zip2fips checks

2024-06-13

Quarter inclusion?

We run a short sensitivity analysis to examine changes in the final crosswalk when input crosswalks from all quarters are used.

```
# If we run just Q4 analysis, what changes? -------------------------------
library(data.table)
library(dplyr)
library(ggplot2)

setwd("~/desktop/nsaph/data_team/zip-fips-crosswalk")
# load full data frame
xwalk_full <- data.table::fread("./data/intermediate/zip2fips_xwalk_clean_allquarters.csv",
                                colClasses = c(zip="character",
                                               fips="character"))

# perform basic matching with just Q4 data
fips_matches_q4 <- xwalk_full %>%
  filter(quarter == 4) %>%
  group_by(zip, year) %>%
  slice_max(tot_ratio) %>%
  ungroup() %>%
  group_by(zip, fips) %>%
  summarize(min_year = min(year),
            max_year = max(year),
            total_matches = n(),
            tot_ratio_min = min(tot_ratio),
            tot_ratio_max = max(tot_ratio),
            tot_ratio_mean = mean(tot_ratio))

# do matching with full dataset
fips_matches_all <- xwalk_full %>%
  group_by(zip, year) %>%
  slice_max(tot_ratio) %>%
  ungroup() %>%
  group_by(zip, fips) %>%
  summarize(min_year = min(year),
            max_year = max(year),
            total_matches = n(),
            tot_ratio_min = min(tot_ratio),
            tot_ratio_max = max(tot_ratio),
            tot_ratio_mean = mean(tot_ratio))

fips_matches_q4$match_str <- paste0(fips_matches_q4$zip, ",", fips_matches_q4$fips)
```

```r
fips_matches_all$match_str <- paste0(fips_matches_all$zip, ",", fips_matches_all$fips)

# what fraction of zip/fips matches are NOT in the q4 dataset?
print(sum(!unique(fips_matches_all$match_str) %in%
      unique(fips_matches_q4$match_str))/length(unique(fips_matches_all$match_str)))
```

```
## [1] 0.0009296636
```

```r
# raw number
print(sum(!unique(fips_matches_all$match_str) %in%
      unique(fips_matches_q4$match_str)))
```

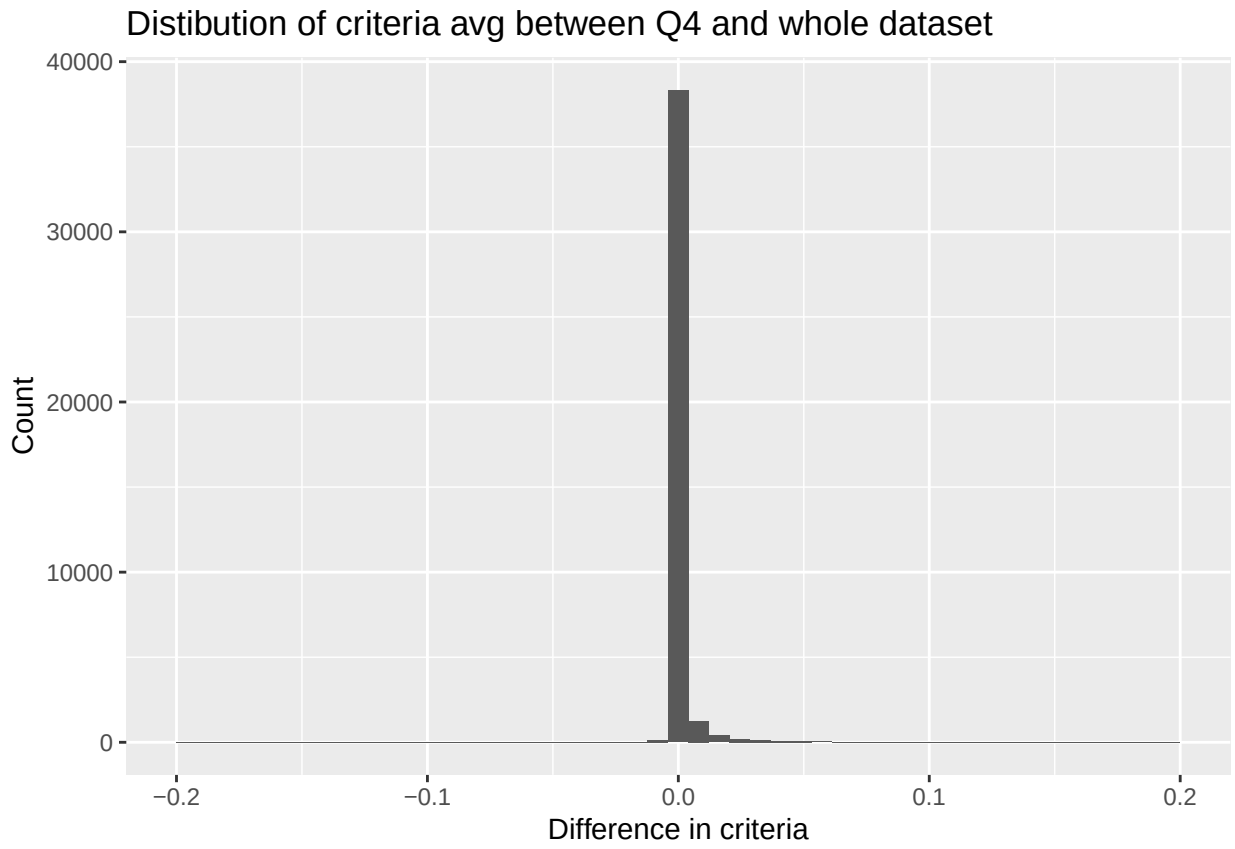
```
## [1] 38
```

The above code chunk calculates the total number of total zip/fips matches that are in the quarterly crosswalks but not in the yearly-compiled crosswalks. There are just 0.09% missing! Only 38 out of nearly 40,000 unique zip codes.

We now run an analysis to tell if there is a noticeable change in criteria averages if we use Q4 vs the whole dataset. Note that these matches are not exactly the same as would be produced by the pipeline, but they are close enough for a quick comparison.

```r
# now comparing criteria means between the two groups
fips_matches_all_mg <- fips_matches_all %>%
  merge(fips_matches_q4 %>% select(zip, fips, tot_ratio_mean),
        by=c("zip", "fips"),
        suffixes=c(".all", ".q4")) %>%
  mutate(criteria_diff = tot_ratio_mean.all - tot_ratio_mean.q4)

ggplot(fips_matches_all_mg, aes(x=criteria_diff)) +
  geom_histogram(bins=50) +
  xlab("Difference in criteria") +
  ylab("Count") +
  xlim(c(-0.2, 0.2)) +
  ggtitle("Distibution of criteria avg between Q4 and whole dataset")
```




```
# summary statistics
print(paste0("Mean difference: ", mean(fips_matches_all_mg$criteria_diff)))
```

```
## [1] "Mean difference: 0.00137617393789681"
```

```
print(paste0("SD difference: ", sd(fips_matches_all_mg$criteria_diff)))
```

```
## [1] "SD difference: 0.00844233996447902"
```

```
print(quantile(fips_matches_all_mg$criteria_diff, c(0.025, 0.5, 0.975)) %>% round(5))
```

```
##     2.5%      50%    97.5%
## -0.00008  0.00000  0.01355
```