Supporting Information:

Biodiversity monitoring for a just planetary future

Abstract:

While rapidly growing repositories of biodiversity data provide unprecedented insight into ecological patterns at global scales, the application of species observations often belies the reality that the species these data tell us most about is the one they were never intended to include: humans. Biodiversity data trace not only cities and roads but the rise of surveillance technology, shadows of colonial histories, and echoes of contemporary racial and economic disparities. However, these same data are increasingly used as the starting point to inform the implementation of global policy and the investment of billions of dollars to protect and restore nature over the next decade. Effectively leveraging large-scale biodiversity data to benefit both people and nature requires expertise in social, cultural, and political processes underlying data infrastructures and their histories, just as much as it requires more data and increasingly complex statistical methods.

In this SI document, we synthesize examples of the social, political and economic dimensions of human society reflected in global biodiversity data. We provide data, code, and citations for reproducing the figure in "Biodiversity monitoring for a just planetary future" (doi:10.1126/science. adh8874)

Exploring social and political dimensions of biodiversity data

In this figure, we leverage the Global Biodiversity Information Facility (GBIF) occurrence data set to reproduce, visualize, update, and/or expand upon the cited social and political dimensions of biodiversity data presented in the paper.

GBIF releases full occurrence "snapshots" monthly. In this paper, we leverage the Sept 31, 2023 Snapshot, which was the most recent at the time of submission (https://doi.org/10.15468/dl.ua9nww) (GBIF.Org User 2023). This snapshot has approximate 2.6 billion occurrence records.

We provide code to reproduce each panel of Figure 1. All data used is accessed within the code and is openly available.

We use the following packages:

```r
library(geomtextpath)
library(duckdbfs)
library(gbifdb)
library(tidyverse)
library(fst)
library(sf)
library(terra)
#library(raster)
library(MetBrewer)
library(rnaturalearth)
library(countrycode)
library(arrow)
library(usmap)
library(svglite)
```

Package citations: (Cameron and Brand 2022; Mühleisen and Raasveldt 2023; Boettiger 2023; "Global Biodiversity Information Facility (GBIF) Species Occurrences" 2021; Wickham et al. 2019; Pebesma and Bivand 2023; Hijmans 2023b, 2023a; Mills 2022; Massicotte and South 2023; Arel-Bundock, Enevoldsen, and Yetman 2018; Richardson et al. 2023; Di Lorenzo 2023):

Connect to a GBIF snapshot

We use a local copy of the Sept 31, 2023 GBIF snapshot and the `gbifdb` package ("Global Biodiversity Information Facility (GBIF) Species Occurrences" 2021) to query the >2.6 billion observations in the database.

```r
gbif <- gbif_local("/home/shared-data/gbif/occurrence/2023-10-01/occurrence.parquet/",
                   backend="duckdb")
```

All analysis here can alternatively be done by querying the GBIF AWS snapshot (leveraging the `arrow` package (Richardson et al. 2023)) using the following code:

```r
#snapshot <- "s3://gbif-open-data-eu-central-1/occurrence/2023-10-01/occurrence.parquet"
#gbif <- open_dataset(gbif_snapshot)
```

Panel A: Global map

We summarize the count of observations at 0.1 decimal degrees. All observations in GBIF with coordinates are included in this map.

```
df <- gbif |>
  mutate(latitude = round(decimallatitude,2),
         longitude = round(decimallongitude,2)) |>
  #filter(year >1800) |>
  count(longitude, latitude) |>
  collect()
```

We convert the lat/long to spatial points using the `sf` package (Pebesma and Bivand 2023).

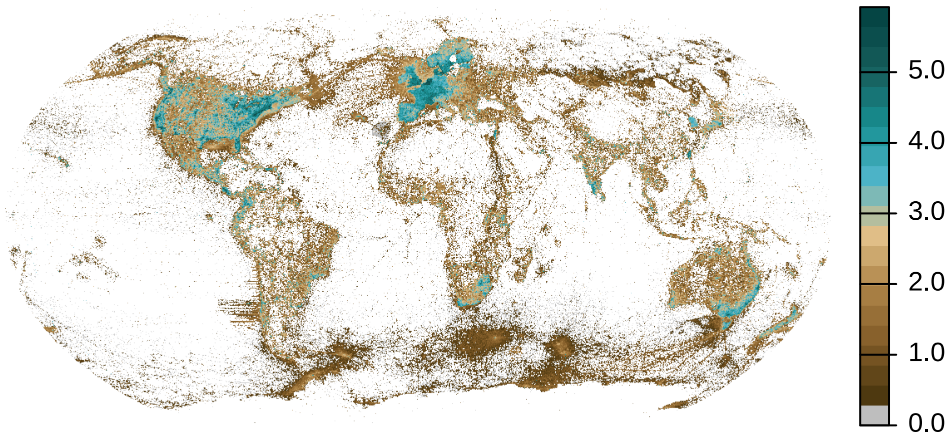
```
df_spatial <- df |>
  filter(!is.na(latitude),
         !is.na(longitude)) |>
  st_as_sf(coords = c("longitude", "latitude"),
           crs = "+proj=robin +lon_0=0 +x_0=0 +y_0=0 +datum=WGS84 +units=m")
```

