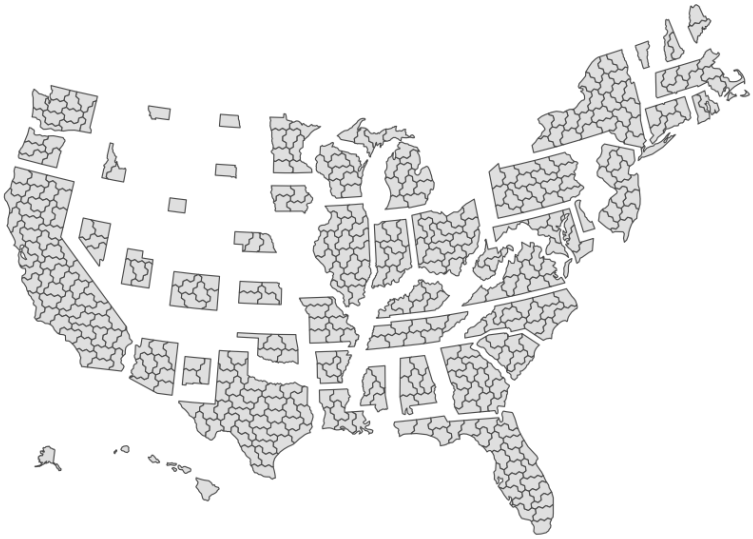
Vote Choice and Population Size by Geography, Demographics, and Turnout

AUTHOR
Shiro Kuriwaki, Stephen Ansolabehere, Angelo Dagonel, and Soichiro Yamauchi



Shapefile: Daily Kos

Dataverse Version 3 (May 2023)

Contact: Shiro Kuriwaki (https://www.shirokuriwaki.com; ORCID: 0000-0002-5687-2647)

Purpose

This Dataverse contains estimates of vote choice and electoral size at the congressional district level. We estimate a high-dimensional joint post-stratification table for weighting, and use a hierarchical regression to effectively use sparse survey data. We report the open-source software we used to estimate this model.

For 2016 and 2020 estimates, please cite:

> Kuriwaki, Shiro, Stephen Ansolabehere, Angelo Dagonel, and Soichiro Yamauchi. 2023. "The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level." *American Political Science Review*.

This Dataverse (https://doi.org/10.7910/DVN/MAZNJ6) contains a copy of the post-stratification tables and estimates from the article, but also will add estimates years other than the ones in the article (2016 and 2020). The APSR Replication Dataverse (https://doi.org/10.7910/DVN/VX5N1V) includes the code to analyze data like the one here and make the figures in the *APSR* articles.

These estimates were supported by NSF Grant 1926424, and developed with input and collaboration from Andrew Gelman, Lauren Kennedy, and Mitzi Morris.

The data are organized in three types of folders.

Poststratification counts

The counts describe the size of each demographic cell in the electorate (as well as the non-turnout voting age population). They can be summed up in various ways, e.g. at the district level, race-district level, or education-district level. They serve as the weighting target for survey estimates that contain vote choice probabilities.

Dataset `poststrat.csv`

Estimates of the population size of demographic (post-stratification) cells. Each row defines an year, state/congressional district, age (5-way), sex, race (4-way), education (4-way), and turnout (0 or 1). Here is a sample of the table:

```
# A tibble: 278,400 × 11
    year st    cd    race  age      female educ  turnout count_calib denominator
   <dbl> <chr> <chr> <chr> <chr>     <dbl> <chr>   <dbl>       <dbl>       <dbl>
 1  2016 CT    CT-01 White 18 to 2…      0 HS o…       1       2369.       6317.
 2  2016 CT    CT-01 White 18 to 2…      0 Some…       1       3501.       6560.
 3  2016 CT    CT-01 White 18 to 2…      0 4-Ye…       1       2197.       4060.
 4  2016 CT    CT-01 White 18 to 2…      0 Post…       1        312.        549.
 5  2016 CT    CT-01 White 18 to 2…      1 HS o…       1       2971.       7868.
 6  2016 CT    CT-01 White 18 to 2…      1 Some…       1       2788.       5198.
 7  2016 CT    CT-01 White 18 to 2…      1 4-Ye…       1       1967.       3617.
 8  2016 CT    CT-01 White 18 to 2…      1 Post…       1        139.        244.
 9  2016 CT    CT-01 White 25 to 3…      0 HS o…       1       1563.       4143.
10  2016 CT    CT-01 White 25 to 3…      0 Some…       1       2976.       5552.
# ℹ 278,390 more rows
# ℹ 1 more variable: count_twway <dbl>
```

These estimates come from a synthetic method for estimating the joint distribution of demographics. See our APSR article and Appendix A.6 for details. These estimates were built

with data loaded in the packages ccesMRPprep (https://github.com/kuriwaki/ccesMRPprep) and synthjoint (https://github.com/kuriwaki/synthjoint).

Main variables:

- The demographic, geographic, and turnout variables indicate the partition of the cell.
- `count_calib`: counts calibrated one-way to district turnout.
- `count_twway`: counts calibrated two-way to district turnout and national race-level turnout estimated from the CPS. We use **count_twway** in the APSR article but doing this relies on the assumption that the national race-level turnout computed from the CPS is accurate.

