Some exploration of MAG data quality

Flavio Hafner (minor changes by Christoph)

18 August, 2022

Contents

Aggregate statistics by cohort . . . 2
Average number of new authors per year . . . 2
Average career length . . . 3
Average career publication rate . . . 4
Career length . . . 4
Author count and gender share by field-cohort and region-cohort . . . 5
By field-cohort . . . 6
By region-cohort . . . 8
How good is the assignment of authors to fields? Aggregate statistics by cohort and FieldClass . . 10
Average count of FieldOfStudyId per AuthorId-FieldClass by cohort . . . 10
Average Herfindahl index per author . . . 10
Share of authors with "moderate" or "high" concentration in a field . . . 12
Normalized HHI . . . 12
Who are these authors with very low HHI? . . . 13

MAG institution coverage over time **14**
Fraction of author-paper combinations with non-missing affiliation . . . 16
Fraction of papers with non-missing affiliation . . . 17
Fraction of authors with non-missing affiliation . . . 18
What do we learn? . . . 19

```
cat("Distribution of authors across FieldClass by missing FieldOfStudyId: \n")
```

```
## Distribution of authors across FieldClass by missing FieldOfStudyId:
```

```
print(missing_fields)
```

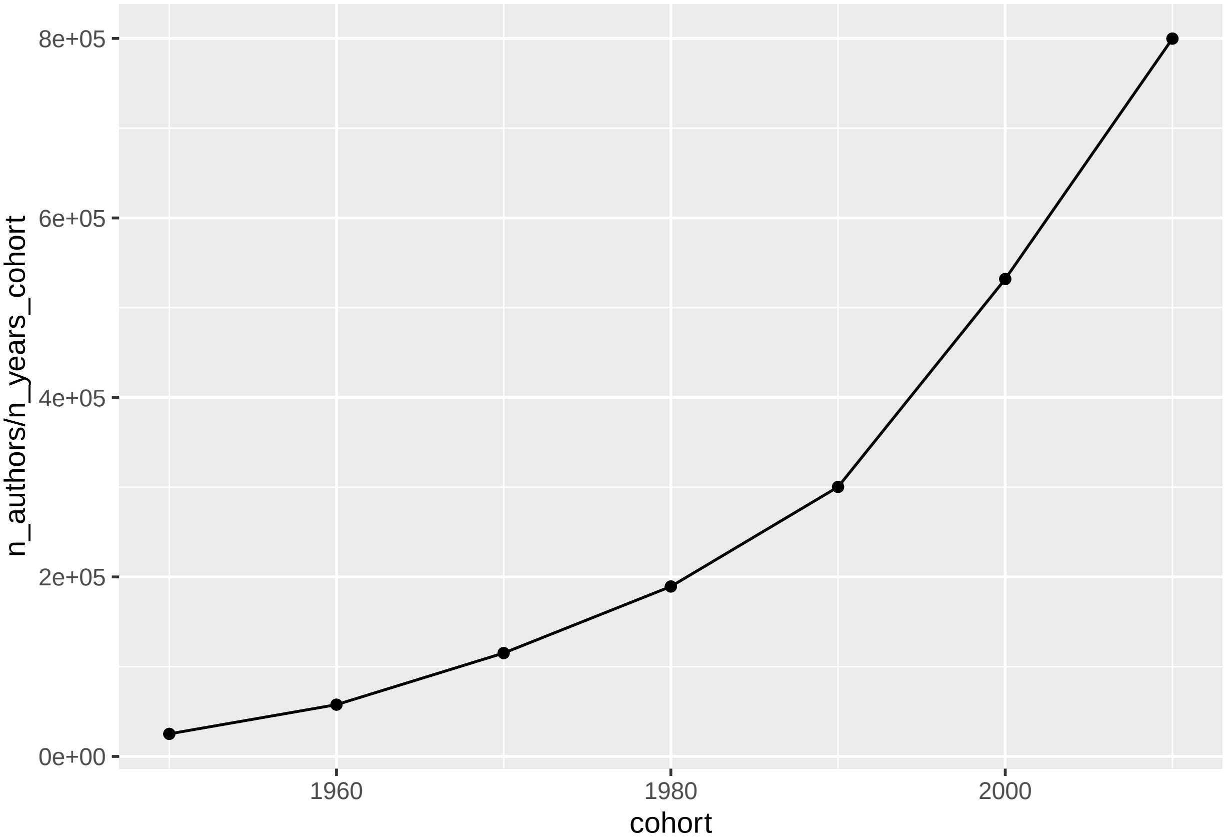
```
## # A tibble: 5 x 5
##   FieldClass field_missing n_authors mean_career_length mean_paper_count
##   <chr>              <dbl>     <int>              <dbl>            <dbl>
## 1 first                  0  16898470               8.25             12.2
## 2 first                  1     96123              15.0               5.85
## 3 last                   0  16921034               8.27             12.2
## 4 last                   1     73578              12.5               3.84
## 5 main                   0  16995166               8.28             12.2
```

```
cat("Authors with missing fields are dropped from now.")
```

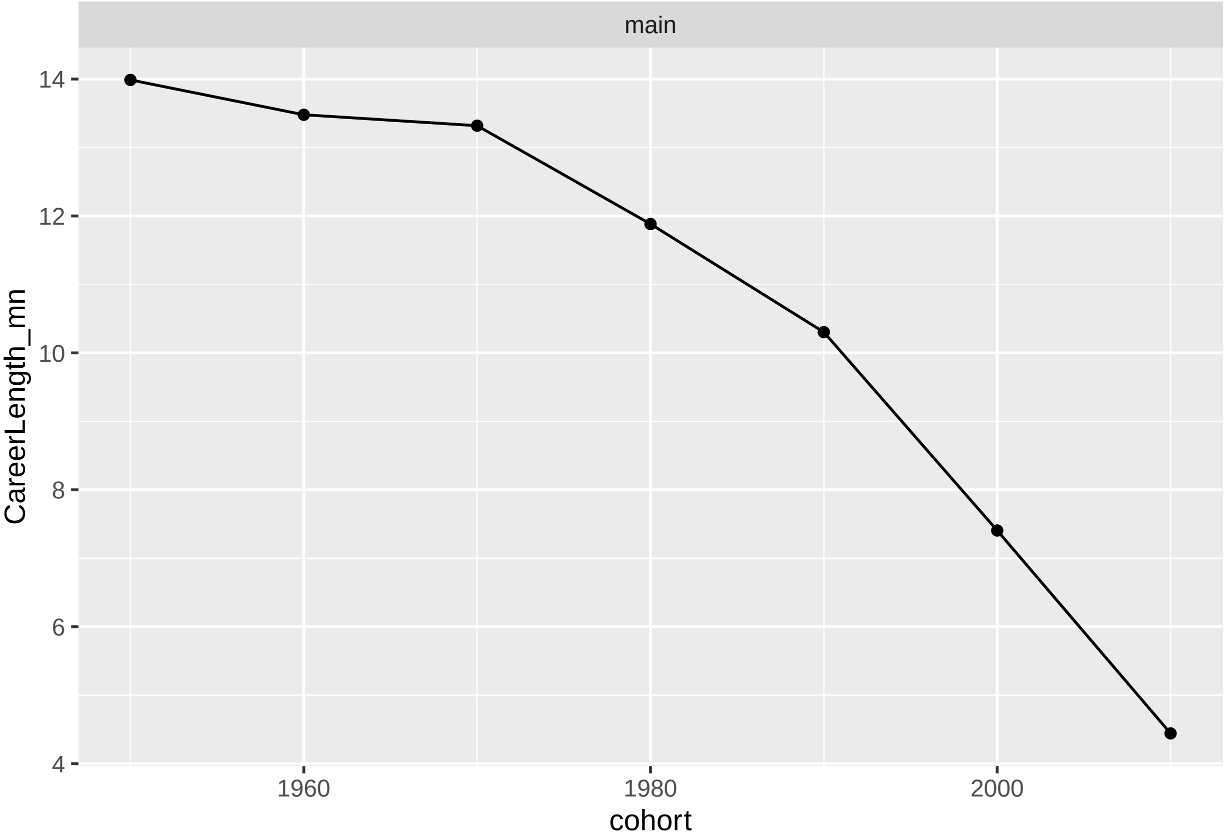
```
## Authors with missing fields are dropped from now.
```