The log of the sum of observations at each point is converted into a global raster at 0.1 degrees.

```
library(raster)
ras_temp <-raster(xmn=-180, xmx=180, ymn=-90, ymx=90,
                  resolution=c(0.1,0.1), vals=NA)
global_plot_all <- rasterize(df_spatial, ras_temp,
                             field = "n", fun='sum')
#rm(df_spatial) #remove unnecessary data
rm(ras_temp) #remove unnecessary data
```

Reproject to the Robinson projection and plot using the terra package (Hijmans 2023b).

```
crs <- "+proj=robin +lon_0=0 +x_0=0 +y_0=0 +datum=WGS84 +units=m"
global_plot_all <- terra::rast(global_plot_all)
global_plot <- global_plot_all * 1 # to deal with NAs in this dataset
# reproject for viz
global_plot_r <- terra::project(global_plot, crs, mask=TRUE)
# define color gradient
colors <- c("grey", met.brewer(name="Isfahan1",n=20,type="continuous"))
# take log for viz
terra::plot(log(global_plot_r,10), col = colors, axes = FALSE)
```




```
writeRaster(global_plot_r, "../data/panels/PanelA_data.tif", overwrite=TRUE)
```

Panel B: Macroeconomic patterns

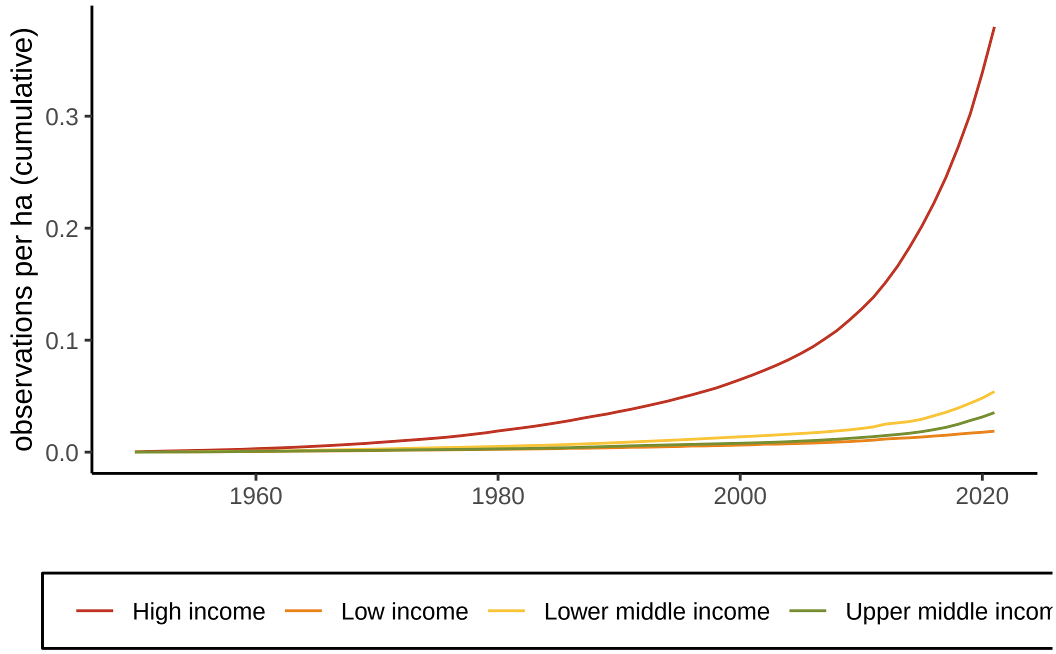
In panel B, we show the cumulative number of observations per hectare collected in countries across different income groups. The World Bank classifies economies for analytical purposes into four income groups: **low, lower-middle, upper-middle, and high income**. We use these for our analysis.

```
world <- ne_countries(type = "countries", scale = "medium")
world <- st_as_sf(world) |>
  dplyr::select(iso_a2, income_grp) |>
  st_make_valid() |>
  mutate(area = st_area(geometry)) |>
  as_tibble() |>
  dplyr::select(-geometry) |>
  mutate(area = as.numeric(area)/10000) |>
  rename(countrycode = iso_a2)
```

