The Geography of Racially Polarized Voting

This Dataverse Replication is for Kuriwaki et al., “The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level”.

Please cite this dataset using the article:

> 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*

Corresponding Author: Shiro Kuriwaki (ORCID 0000-0002-5687-2647).

Note: The article analyzes estimates for 2016 and 2020. For future or past estimates, you may find more estimates at Kuriwaki et al., “Modeled vote choice of geographic and racial subgroups with survey and ecological data” also at the Dataverse https://doi.org/10.7910/DVN/MAZNJ6.

Navigation Instructions

1. Download the replication contents (https://doi.org/10.7910/DVN/VX5N1V) in its entirety as a zipfile (in “original” format rather than archival format) to retain the folder structure.

2. Download the necessary R packages, which are given in the next section.

3. To make figures and tables out of computed estimates or cleaned datasets, see the files and data in `analyze`.

4. To generate the computed estimates from raw data, see the files and data in `build`.

Computing Environment

- These should be computed in R 4.1 or above.
- Packages: The beginning of each file reports the packages necessary to run the scripts. Most of the scripts rely on `tidyverse` and `haven`.
- The latest versions of the main packages used were:
    - `cmdstan` : v2.31.0 (See `cmdstanr::cmdstan_version()` in R)
    - `brms` : v2.17.5
    - The underlying software for all of this is Stan. We use rstan version v2.21.7.

  The cmdstan software was installed from https://mc-stan.org/cmdstanr/ instead of CRAN.
- The packages `ccesMRPprep`, `ccesMRPrun`, `ccesMRPviz`, and `synthjoint` are custom packages used throughout the replication to facilitate common procedures and datasets used in the MRP workflow. The current versions of those can be downloaded from a Public github repository below. For posterity, a archival version of the packages at the time of submission have been shared in the `src` folder.

```
remotes::install_github("kuriwaki/ccesMRPprep")
remotes::install_github("kuriwaki/ccesMRPrun")
remotes::install_github("kuriwaki/ccesMRPviz")
remotes::install_github("kuriwaki/synthjoint")
```

Data and scripts in `analyze`

The `analyze` half of the project constructs the figures and tables of our paper. The scripts should be run in order from the Rstudio project in the dataverse. Be sure to install the dataverse package in its entirety to preserve the file structure.

- R code with the numbers `01` to `10` reproduce figures and tables in the main text, as well as some other appendix tables. Each script is named with the name of the Figure or Table it reproduces.
- R code with the numbers `11` to `17` create data related to the appendix, for example EI estimates for all 435 districts.

The following explains the datasets and their variables, along with the relevant scripts. Many of the scripts also have their independent comments.

- `mrp_ests_by_cd_race.csv` : This is the main MRP output. There are 3480 rows, one for each year (2) x district (435) x race.
    - `p_mrp_nofix` , `p_mrp_ggfix` , `p_mrp_twway` is the estimated two-party Republican 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 our paper.**
    - `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 Trump vote in the sample.
    - `p_wt` is the direct estimator with the survey weights that come with the national survey. As we discuss in the paper, these weights do not weight to the district level.
    - `n_raw` , `n_wt` , `se_raw` , and `se_wt` are sample sizes `n` and standard errors `se_` that correspond to the raw and weighted estimators.
    - `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. See `cd_presvote_2016_2020.csv`
    - `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.
    - `mrp_ests_by_state_race.csv` is the same format of the dataset 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. This file is created in the `build` project, `07b_summ_sims_export.R`. In `04_tabs_table_01_appendix_tables.R`, we use this dataset to report the state-level standard errors.
    - `mrp_ests_by_cd_race_asian.csv` estimates the voteshares in each district but including Asian Americans as a fifth racial category. For the reasons explained in Appendix B of the paper, this is estimated with data from four states.

- `cces_by_agg_level.dta`: A dataset similar to `mrp_ests_by_cd_race.csv` but an additional few rows that provide estimates for different levels of geography. Script `02_create_aggregates_by_level.R` creates this dataset. **This dataset is the most commonly used file for creating figures and tables in the subsequent scripts.** The `level` variable indicates the level of the geography in question and constitutes of `"nation"` (the entire US), `"state"` (state-level), `"region"` (Census region level), `"division"` (Census division level), and `"cd"`. For example, for rows with `level == "state"`, the MRP estimate refers to the two-way Republican voteshare for the state specified in `state` for race defined in `race`.