Aggregate statistics by cohort

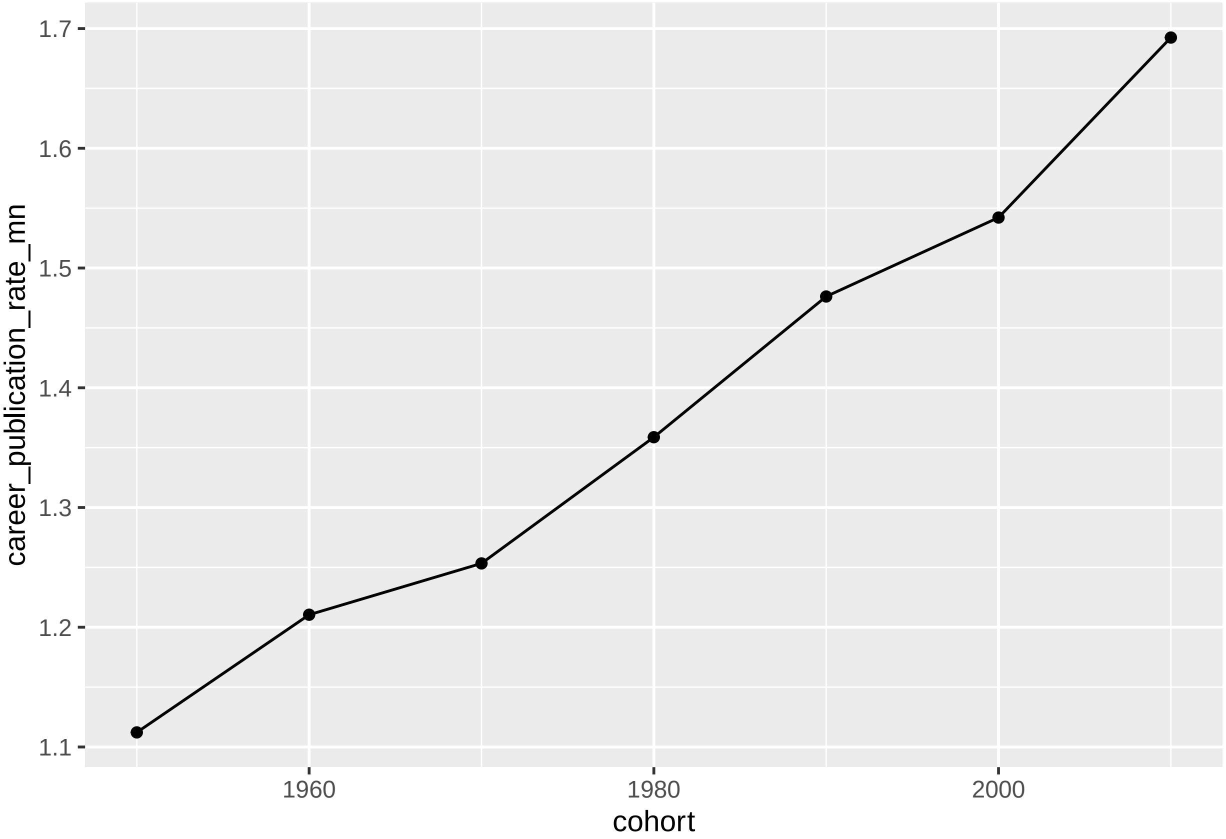
Average number of new authors per year



Average career length



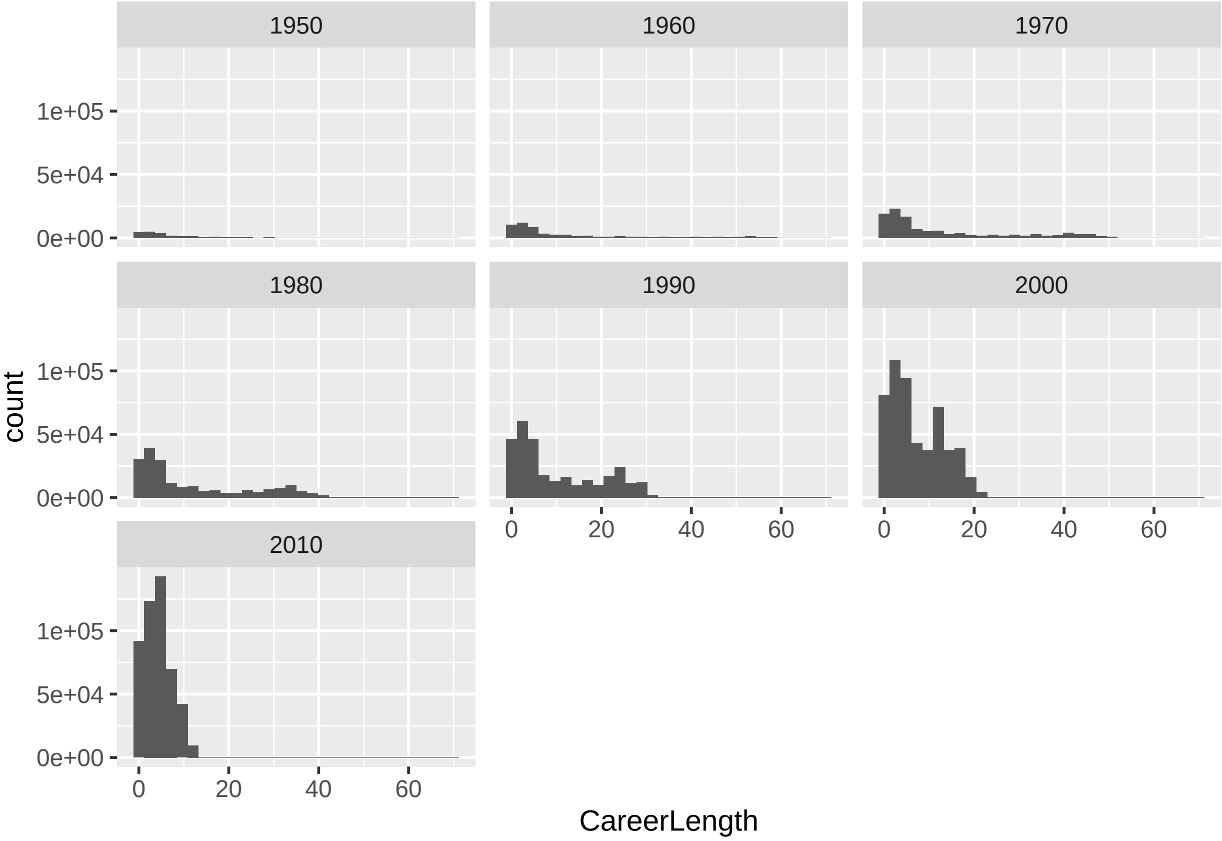
Average career publication rate



Career length

- 10 percent subsample of authors
- The “discontinuous” drop in career length density is at 6 years