Collect count of observations per year per country

```r
country_year <-  gbif |>
    count(countrycode, year) |>
    collect()
```

```r
macroeconomics_density_cumulative <- country_year |>
  filter(year >1949 & year < 2022) |>
  mutate(n = replace_na(n, 0)) |>
  left_join(world) |>
  mutate(income_grp = str_sub(income_grp, 4, -1),
         #INCOME_GRP = gsub("\\s", "\n", INCOME_GRP),
         income_grp = gsub("\\:.*","",income_grp)) |>
  group_by(year, income_grp) |>
  summarise(n = sum(n, na.rm = TRUE),
            area = sum(area)) |> ungroup()|>
  group_by(income_grp) |>
  mutate(cum_obs = cumsum(n)) |>
  ungroup() |>
  mutate(density = cum_obs/area) |>
  drop_na() |>
  ggplot(aes(year, density, color = income_grp, label = income_grp)) +
  geom_line() +
  #geom_textline(size = 3, fontface = 2, spacing = 30, text_smoothing = 50) +
  theme_classic() +
  theme(legend.position = "bottom", legend.title = element_blank()) +
  labs(x = "", y = "observations per ha (cumulative)") +
  scale_color_manual(values=met.brewer("Homer2", 4)) +
  theme(legend.background =
          element_rect(colour = 'black', fill = 'white', linetype='solid'))
macroeconomics_density_cumulative
```



Panel C: Redlining

Redlining data

Redlining data is downloaded from Mapping Inequality (https://dsl.richmond.edu/panorama/redlining/) (Robert K. Nelson, n.d.). Here we reproduce the patterns found in Ellis-Soto et al., 2023 (Ellis-Soto, Chapman, and Locke 2023) (with an updated snapshot of GBIF data) showing disparities in the density of bird data throughout redlined neighborhoods in the United States.

In the figure, we compare neightborhoods that were deemed (holc grade A) to those deemed most hazardous (holc grade D).

```
# download redlining geojson
holc <-
  st_read("https://dsl.richmond.edu/panorama/redlining/static/mappinginequality.json") |>
  rename(holc_grade = grade) |>
  dplyr::select(state, city, holc_grade, geometry) |>
  dplyr::filter(!is.na(holc_grade) & holc_grade != 'E') |>
  sf::st_make_valid() |>
  mutate(holc_grade = ifelse(holc_grade == "A ", "A",
                        ifelse(holc_grade == "B ", "B",
```

```
                                    ifelse(holc_grade == "C ", "C",
                                           ifelse(holc_grade == "D ", "D", holc_grade)))))
```

```
Reading layer `mappinginequality' from data source
  `https://dsl.richmond.edu/panorama/redlining/static/mappinginequality.json'
  using driver `GeoJSON'
Simple feature collection with 10150 features and 11 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: -122.7675 ymin: 25.70537 xmax: -69.60044 ymax: 48.2473
Geodetic CRS:  WGS 84
```

```
# remove inalid polygons
holc <- holc |>
  dplyr::mutate(valid =st_is_valid(holc)) |>
  dplyr::filter(valid=="TRUE")
# calculate area per neighborhood
holc_area  <- holc |>
  mutate(area = st_area(geometry)) |>
  as_tibble() |> drop_na() |>
  dplyr::select(-geometry) |>
  group_by(holc_grade) |>
  summarise(area = sum(area)) |>
  group_by(holc_grade) |>
  summarise(area = sum(area)) |>
  mutate(area = as.numeric(area)/10000) # area in ha
```

US observations

Species observations in class "Aves" in the US are queried and summerized at 0.001 decimal degrees.

