calculate_ratio_2000_2020: Disparity A/D ratio change (2000–2020)

Diego Ellis Soto

2026-01-14

Overview

This document produces the calculation of the 2000–2020 A/D disparity using two parallel pipelines that yield the same percentage.

(Ellis-Soto, Diego, Melissa Chapman, and Dexter H. Locke. "Historical redlining is associated with increasing geographical disparities in bird biodiversity sampling in the United States." Nature Human Behaviour 7.11 (2023): 1869-1877.)

[1] Tidyverse style coding for our 2023 manuscript

[2] A replicator-style implementation using data.table, closely matching the structure and objects used in the replication report.

The goal is to make explicit how the A/D ratio is computed, and to clarify how a mixed-area expression could have arrised by the replicators (e.g. the 39.8%).

First pipeline (original 2023 manuscript implementation)

This section reproduces the 2000–2020 A/D disparity using tidyverse code.

Although the data manipulation steps differ (plyr/tidyverse vs. data.table), the final ratio is computed using the same definition: year-specific sampling density (n_obs / area_sum) yielding the same 39.5%.

Replicator-style calculation (data.table)

Replicator-style calculation (data.table)

This section reproduces the A/D percent-change calculation using the replication approach (data.table + HOLC-grade area totals) which yield 39.5%.

We also proceed to print the mixed-area expression as text (not evaluated) to make transparent where the object ttb enters the ratio in the replication script (unintentionally?) which leads to a 39.8% estimate, as opposed to 39.5%. (e.g., "rep_SI_v5.R to L1413"),

Second pipeline (original 2023 manuscript implementation)

This section reproduces the same 2000–2020 A/D disparity using the original analysis workflow from the 2023 manuscript.

Although the data manipulation steps differ (plyr/tidyverse vs. data.table), the final ratio is computed using the same definition: year-specific sampling density (n_obs / area_sum).

```
suppressPackageStartupMessages({
require(sf)
require(tidycensus)
require(readr)
require(dplyr)
require(tidyr)
  })


holc_area = read_csv('../../indir/Biodiv_Greeness_Social/main_combined_2022-05-27.csv')  %>%
  dplyr::select(city, holc_grade, area_holc_km2) %>%
  dplyr::group_by(holc_grade) %>%
   dplyr::filter(holc_grade != 'E') %>%
  dplyr::summarise(area_sum = sum(area_holc_km2))
```

```
## Rows: 9851 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): id, state, city, holc_id, holc_grade, city_state, msa_NAME
## dbl (25): area_holc_km2, holc_tot_pop, msa_GEOID, msa_M, msa_p, msa_H, msa_e...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
temporal_trend = read.table('../../indir/Biodiv_Greeness_Social/R1_biodiv_trend_by_time_holc_id_1933_202

names(temporal_trend) <- c('Year','holc_grade', 'Sum')
temporal_trend = temporal_trend %>% dplyr::filter(holc_grade != 'E')

temporal_all_data <- temporal_trend %>%
  group_by(holc_grade, Year) %>%
  dplyr::summarise(
    n_obs = sum(Sum, na.rm = TRUE),
    .groups = "drop"
  ) |>
    dplyr::left_join(holc_area, by = "holc_grade")

# Nor rounding:
ratios <- temporal_all_data %>%
  dplyr::filter(holc_grade %in% c("A","D"), Year %in% c(2000, 2020)) %>%
  dplyr::mutate(year_density = n_obs / area_sum) %>%
  dplyr::select(Year, holc_grade, year_density) %>%
  tidyr::pivot_wider(names_from = holc_grade, values_from = year_density) %>%
  dplyr::filter(!is.na(A), !is.na(D)) %>%
  dplyr::mutate(A_D_ratio = A / D) %>%
  dplyr::select(Year, A_D_ratio)

ratios %>%
  tidyr::pivot_wider(names_from = Year, values_from = A_D_ratio,
                     names_prefix = "ratio_") %>%
  dplyr::mutate(
    percent_change = round((ratio_2020 / ratio_2000 - 1) * 100, 1),
    ratio_2000 = round(ratio_2000, 2),
    ratio_2020 = round(ratio_2020, 2)
  ) |>
  dplyr::pull(percent_change)
```

```
## [1] 39.5
```

Replicator data.table based disparity calculation.