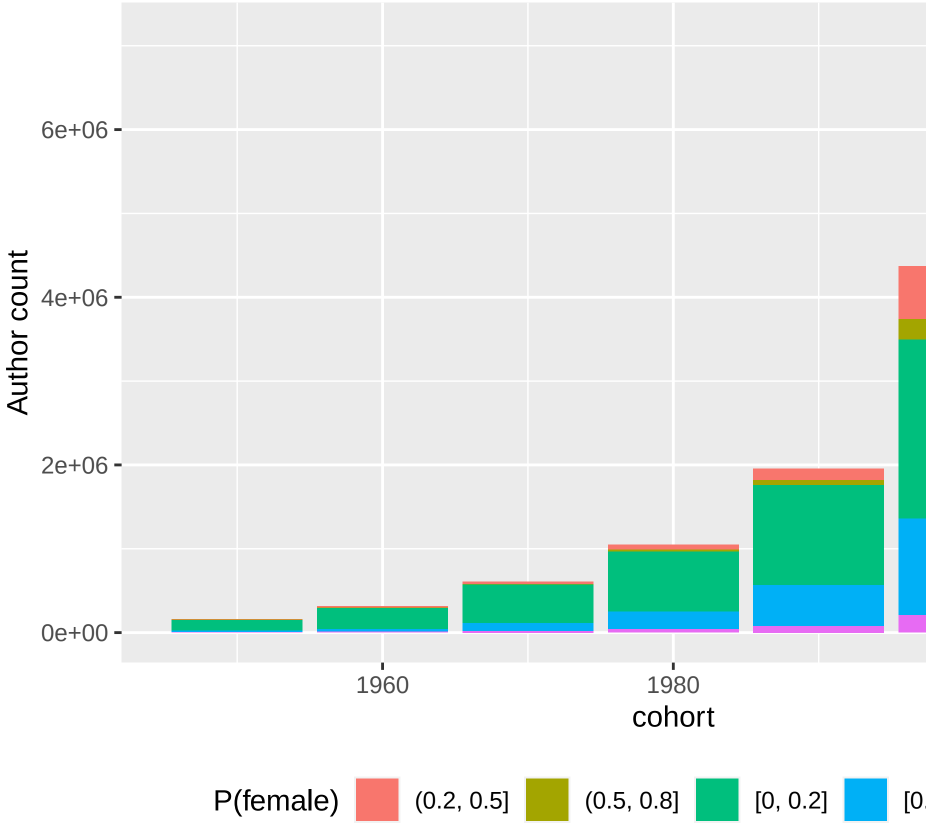
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



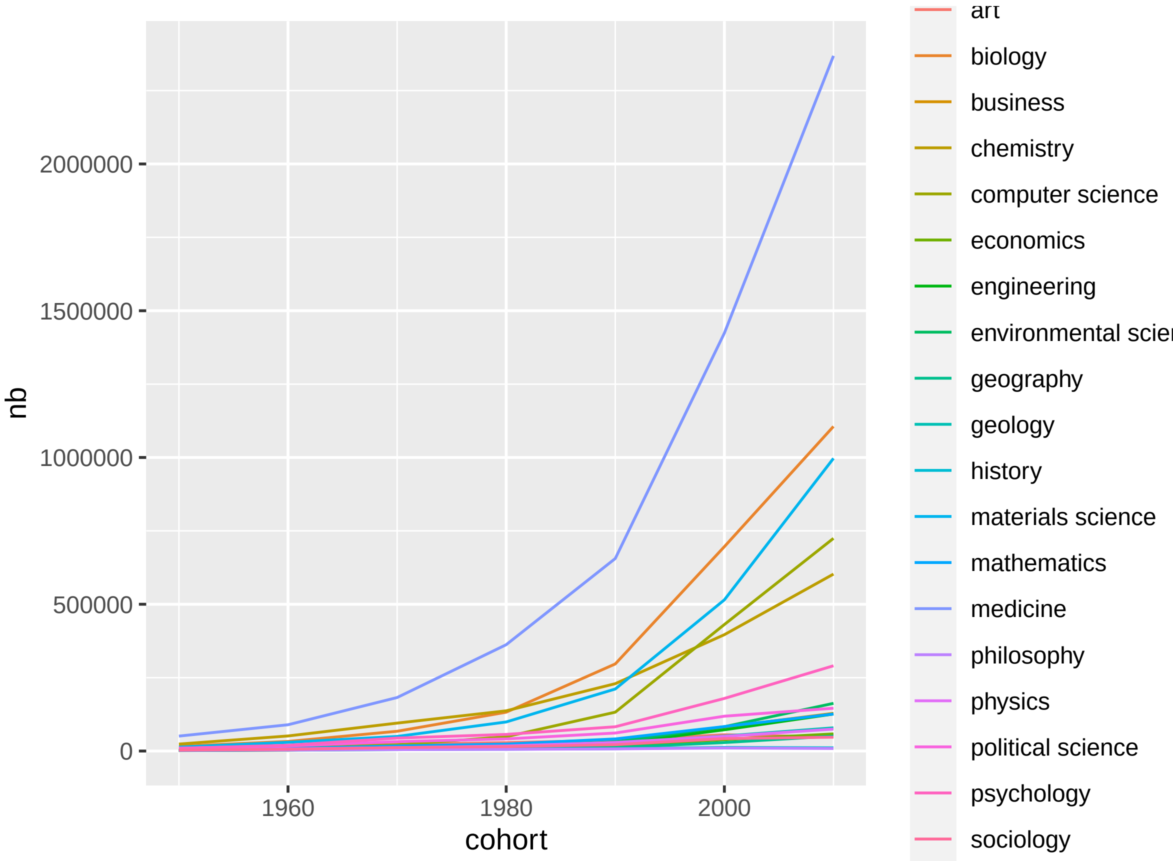
Author count and gender share by field-cohort and region-cohort