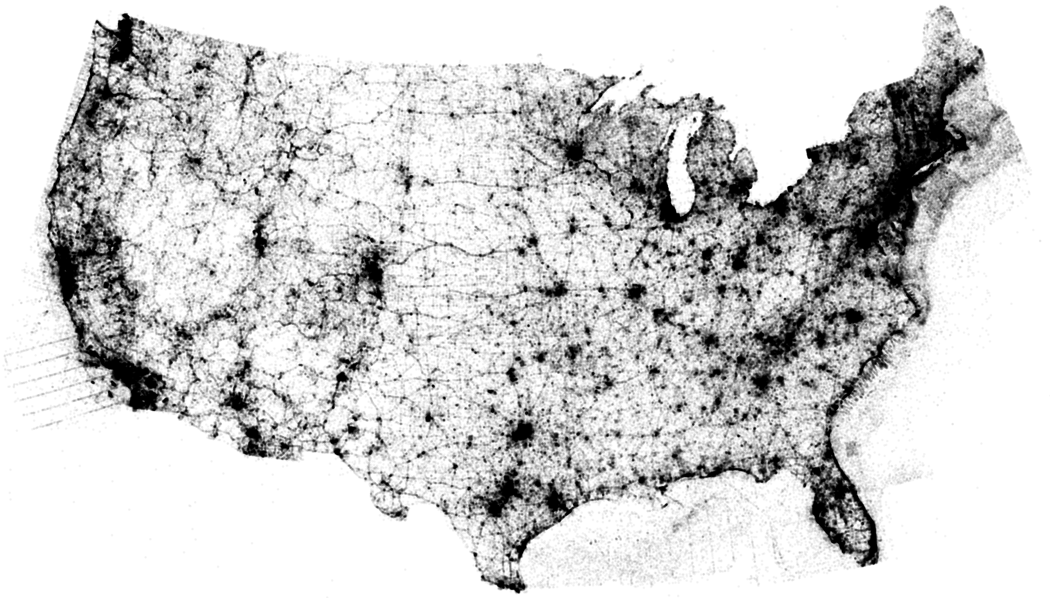
```
states <- sf::st_as_sf(maps::map("state",
                                 plot = FALSE, fill = TRUE))
ext_states <- ext(states)

US_pts <- gbif |>
   filter(countrycode == "US") |>
   filter(class == "Aves") |>
   mutate(latitude = round(decimallatitude,3),
```

```
        longitude = round(decimallongitude,3)) |>
  count(latitude, longitude) |>
  collect() |>
  filter(longitude > ext_states[1] & longitude < ext_states[2] &
           latitude > ext_states[3] & latitude < ext_states[4])
```

These points are converted to a spatial data frame and reprojected. For the US map we plot a subset of 1 million points to allow for visualization of point density in cities and around major roads.

```
US_pts_sf_all <- US_pts |>
  filter(!is.na(latitude),
        !is.na(longitude)) |>
  st_as_sf(coords = c("longitude", "latitude"),
           crs = st_crs(holc))

US_pts_sf <- US_pts_sf_all |>  head(1000000) |>
  st_transform(crs = usmap::usmap_crs())
```

```
plot_gbif <- US_pts_sf |>
  ggplot() +
  geom_sf(aes(geometry = geometry), alpha = 0.05,
            size = 0.001, color= "black") +
  theme(legend.position = "none") +
  theme_void()

plot_gbif
```