We next reproduce the replication team's data.table-based calculation of the A/D disparity. The first `dispar_tta` object implements the correctly matched-area ratio and yields the same 39.5% estimate as the tidyverse pipeline above. For transparency, we also print the replication script's mixed-area ratio expression (not evaluated), in which the 2000 A-grade denominator is drawn from `ttb` while the D-grade denominator is drawn from `tta`. Because `tta` and `ttb` use slightly different HOLC area normalizations, this produces an unintended hybrid ("mixed-area") A/D ratio that slightly alters the percent-change estimate (39.8%).

```
suppressPackageStartupMessages({
library(data.table)
library(here)
  })


holc_a <- fread("../../indir/Biodiv_Greeness_Social/soc_dem_max_2022_03_12 17_31_11.csv")
holc_area_sum_a_dt <- holc_a[, .(sum_area_holc_km2 = sum(area_holc_km2)), by = holc_grade]

t <- fread("../../indir/Biodiv_Greeness_Social/R1_biodiv_trend_by_time_holc_id_1933_2022.csv")
setnames(t, c("year", "holc_grade", "Sum"))
t <- t[holc_grade != "E"]

tt <- t[, .(n_obs = sum(Sum)), by = .(year, holc_grade)]
setorder(tt, holc_grade, year)

tta <- merge(tt, holc_area_sum_a_dt, all.x = TRUE)
tta[, sampling_density := n_obs/sum_area_holc_km2]

tta[, holc_grade_D := factor(holc_grade, levels = c("D","B","C","A"))]

dispar_tta <- round(
(
(tta[year == 2020 & holc_grade == "A", sampling_density] /
tta[year == 2020 & holc_grade == "D", sampling_density]) /
(tta[year == 2000 & holc_grade == "A", sampling_density] /
tta[year == 2000 & holc_grade == "D", sampling_density])
- 1
) * 100,
1
)
dispar_tta
```

```
## [1] 39.5
```

```r
cat(
  "Mixed-area A/D disparity (2000-2020):\n",
  "round((((",
  "tta[year %in% c(2020) & holc_grade %in% c('A'), sampling_density] / ",
  "tta[year %in% c(2020) & holc_grade %in% c('D'), sampling_density]) / ",
  "(ttb[year %in% c(2000) & holc_grade %in% c('A'), sampling_density] / ",
  "tta[year %in% c(2000) & holc_grade %in% c('D'), sampling_density])",
  ") - 1) * 100, 1\n",
  sep = ""
)
```

```
## Mixed-area A/D disparity (2000-2020):
## round((((tta[year %in% c(2020) & holc_grade %in% c(’A’), sampling_density] / tta[year %in% c(2020) &
```

```r
sessionInfo()
```

```
## R version 4.4.1 (2024-06-14)
## Platform: aarch64-apple-darwin20
## Running under: macOS Sonoma 14.6
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/Los_Angeles
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] here_1.0.1        data.table_1.17.0 tidyr_1.3.1       dplyr_1.1.4
## [5] readr_2.1.5       tidycensus_1.7.1  sf_1.0-21
##
## loaded via a namespace (and not attached):
##  [1] rappdirs_0.3.3     generics_0.1.4     class_7.3-23       xml2_1.3.8
##  [5] KernSmooth_2.23-26 stringi_1.8.7      hms_1.1.3          digest_0.6.37
##  [9] magrittr_2.0.4     evaluate_1.0.3     grid_4.4.1         fastmap_1.2.0
## [13] rprojroot_2.0.4    jsonlite_2.0.0     processx_3.8.6     e1071_1.7-16
## [17] chromote_0.4.0     DBI_1.2.3          ps_1.9.1           promises_1.3.3
## [21] httr_1.4.7         rvest_1.0.4        purrr_1.2.0        cli_3.6.5
## [25] rlang_1.1.6        tigris_2.1         crayon_1.5.3       units_0.8-7
## [29] bit64_4.6.0-1      withr_3.0.2        yaml_2.3.10        parallel_4.4.1
## [33] tools_4.4.1        tzdb_0.5.0         uuid_1.2-1         vctrs_0.6.5
## [37] R6_2.6.1           proxy_0.4-27       lifecycle_1.0.4    classInt_0.4-11
## [41] stringr_1.6.0      bit_4.6.0          vroom_1.6.5        pkgconfig_2.0.3
## [45] pillar_1.11.1      later_1.4.2        glue_1.8.0         Rcpp_1.1.0
## [49] xfun_0.52          tibble_3.3.0       tidyselect_1.2.1   rstudioapi_0.17.1
## [53] knitr_1.50         htmltools_0.5.8.1  websocket_1.4.2    rmarkdown_2.29
## [57] compiler_4.4.1
```

This reproduces the replication script's mixed-area ratio (tta vs ttb).

Using different area denominators across years creates an unintended hybrid ratio,

which slightly alters the percent-change estimate relative to the matched-area calculation.