-based Scrubbing (gatoRs): An R Package and Reproducible Workflow for Processing Biodiversity Data

Natalie Patten, Shelly Gaynor, Doug Soltis, & Pam Soltis