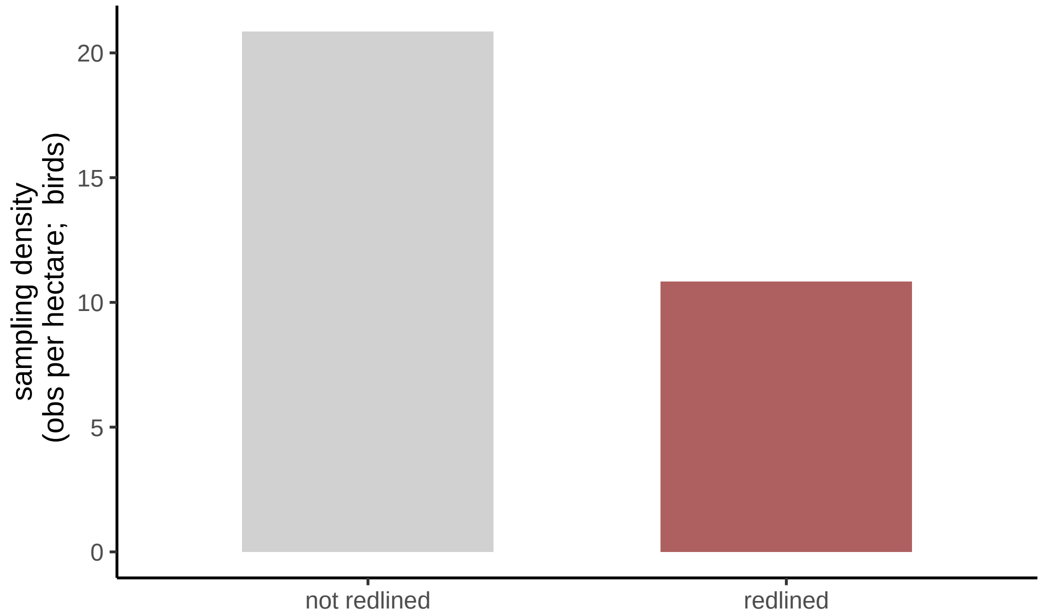


Redlining observations summary

The bar chart in Panel C shows the number of aves observations per unit area in grade A vs holc grade D across all cities included in the Mapping Inequality dataset.

```
holc_obs <- st_join(US_pts_sf_all, holc, join = st_within)
```

```r
redline_area <- holc_area |>
  filter(holc_grade %in% c("A", "D")) |>
  #filter(holc_grade %in% c("A", "B", "C", "D")) |>
  mutate(redlined = ifelse(holc_grade == "D", "redlined", "not redlined")) |>
  group_by(redlined) |>
  summarise(area = sum(area))

redlining <- as_tibble(holc_obs) |>
  drop_na() |>
  filter(holc_grade %in% c("A", "D")) |>
  #filter(holc_grade %in% c("A", "B", "C", "D")) |>
  mutate(redlined = ifelse(holc_grade == "D", "redlined", "not redlined")) |>
  group_by(redlined) |>
  summarise(counts = sum(n)) %>% ungroup() |>
  left_join(redline_area) |>
  mutate(density = counts/area) |> drop_na() |>
  ggplot(aes(x = redlined, y = density, fill = redlined)) +
  scale_fill_manual(
    values=c("grey", "firebrick4")) +
  geom_col(width = 0.6, alpha = 0.7) + theme_classic() +
  theme(legend.position = "none") +
  labs(y = " sampling density \n (obs per hectare;  birds)", x = "")

as_tibble(holc_obs) |>
  drop_na() |>
  filter(holc_grade %in% c("A", "D")) |>
  #filter(holc_grade %in% c("A", "B", "C", "D")) |>
  mutate(redlined = ifelse(holc_grade == "D", "redlined", "not redlined")) |>
  group_by(redlined) |>
  summarise(counts = sum(n)) %>% ungroup() |>
  left_join(redline_area) |>
  mutate(density = counts/area) |> drop_na() |>
  write_csv("../data/panels/redlining_simplified.csv")

redlining
```