- region is assigned based on the Iso3166 Code of the author's first affiliation

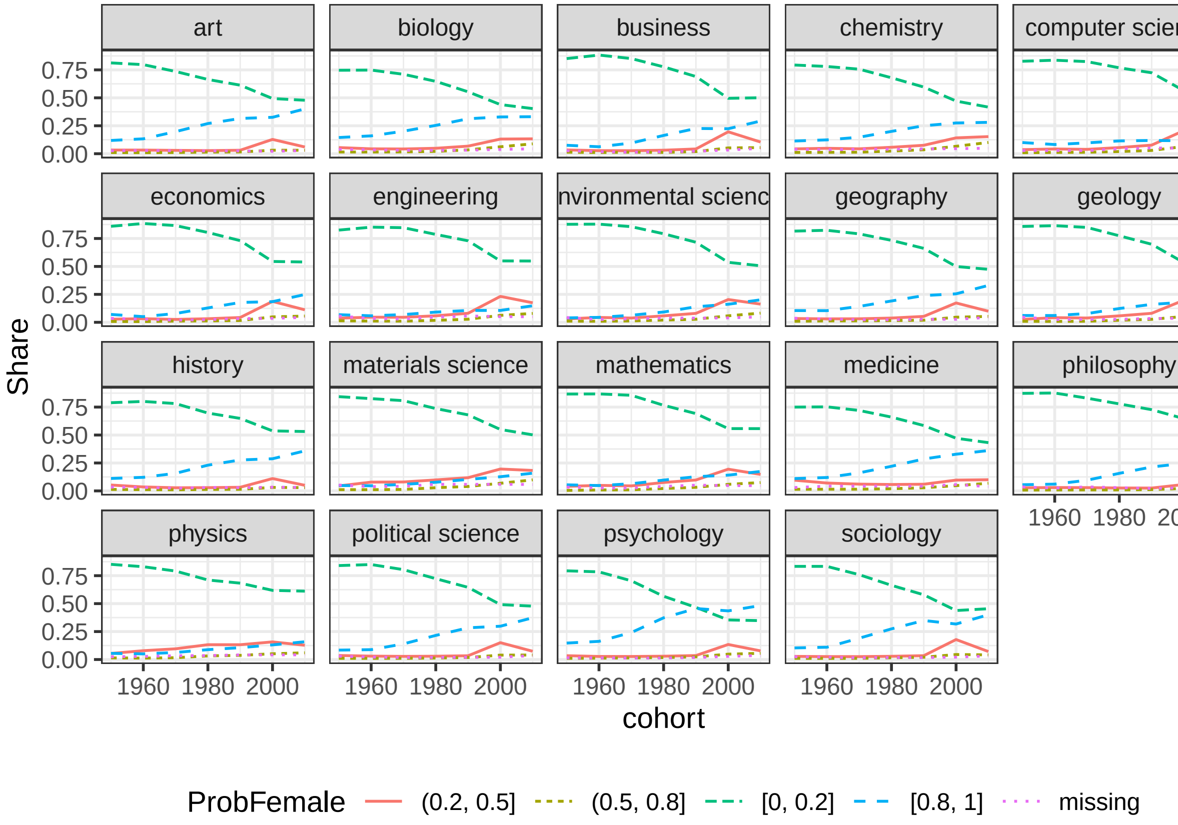
By field-cohort



Number of authors by assigned gender

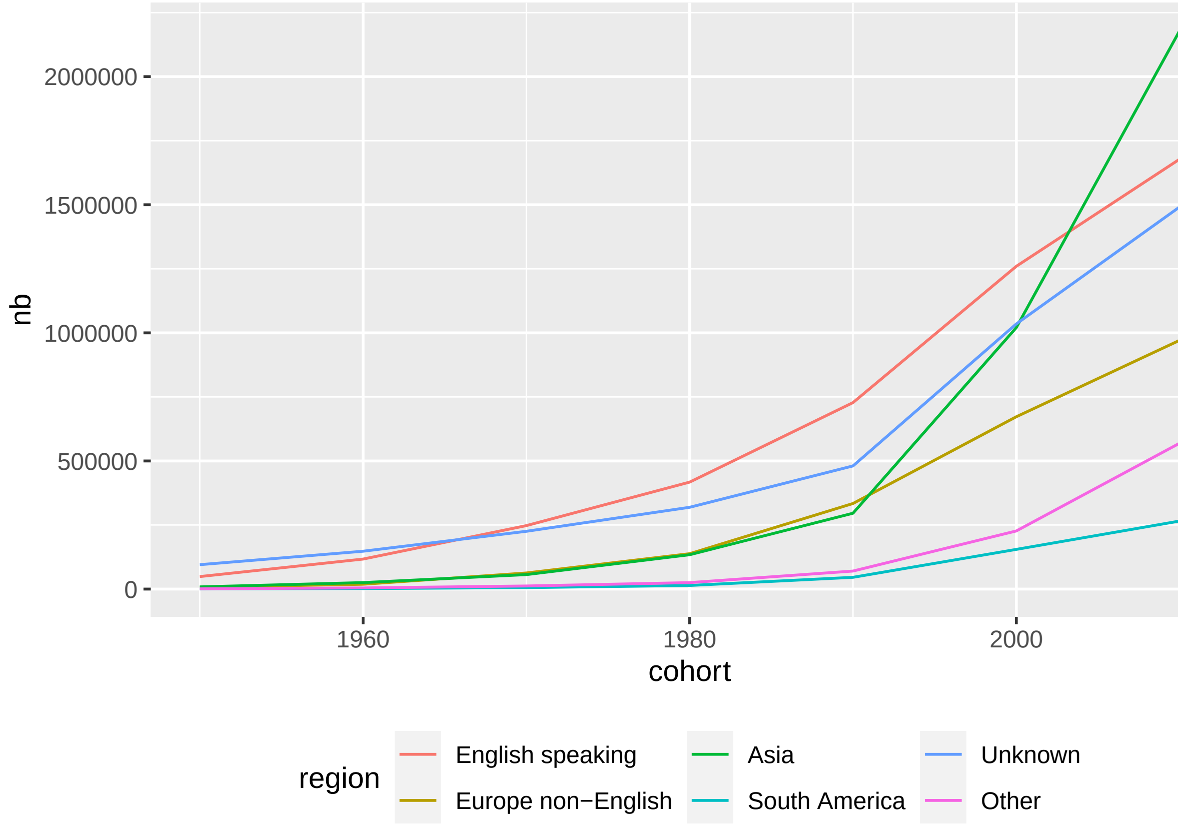


Number of authors

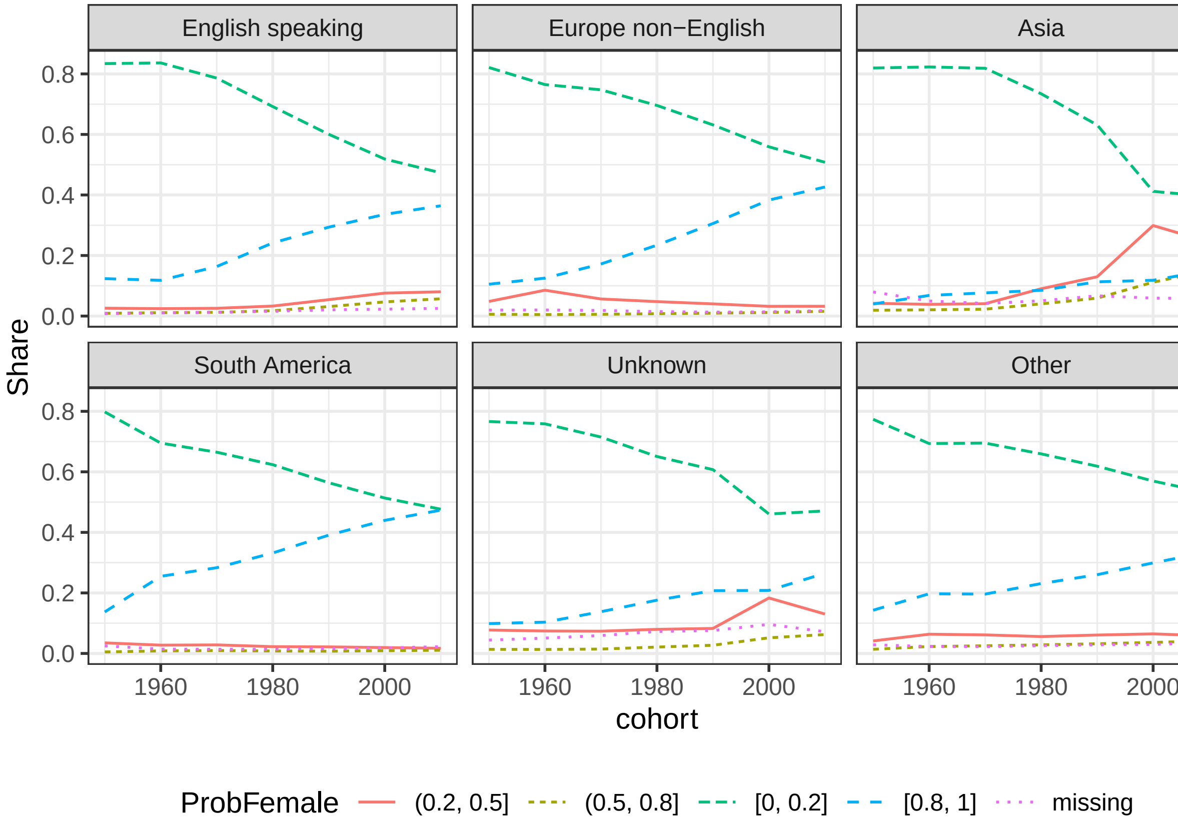


Fraction by gender