Survey-based estimates (MRP)

These estimates provide vote choice (i.e., Republican vs. Democrat) estimates at the district and race level.

Dataset `by-cd-race_MRP-ests.csv`

Estimates of two-party Presidential voteshare by CD by race by year by estimator. Each row is a district-race combination. It includes three sets of estimates of the Trump vote in each of those combinations.

```
# A tibble: 3,480 × 18
    year cd    race          N p_mrp_nofix p_mrp_ggfix p_mrp_twway p_mrp_twway_se
   <dbl> <chr> <chr>     <dbl>       <dbl>       <dbl>       <dbl>          <dbl>
 1  2016 AK-01 Black      6563       0.145       0.173       0.161         0.0736
 2  2016 AK-01 Hispan…    9565       0.451       0.503       0.446         0.104
 3  2016 AK-01 Other     59169       0.459       0.512       0.507         0.0518
 4  2016 AK-01 White    239585       0.564       0.616       0.625         0.0129
 5  2016 AL-01 Black     67979       0.0446      0.0773      0.0723        0.0312
 6  2016 AL-01 Hispan…    5617       0.550       0.683       0.642         0.0939
 7  2016 AL-01 Other      7727       0.655       0.772       0.777         0.0513
 8  2016 AL-01 White    222617       0.712       0.821       0.829         0.00957
 9  2016 AL-02 Black     73747       0.0467      0.0997      0.0942        0.0345
10  2016 AL-02 Hispan…    4105       0.525       0.700       0.671         0.0948
# ℹ 3,470 more rows
# ℹ 10 more variables: p_mrp_twway_050 <dbl>, p_mrp_twway_100 <dbl>,
#   p_mrp_twway_900 <dbl>, p_mrp_twway_950 <dbl>, p_raw <dbl>, p_wt <dbl>,
#   n_raw <dbl>, se_raw <dbl>, n_wt <dbl>, se_wt <dbl>
```

Main variables

- `year` is the election year, `race` is a four-way categorization of race, and `cd` is the congressional district in that election. Currently, all the estimates `p_` are for the

Republican two-party vote for President.

- `p_mrp_nofix`, `p_mrp_ggfix`, `p_mrp_twway` is the estimated two-party Republican voteshare in each group. The suffixes represent different estimators. `nofix` is the estimates without any calibration, `_ggfix` is the estimates with a one-way fix a la Ghitza Gelman, and `_twway` is the two-way calibration that we propose. **`_twway` is the preferred specification we use in the APSR article.**
- `p_mrp_twway_050` indicates the 5% percentile in the posterior distribution of the two-way calibrated estimate. Similarly, `p_mrp_twway_900` is the 90th percentile, and so on.
- `p_raw` is the direct estimator (raw average) of the Republican vote in the sample.
- `p_wt` is the direct estimator of the Republican vote with the survey weights that come with the CCES/CES survey. As we discuss in the article, these weights do not weight to the district level but only to the national/state level.
- `n_raw`, `n_wt`, `se_raw`, and `se_wt` are the sample sizes (`n_`) and standard errors (`se_`) that correspond to the raw, unweighted (`_raw`) and simple weighted estimators (`_wt`).
- `pct_trump` is at the district-level (rather than district x race level) covariates. They measure the percent of the two-party vote won by Trump in that year and district. Computed by Daily Kos.
- `N` is the estimated population size of the district x race combination that turned out. Use this to aggregate estimates up to higher levels of geography.

To generate estimates at different levels of aggregation, one can use the hierarchical regression models in the APSR Dataverse and predict them onto the post-stratification cell.

Dataset `by-state-race_MRP-ests.csv`

This has the same format as `by-cd-race_MRP-ests.csv` but for each state. The reason a different dataset is provided here is that the standard error of the state-level estimate must be recomputed from the state x iteration level.

```
# A tibble: 400 × 20
    year st    race         N p_mrp_nofix p_mrp_ggfix p_mrp_twway p_mrp_nofix_se
   <dbl> <chr> <chr>    <dbl>       <dbl>       <dbl>       <dbl>          <dbl>
 1  2016 AK    Black   6.56e3      0.145       0.173       0.158          0.0698
 2  2016 AK    Hispan… 9.56e3      0.451       0.503       0.428          0.108
 3  2016 AK    Other   5.92e4      0.459       0.512       0.491          0.0676
 4  2016 AK    White   2.40e5      0.564       0.616       0.630          0.0303
 5  2016 AL    Black   4.81e5      0.0416      0.0722      0.0663         0.0178
 6  2016 AL    Hispan… 4.62e4      0.506       0.639       0.580          0.0784
 7  2016 AL    Other   4.91e4      0.645       0.762       0.756          0.0538
 8  2016 AL    White   1.53e6      0.710       0.821       0.829          0.0200
 9  2016 AR    Black   1.33e5      0.0605      0.0852      0.0723         0.0230
10  2016 AR    Hispan… 3.62e4      0.443       0.507       0.440          0.0805
```

```
# ℹ 390 more rows
# ℹ 12 more variables: p_mrp_ggfix_se <dbl>, p_mrp_twway_se <dbl>,
#   p_mrp_twway_050 <dbl>, p_mrp_twway_100 <dbl>, p_mrp_twway_900 <dbl>,
#   p_mrp_twway_950 <dbl>, p_raw <dbl>, p_wt <dbl>, n_raw <dbl>, se_raw <dbl>,
#   n_wt <dbl>, se_wt <dbl>
```

Ecological inference estimates (EI)

These are estimates of vote choice by race using only aggregate electoral and Census data. They are roughly in the same format as the MRP estimates.