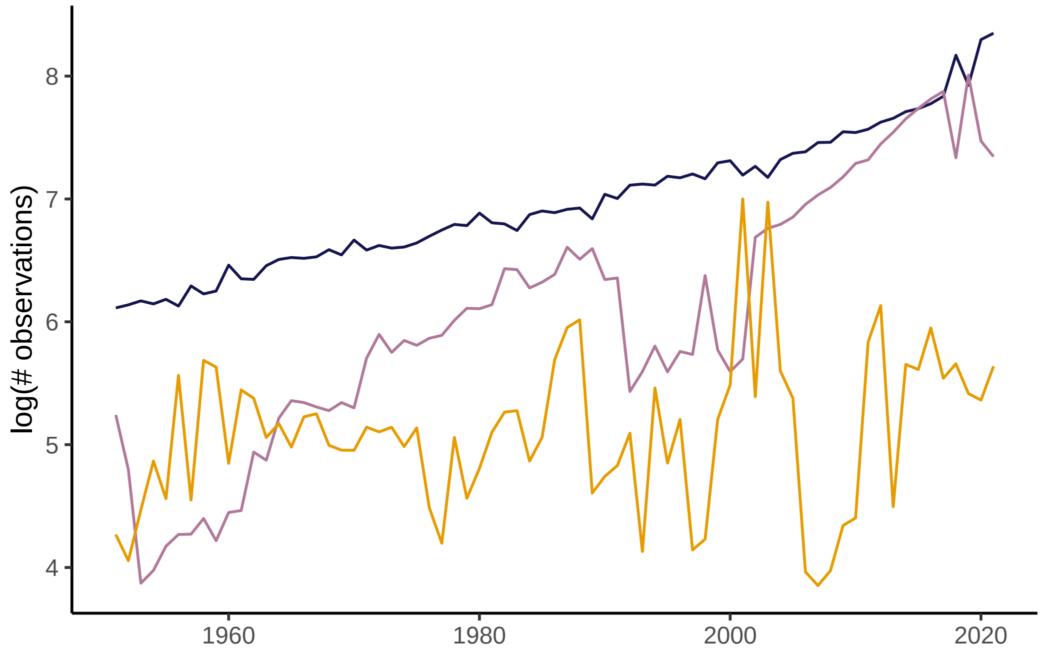
Panel D: Conflict

Following analysis in (Zizka et al. 2021) and leveraging the yearly conflict data from the Uppsala Conflict Data Program (UCDP) (Davies, Pettersson, and Öberg 2023; GLEDITSCH et al. 2002), we show how biodiversity data observations track conflict both (i) globally and (ii) in Cambodia (Zizka et al. 2021) since 1950.

```
country_year <-  gbif |>
    count(countrycode, year) |>
    collect()
```

```
#download.file(url = "https://ucdp.uu.se/downloads/ucdpprio/ucdp-prio-acd-231-rds.zip", dest
#unzip("data/conflict.zip", exdir = "data/")

conflict <- readRDS("../data/UcdpPrioConflict_v23_1.rds") |>
  dplyr::select(location, year, intensity_level) |>
  separate_rows(location, sep = ",") |>
  mutate(countrycode =
           countrycode(location,
                       origin = 'country.name',
                       destination = 'iso2c'))
```

```
conflict_plot_i <- country_year |>
  left_join(conflict) |>
  dplyr::select(-location) |>
  filter(year >1950 & year <2022) |>
  mutate(intensity_level = replace_na(intensity_level, 0)) |>
  mutate(n = replace_na(n, 0)) |>
  group_by(year, intensity_level) |>
  summarise( n = sum(n, na.rm = TRUE)) |>
  unique() |>
  ggplot() +
  geom_line(aes(year, log(n,10), col = as.factor(intensity_level), group=as.factor(intensity_
  #scale_color_manual(values= c("darkgrey", "#FF4433", "darkred")) +
  scale_color_manual(values=met.brewer("Renoir", 4)) +
  #geom_line(aes(year, v2xcl_dmove*10), lwd = 1.5, color = "black") +
  theme_classic() +
  theme(legend.position = "none", axis.title.x = element_blank()) +
  scale_y_continuous(
    # Features of the first axis
    name = "log(# observations)"#,
    # Add a second axis and specify its features
    #sec.axis = sec_axis( trans=~./10, name="freedom of movement")
  ) #+ geom_line(aes(x = year, y = intensity_level*5)) +

conflict_plot_i
```




```
conflict_plot_ii <- country_year |>
  filter(countrycode == "KH") |> filter(year >1949 & year < 2022) |> arrange(-year) |>
  left_join(conflict, by = c("countrycode", "year")) |>
  dplyr::select(-location) |>
  filter(year >1949 & year < 2022) |>
  mutate(intensity_level = replace_na(intensity_level, 0)) |>
  mutate(n = replace_na(n, 0)) |>
  group_by(countrycode, year) |> count()
  summarise(intensity_level = max(intensity_level),
            n = mean(n, na.rm = TRUE)) |>
  unique() |>
  ggplot() +
  geom_col(aes(year, (log(n)), fill = as.factor(intensity_level))) +
  #scale_fill_manual(values= c("darkgrey", "#FF4433", "darkred")) +
  scale_fill_manual(values=met.brewer("Renoir", 4)) +
  #geom_line(aes(year, v2xcl_dmove*10), lwd = 1.5, color = "black") +
  theme_classic() +
  theme(legend.position = "none", axis.title.x = element_blank()) +
  scale_y_continuous(
    # Features of the first axis
    name = "log(# observations)"#,
    # Add a second axis and specify its features
```

```
    #sec.axis = sec_axis( trans=~./10, name="freedom of movement")
  ) #+ geom_line(aes(x = year, y = intensity_level*5)) +
```

```
Error in h(simpleError(msg, call)): error in evaluating the argument 'x' in selecting a metho
```

```
conflict_plot_ii
```

```
# A tibble: 69 x 3
# Groups:   countrycode, year [69]
   countrycode  year     n
   <chr>       <dbl> <int>
 1 KH           1950     1
 2 KH           1951     1
 3 KH           1952     1
 4 KH           1953     1
 5 KH           1954     1
 6 KH           1955     1
 7 KH           1957     1
 8 KH           1958     1
 9 KH           1959     1
10 KH           1960     1
# i 59 more rows
```

```
country_year |>
  left_join(conflict) |>
  dplyr::select(-location) |>
  filter(year >1949 & year <2022) |>
  mutate(intensity_level = replace_na(intensity_level, 0)) |>
  mutate(n = replace_na(n, 0)) |>
  group_by(year, intensity_level) |>
  summarise(n = sum(n, na.rm = TRUE)) |>
  unique() |>
  write_csv("../data/panels/panelD_i_data.csv")
```