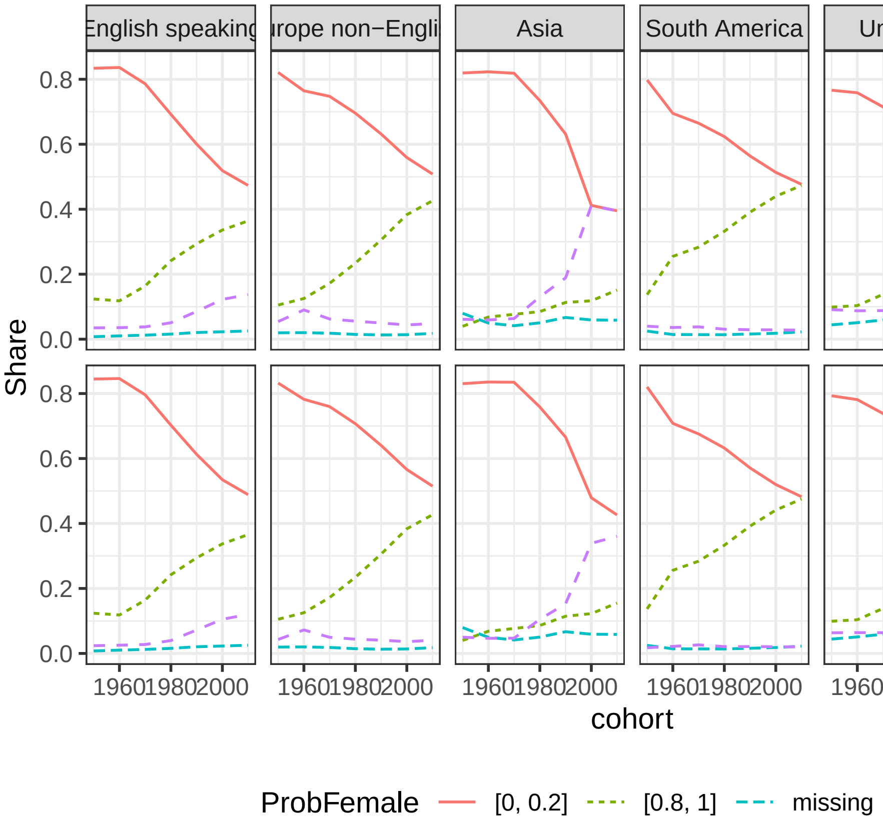
By region-cohort



Number of authors



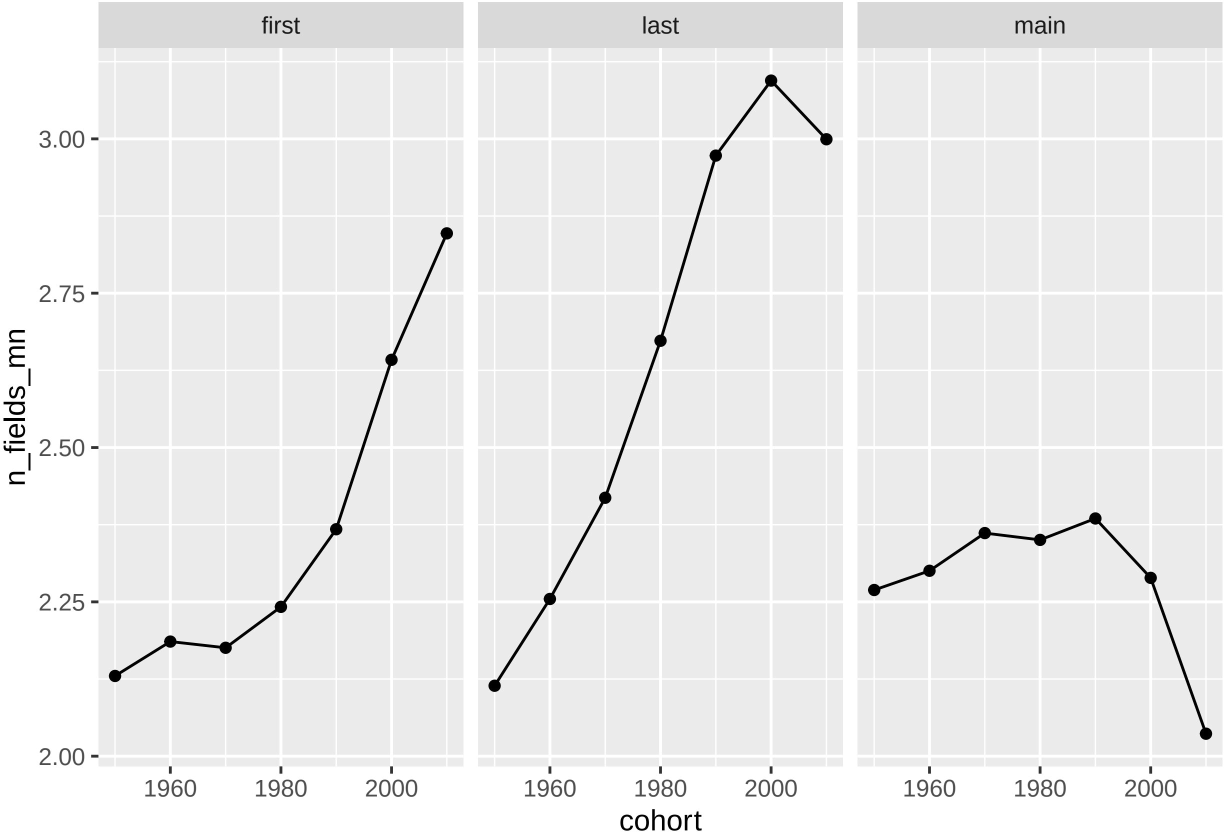
Fraction by gender



Comparing new and old gender assigment

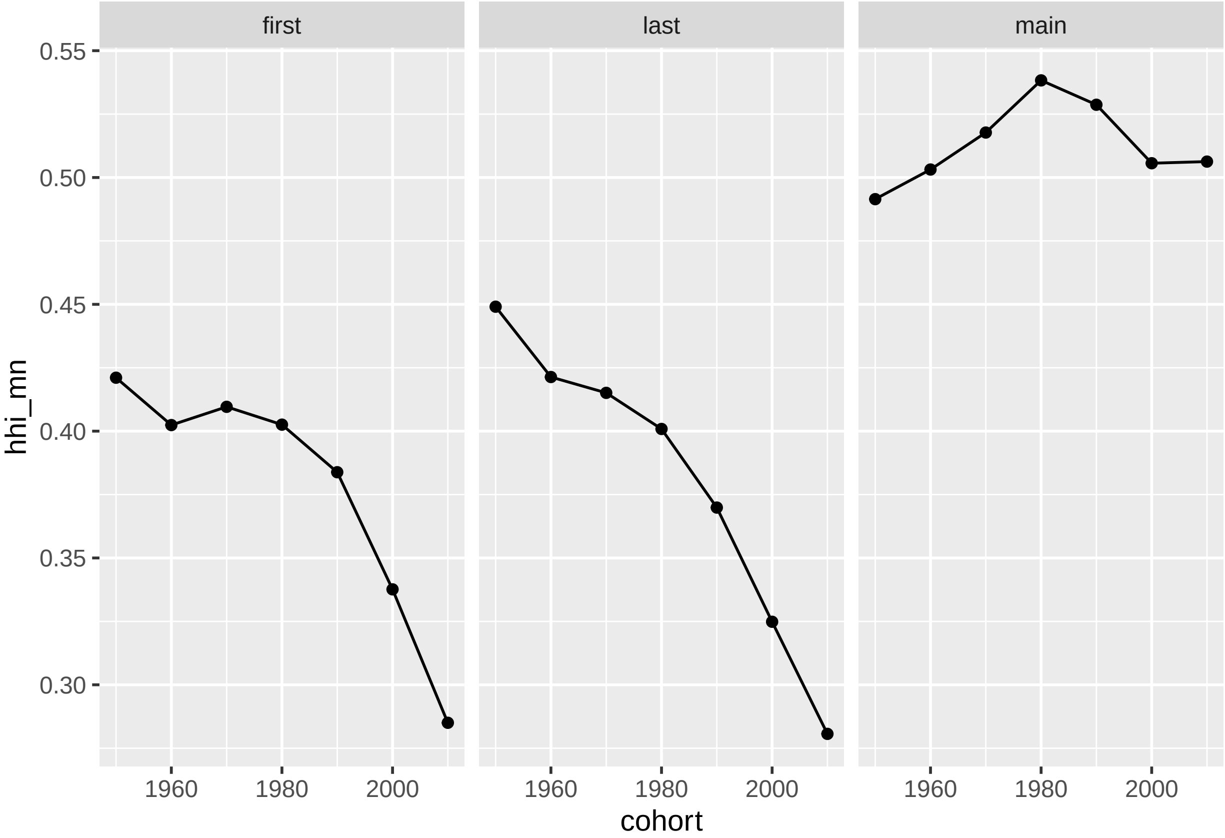
How good is the assignment of authors to fields? Aggregate statistics by cohort and FieldClass