Dataset `by-cd-race_EI-ests.csv`

Estimates of two-party Presidential voteshare by CD by race by year.

```
# A tibble: 3,480 × 9
   cand       race    cd    ei_mean   ei_sd ei_025 ei_975 ei_nsamples nprecincts
   <chr>      <chr>   <chr>   <dbl>   <dbl>  <dbl>  <dbl>       <dbl>      <dbl>
 1 pres2_20_R White   AK-01  0.691  0.00881 0.673  0.705         5000        447
 2 pres2_20_R Black   AK-01  0.250  0.0934  0.118  0.492         5000        447
 3 pres2_20_R Hispan… AK-01  0.364  0.0500  0.243  0.472         5000        447
 4 pres2_20_R Other   AK-01  0.305  0.0141  0.278  0.340         5000        447
 5 pres2_20_D White   AK-01  0.309  0.00881 0.295  0.327         5000        447
 6 pres2_20_D Black   AK-01  0.750  0.0934  0.508  0.882         5000        447
 7 pres2_20_D Hispan… AK-01  0.636  0.0500  0.528  0.757         5000        447
 8 pres2_20_D Other   AK-01  0.695  0.0141  0.660  0.722         5000        447
 9 pres2_20_R White   AL-01  0.903  0.00554 0.892  0.915         5000        214
10 pres2_20_R Black   AL-01  0.0508 0.0105  0.0343 0.0820        5000        214
# ℹ 3,470 more rows
```

Main estimates:

- `cand`: The candidate of interest. Either the Republican or Democrat in the 2020 Presidential race
- `p_ei_est`: Republican/Democratic two-party vote estimate by EI, among that cd and racial group.
- `ei_sd`: The standard deviation across EI samples.
- `ei_025` and `ei_975` represent the 95% credible interval.
- For other variables, see the APSR article replication package.

Dataset `alarm-vest_precinct_for-EI.csv.gz`

The underlying precinct-level data used to compute the EI estimates. This data combines election data from VEST (https://dataverse.harvard.edu/dataverse/electionscience) and

decennial Census race and population data formatted by ALARM (https://dataverse.harvard.edu/dataverse/alarm; https://github.com/alarm-redist/census-2020).

```
# A tibble: 167,728 × 35
   cd    GEOID20     state county         cd_2010   pop   vap vap_hisp_frac
   <chr> <chr>       <chr> <chr>          <chr>   <dbl> <dbl>         <dbl>
 1 AK-01 0201337-700 AK    Aleutians Eas… 1        1589  1567        0.357
 2 AK-01 0201337-712 AK    Aleutians Eas… 1         450   420        0.1
 3 AK-01 0201337-720 AK    Aleutians Eas… 1         759   655        0.0641
 4 AK-01 0201337-744 AK    Aleutians Eas… 1         622   478        0.0293
 5 AK-01 0201637-704 AK    Aleutians Wes… 1         498   457        0.116
 6 AK-01 0201637-706 AK    Aleutians Wes… 1        4254  3768        0.140
 7 AK-01 0201637-750 AK    Aleutians Wes… 1          67    54        0.0370
 8 AK-01 0201637-752 AK    Aleutians Wes… 1         413   300        0.00667
 9 AK-01 0202012-225 AK    Anchorage Mun… 1        2353  1816        0.0523
10 AK-01 0202012-230 AK    Anchorage Mun… 1        3297  2395        0.0409
# ℹ 167,718 more rows
# ℹ 27 more variables: vap_white_frac <dbl>, vap_black_frac <dbl>,
#   vap_other_frac <dbl>, pres2_16_R <dbl>, pres2_16_D <dbl>,
#   pres2_20_R <dbl>, pres2_20_D <dbl>, pop_hisp <dbl>, pop_white <dbl>,
#   pop_black <dbl>, pop_aian <dbl>, pop_asian <dbl>, pop_nhpi <dbl>,
#   pop_other <dbl>, pop_two <dbl>, vap_hisp <dbl>, vap_white <dbl>,
#   vap_black <dbl>, vap_aian <dbl>, vap_asian <dbl>, vap_nhpi <dbl>, …
```

See the documentation in the APSR Replication Dataverse (https://doi.org/10.7910/DVN/VX5N1V), in particular `14_EI-435-districts.R` (https://dataverse.harvard.edu/file.xhtml?fileId=6909048) to apply EI to this dataset to replicate the result in the APSR paper.