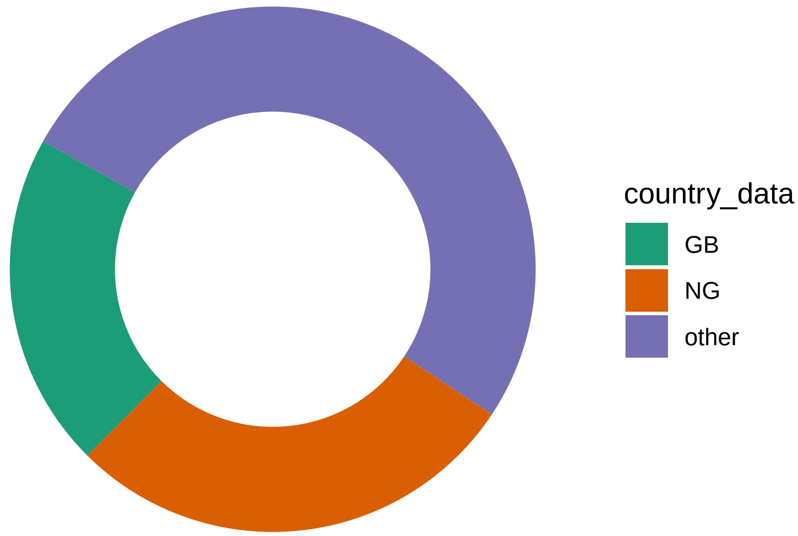
Panel E: Colonialism

As explored in (Zizka et al. 2021), social and political factors impact who has collected biodiversity data. We reproduce this analysis using the update GBIF snapshot. We can see that the publishing country before and after Nigeria's independence (1960) is drastically different (Zizka et al. 2021).

```r
NG_year <- gbif |>
    filter(countrycode == "NG") |>
    count(year, basisofrecord, datasetkey) |>
    collect()

# download dataset keys to keep track of publishing country
orgs <- read_tsv("https://api.gbif.org/v1/dataset/search/export?format=TSV&") |>
  dplyr::select(publishing_country, dataset_key, title) |>
  rename(datasetkey = dataset_key)

NG_year_summary <- NG_year |>
  left_join(orgs) |>
  mutate() |>
  mutate(country_data = ifelse(publishing_country == "GB", "GB",
                               ifelse(publishing_country == "NG", "NG", "other"))) |>
  mutate(precol = ifelse(year <1961,"1","2")) |>
  group_by(country_data, precol) |>
  summarise(n = sum(n)) |>
  drop_na()
```

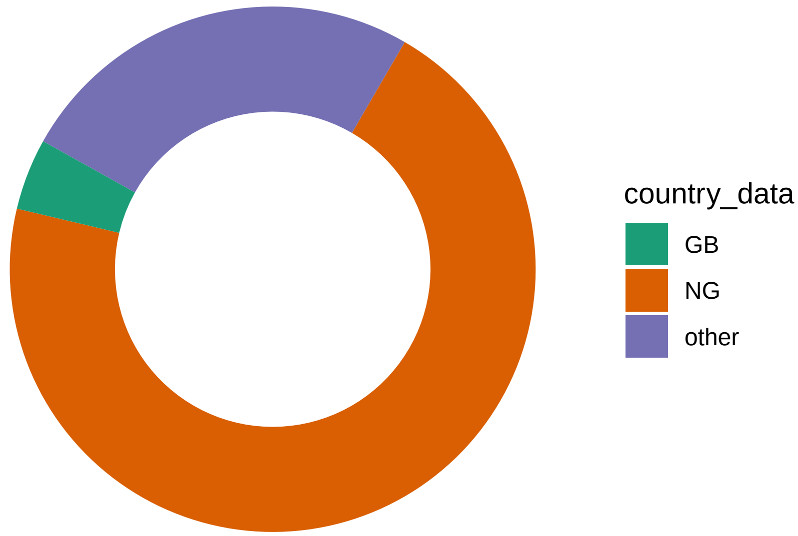
Observations in Nigeria pre-independence

```r
t1 <- NG_year_summary |>
  group_by(precol) |> mutate(total_obs = sum(n)) |> ungroup()|>
  mutate(perc_obs = n/total_obs) |> filter(precol == "1") |>
  arrange(desc(perc_obs)) %>%
  mutate(lab.pos = cumsum(perc_obs)-.5*perc_obs)

panelE_i <- ggplot(data = t1,
       aes(x = 2, y = perc_obs, fill = country_data))+
  geom_bar(stat = "identity")+
  coord_polar("y", start = 200) +
  theme_void() +
  scale_fill_brewer(palette = "Dark2") + xlim(.2,2.5)

panelE_i
```