Average count of FieldOfStudyId per AuthorId-FieldClass by cohort

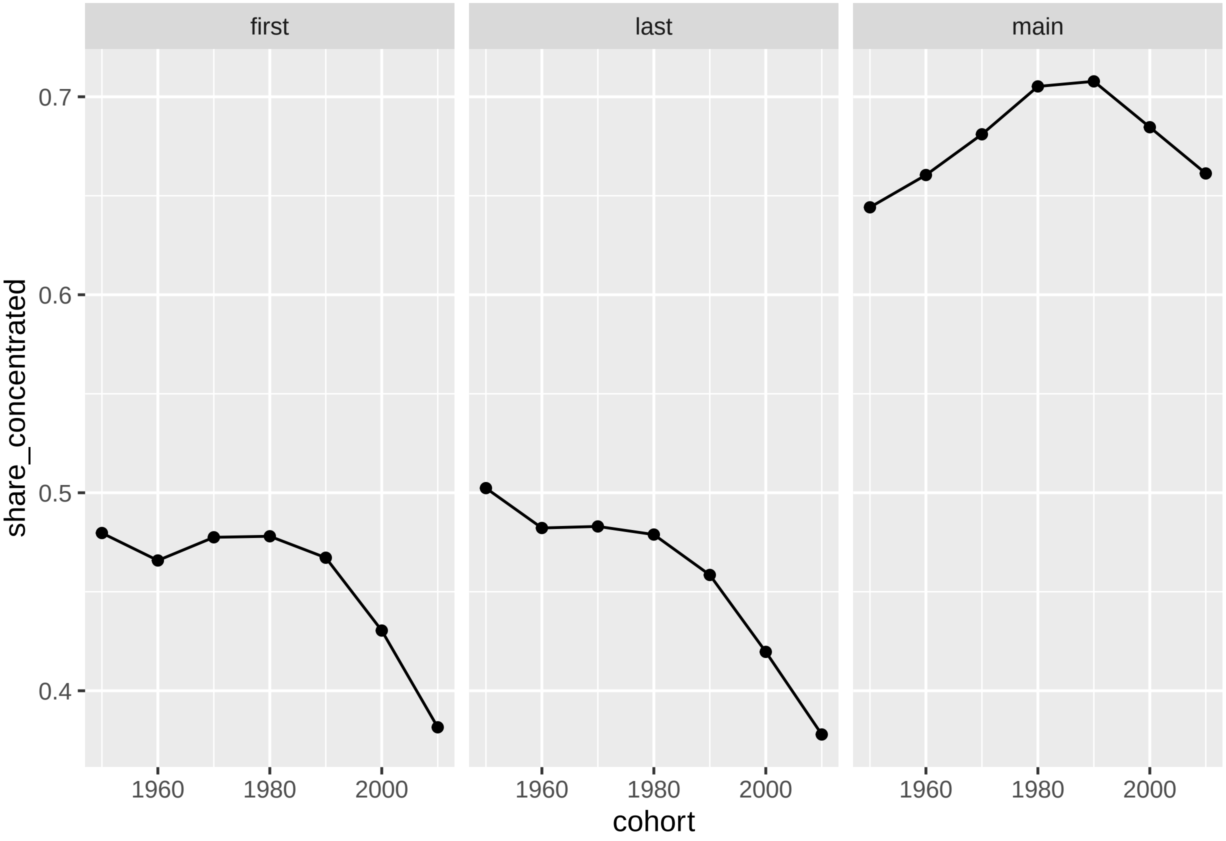


Average Herfindahl index per author

- The index measures how much an author specializes in a specific field
- The figure plots the normalized HHI


first
last
main
0.55
0.50
0.45
0.40
0.35
0.30
hhi_mn
1960
1980
2000
cohort

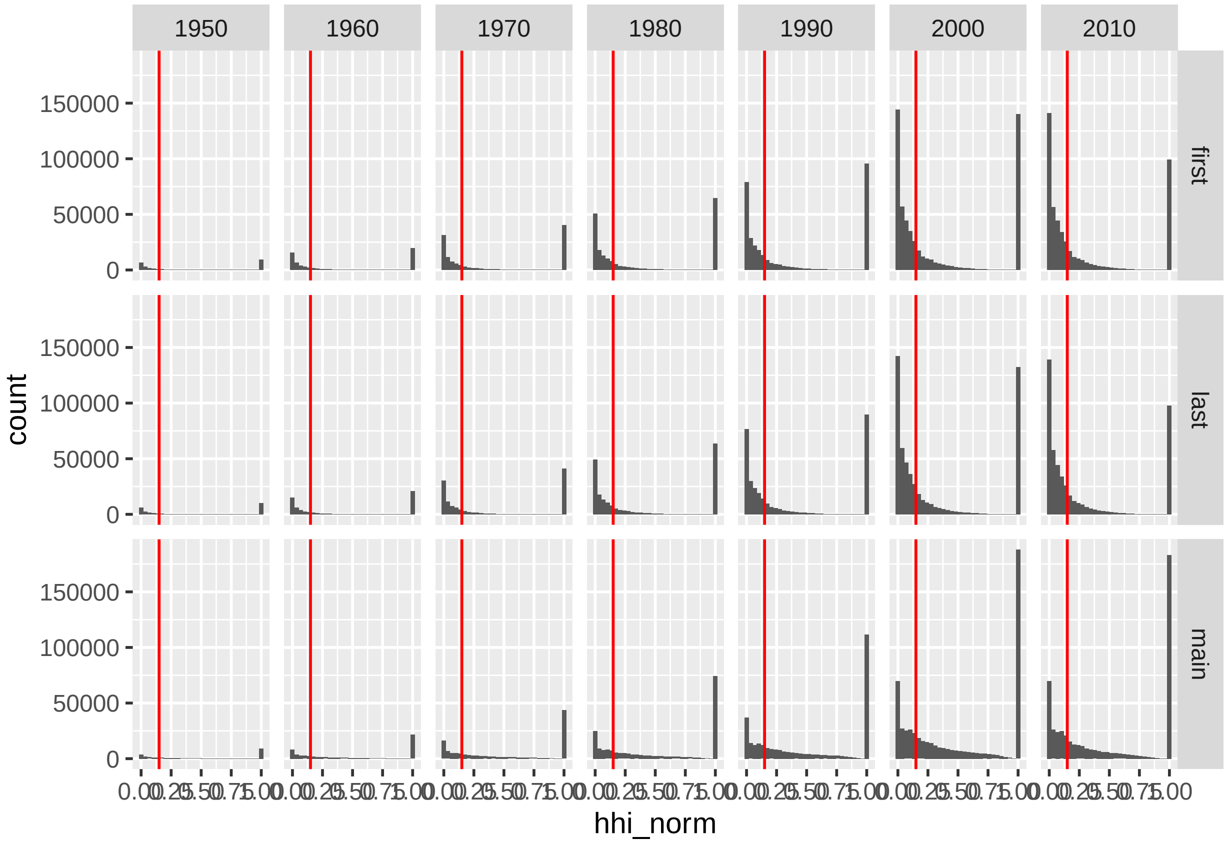
Share of authors with “moderate” or “high” concentration in a field



Normalized HHI

- 10 percent subsample of authors by FieldClass-cohort
- The red line indicates the threshold for moderate concentration

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Who are these authors with very low HHI?