Observations in Nigeria post-independence

```
t2 <- NG_year_summary |>
  group_by(precol) |> mutate(total_obs = sum(n)) |> ungroup()|>
  mutate(perc_obs = n/total_obs) |> filter(precol == "2") |>
  arrange(desc(perc_obs)) %>%
  mutate(lab.pos = cumsum(perc_obs)-.5*perc_obs)

panelE_ii <- ggplot(data = t2,
       aes(x = 2, y = perc_obs, fill = country_data))+
  geom_bar(stat = "identity")+
  coord_polar("y", start = 200) +
  theme_void() +
  scale_fill_brewer(palette = "Dark2") + xlim(.2,2.5)

panelE_ii
```



References

Arel-Bundock, Vincent, Nils Enevoldsen, and CJ Yetman. 2018. "Countrycode: An r Package to Convert Country Names and Country Codes" 3: 848. https://doi.org/10.21105/joss.00848.

Boettiger, Carl. 2023. "Duckdbfs: High Performance Remote File System Access Using 'Duckdb'." https://github.com/cboettig/duckdbfs.

Cameron, Allan, and Teun van den Brand. 2022. "Geomtextpath: Curved Text in 'Ggplot2'." https://allancameron.github.io/geomtextpath/.

Davies, Shawn, Therése Pettersson, and Magnus Öberg. 2023. "Organized Violence 1989–2022, and the Return of Conflict Between States." *Journal of Peace Research* 60 (4): 691–708. https://doi.org/10.1177/00223433231185169.

Di Lorenzo, Paolo. 2023. "Usmap: US Maps Including Alaska and Hawaii." https://usmap.dev.

Ellis-Soto, Diego, Melissa Chapman, and Dexter H. Locke. 2023. "Historical Redlining Is Associated with Increasing Geographical Disparities in Bird Biodiversity Sampling in the United States." *Nature Human Behaviour*, September. https://doi.org/10.1038/s41562-023-01688-5.

GBIF.Org User. 2023. "Occurrence Download." The Global Biodiversity Information Facility. https://doi.org/10.15468/DL.UA9NWW.

GLEDITSCH, NILS PETTER, PETER WALLENSTEEN, MIKAEL ERIKSSON, MARGARETA SOLLENBERG, and HÅVARD STRAND. 2002. "Armed Conflict 1946-2001: A New Dataset." *Journal of Peace Research* 39 (5): 615–37.

https://doi.org/10.1177/0022343302039005007.

"Global Biodiversity Information Facility (GBIF) Species Occurrences." 2021. https://registry.opendata.aws/gbif.

Hijmans, Robert J. 2023a. "Raster: Geographic Data Analysis and Modeling." https://rspatial.org/raster.

———. 2023b. "Terra: Spatial Data Analysis." https://rspatial.org/.

Massicotte, Philippe, and Andy South. 2023. "Rnaturalearth: World Map Data from Natural Earth." https://docs.ropensci.org/rnaturalearth/ https://github.com/ropensci/rnaturalearth.

Mills, Blake Robert. 2022. "MetBrewer: Color Palettes Inspired by Works at the Metropolitan Museum of Art."

Mühleisen, Hannes, and Mark Raasveldt. 2023. "Duckdb: DBI Package for the DuckDB Database Management System." https://duckdb.org/.

Pebesma, Edzer, and Roger Bivand. 2023. "{Spatial Data Science: With Applications in r}." https://doi.org/10.1201/9780429459016.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, and Apache Arrow. 2023. "Arrow: Integration to 'Apache' 'Arrow'." https://github.com/apache/arrow/.

Robert K. Nelson, Richard Marciano, LaDale Winling. n.d. "Mapping Inequality." https://dsl.richmond.edu/panorama/redlining/#loc=5/39.1/-94.58&text=downloads.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the {Tidyverse}" 4: 1686. https://doi.org/10.21105/joss.01686.

Zizka, Alexander, Oskar Rydén, Daniel Edler, Johannes Klein, Allison Perrigo, Daniele Silvestro, Sverker C. Jagers, Staffan I. Lindberg, and Alexandre Antonelli. 2021. "BioDem, a Tool to Explore the Relationship Between Biodiversity Data Availability and Socio-Political Conditions in Time and Space." *Journal of Biogeography* 48 (11): 2715–26. https://doi.org/10.1111/jbi.14256.