- Short careers? random publications?
  - random publications are taken care of with the sample restriction imposed on `author_sample` table
- Why does the fraction of such authors grow over time?

```
##
## Call:
## lm(formula = HHIAllFields ~ log(CareerLength) + factor(cohort) +
##     log(CareerPaperCount), data = author_fields %>% filter(FieldClass ==
##     "first") %>% slice_sample(prop = 0.01))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.58174 -0.19706 -0.07363  0.23039  0.94896
##
## Coefficients:
##                         Estimate Std. Error  t value Pr(>|t|)
## (Intercept)            0.7310791  0.0051116  143.024  < 2e-16 ***
## log(CareerLength)      0.0786670  0.0008065   97.540  < 2e-16 ***
## factor(cohort)1960     0.0073765  0.0060051    1.228 0.219307
## factor(cohort)1970     0.0180529  0.0055141    3.274 0.001061 **
## factor(cohort)1980     0.0249500  0.0053148    4.694 2.68e-06 ***
## factor(cohort)1990     0.0181968  0.0051993    3.500 0.000466 ***
## factor(cohort)2000    -0.0184196  0.0051180   -3.599 0.000320 ***
## factor(cohort)2010    -0.0580959  0.0051430  -11.296  < 2e-16 ***
## log(CareerPaperCount) -0.1462412  0.0007829 -186.804  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2489 on 168975 degrees of freedom
## Multiple R-squared:  0.1936, Adjusted R-squared:  0.1935
## F-statistic:  5070 on 8 and 168975 DF,  p-value: < 2.2e-16
```

MAG institution coverage over time

Define some functions

- keep authors and their publication when they are in author_sample

```
make_year_query <- function(year) {
  query = paste0("
      SELECT a.PaperId, b.AuthorId, a.Year, a.DocType, c.AffiliationName
      FROM Papers a
      INNER JOIN (
          SELECT PaperId, AuthorId, AffiliationId
          FROM PaperAuthorAffiliations
      ) b USING(PaperId)
      LEFT JOIN (
          SELECT AffiliationId, NormalizedName AS AffiliationName
          FROM Affiliations
      ) c USING(AffiliationId)
      INNER JOIN (
        SELECT AuthorId
        FROM author_sample
      ) USING(AuthorId)
      WHERE Year = ", year, " and DocType in ('Journal', 'Book', 'BookChapter', 'Conference')")
  return(query)
}

summarise_counts <- function(d) {

  # by author
  by_author <- d %>%
    group_by(Year, AuthorId) %>%
    summarise(has_affiliation = any(!is.na(AffiliationName)),
              .groups = "drop") %>%
    group_by(Year, has_affiliation) %>%
    summarise(nb = n(),
              .groups = "drop")


  # by paper-doctype
  by_paper <- d %>%
    group_by(PaperId, Year, DocType) %>%
    summarise(has_affiliation = any(!is.na(AffiliationName)),
              .groups = "drop") %>%
    group_by(Year, DocType, has_affiliation) %>%
    summarise(nb = n(),
              .groups = "drop")
  # by author-paper-doctype
  by_author_paper <- d %>%
    group_by(PaperId, AuthorId, Year, DocType) %>%
    mutate(has_affiliation = ifelse(any(!is.na(AffiliationName)), 1, 0)) %>%
    ungroup() %>%
    filter(!duplicated(paste0(PaperId, AuthorId))) %>%
    group_by(Year, DocType) %>%
    summarise(author_paper_count = n(),
              count_with_affiliation = sum(has_affiliation),
              .groups = "drop")

  out <- list(
    by_author = by_author,
    by_paper = by_paper,
    by_author_paper = by_author_paper
  )

  return(out)
}



get_summary <- function(year) {
  cat(year, "\n--------\n")
  q <- make_year_query(year)
  data <- tbl(con, sql(q)) %>% collect()

  agg <- summarise_counts(data)
  return(agg)
}


get_summary_parallel <- function(year) {
  pcon <- DBI::dbConnect(RSQLite::SQLite(), db_file)
  q <- make_year_query(year)
  data <- tbl(pcon, sql(q)) %>% collect()
  DBI::dbDisconnect(pcon)

  agg <- summarise_counts(data)

  return(agg)
}
```

Parallel queries and summarise

- only querying subset of years should be fine for capturing trends

```
years = seq(1950, 2020, 5)

plan(multisession, workers = n_cores_to_use)

tic()
d_ls <- future_map(.x = years,
                   .f = ~get_summary_parallel(.x),
                   .options = furrr_options(chunk_size = 1, seed = TRUE)
                   )
toc()
```

```
## 508.748 sec elapsed
```

```
plan(sequential)
```

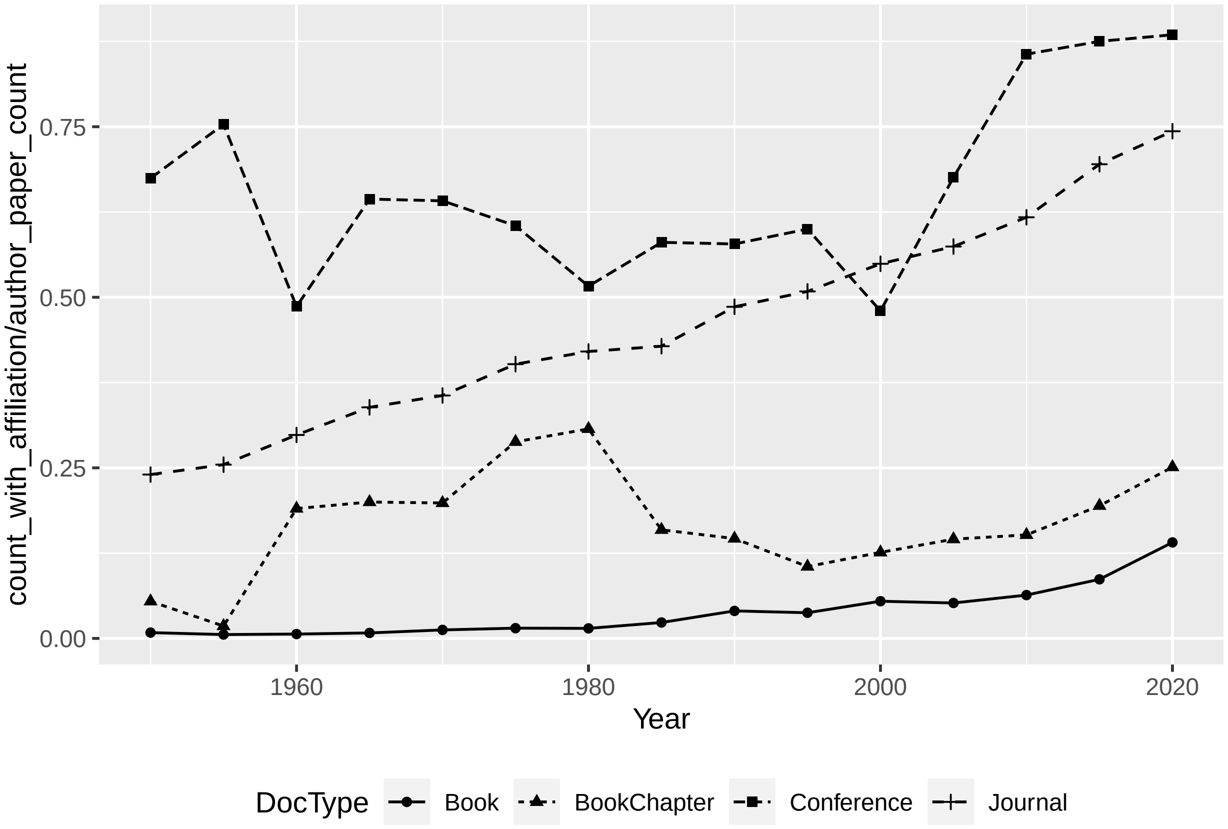
Collect data

```
df_names <- c("by_author", "by_paper", "by_author_paper")
d_collected <- map(.x = df_names,
                   .f = ~map(
                     .x = d_ls,
                     .f = ~.x[[.y]],
                     .y = .x
                   ) %>%
                     bind_rows()
                   )

names(d_collected) <- df_names
```

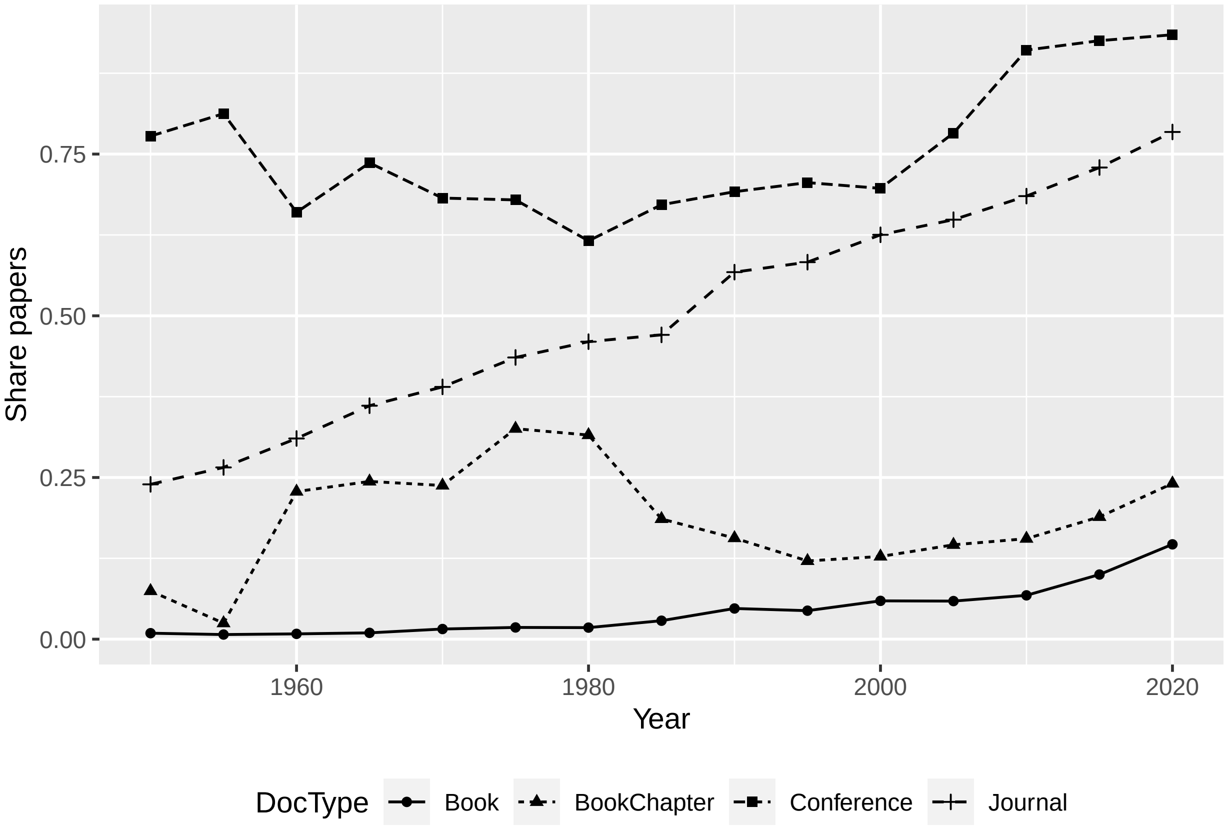
Fraction of author-paper combinations with non-missing affiliation

```
d_collected$by_author_paper %>%
  ggplot(aes(x = Year, y = count_with_affiliation/author_paper_count),
         group = DocType) +
  geom_line(aes(linetype = DocType)) +
  geom_point(aes(shape = DocType)) +
  theme(legend.position = "bottom")
```



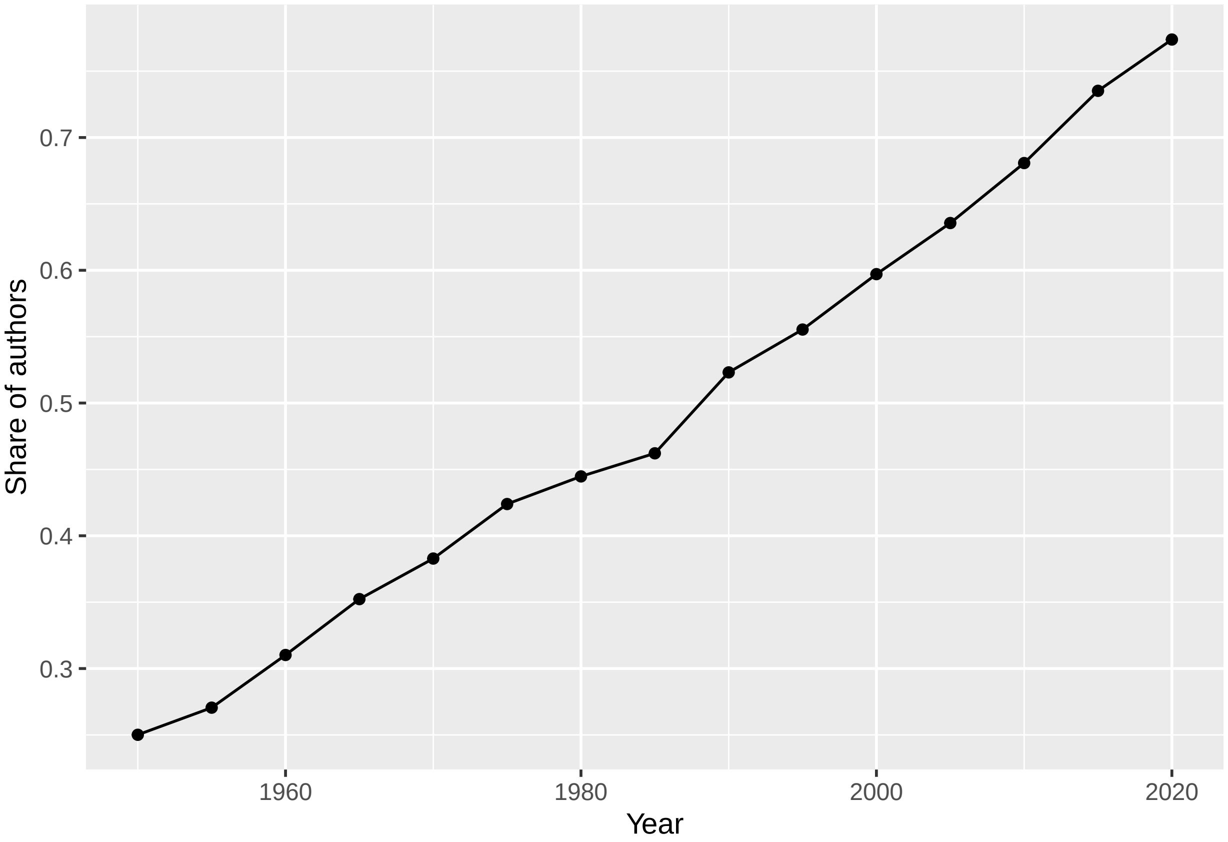
Fraction of papers with non-missing affiliation

```
d_collected$by_paper %>%
  mutate(has_affiliation = ifelse(has_affiliation, "yes", "no")) %>%
  spread(key = has_affiliation, value = nb) %>%
  ggplot(aes(x = Year, y = yes/(yes + no)),
         group = DocType) +
  geom_line(aes(linetype = DocType)) +
  geom_point(aes(shape = DocType)) +
  theme(legend.position = "bottom") +
  labs(y = paste0("Share papers"))
```



Fraction of authors with non-missing affiliation

```
d_collected$by_author %>%
  mutate(has_affiliation = ifelse(has_affiliation, "yes", "no")) %>%
  spread(key = has_affiliation, value = nb) %>%
  ggplot(aes(x = Year, y = yes/(yes + no))) +
  geom_line() +
  geom_point() +
  theme(legend.position = "bottom") +
  labs(y = "Share of authors")
```



What do we learn?

- At first sight, the coverage of affiliations seems low
- But we would like to know the stats for a more selected sample: authors in US
    - Also remember that MAG covers more documents than other sources
- How can we get closer to what we want to measure?
    - perhaps we could measure the fraction of authors in our graduate-mag linked sample that have an affiliation over time?
    - since we did not use the affiliation as a feature for linking, this could work