    - `N_geography`: The population size in that geography (rather than the geography x race combination). `frac_race_in_geo` represents the fraction of the race in the geography.
    - This file is a Stata dta file. In R, use `haven::read_dta` to read the file with proper labels. Use `haven::as_factor` to turn the labelled variables into factors.

- `mrp_full_posterior.csv.gz`: The full posterior distribution that is used to generate the point estimates and standard errors in `mrp_ests_by_cd_race.csv`. These posterior samples are necessary to compute uncertainty intervals in, for example, Figure 3b and the ANOVA results. The variable `iter` indicates 1 of 2000 iterations of

the posterior distribution, and `p_mrp_*` indicate the MRP estimates with and without calibration in that iteration. The rest of the variables are the same as the file `mrp_ests_by_cd_race.csv`. There are 3480*2000 = 6,960,000 observations. Apply `read_csv` on the zip file itself for the fastest performance in R.

- `ei_by_cd_race_comparsion.dta`: A comparison of MRP and EI for 2020 for all 435 congressional districts. These are used in scripts `16_compare_racialgap_EI_MRP.R` and `17_compare_anova_EI_MRP.R` to make comparisons.
    - It is constructed in `15_compare_EI_MRP.R` from the EI output `by_cd_race_EI_ests.csv` and the MRP estimates
    - `by_cd_race_EI_ests.csv` is constructed in `14_EI_435_districts.R` from the precinct-level dataset `vtd_cd10_2020_pres_16_20.csv.gz`. We loop through each district and perform ecological inference, saving intermediate estimates.
- The following two datasets related to the Florida party registration validation are compared in Script `09_FL_partyreg_EI.R`.
    - `FL_val_ei_by_cd.rds`: Estimates of party registration in Florida using precinct data. This dataset is created in script `13_EI_FL_validation.R` with dataset `FL_2016_partyreg_race.csv`.
    - `FL_val_MRP_partyreg_by_race.csv`: Estimates of party registration in Florida using MRP. The variables are named as in the MRP dataset, but the estimates represent party registration rather than voteshare. It is constructed in the "build" project script `08_FL_validation_model.R` as described in the paper.

The following files are source data for the above datasets.

- `totalvoters_by_cd.csv`: A voterfile based measure of total voters in a CD. This is only used when aggregating up CD-level numbers to the state, region, and national levels. The summary statistics come from YouGov's database of Targetsmart voterfiles. Utah's numbers in 2020 were low, likely due to lags in the data update. For this state, we used our own estimates explained in the paper.

- `vtd-cd10-2020_pres-16-20.csv.gz` : A national precinct-level dataset used to compare EI and MRP in Appendix B. As described in the Appendix, this dataset was constructed directly from McCartan et al.'s ALARM data files, which in turn rely on VEST precinct data.

- `FL-2016_partyreg-race.csv` : Precinct-level registration data used for validation (Figure 2). Validation of party-race registration relies on precinct-level registration aggregates from the Catalist Q-tool, a web based query that allows for extracts of aggregated voter file data. Specifically, we extract precinct-level counts of actively registered voters from the 2016 general election, for White, Black, Hispanic, and "Other" (groups not belonging to the former categories) voters registered as Democrats, Republicans, or other partisan affiliations (voters not belonging to the major two parties). Catalist precinct aggregates do not include congressional district identifiers. Therefore, we merge Catalist aggregates to a crosswalk of non-split precincts using a combination of Presidential and House returns data from the MIT Election and Data Science Lab (MEDSL).

- `shp_cd` and `shp_cd` are shapefiles for the cartograms used in the paper. They come from the donnermap package (Kuriwaki and Donner, https://www.github.com/kuriwaki/donnermap).

Data and scripts in `build`

The `build` half of the project constructs the two-way calibrated MRP and post-stratification table. The scripts should be run in order from the Rstudio project in the dataverse. Be sure to install the dataverse package in its entirety to preserve the file structure.

Scripts in `build`

The scripts provided in this project are below. Where the time is long, I have provided a rough sense of the time it takes to run the scripts. All times are from a personal Mac laptop with a M1 chip. So that users do not

need to run all commands, we have provided intermediate output in the dataverse, such as the fitted regression objects and post-stratification table.

- `01_prep-cces.R` : Format CCES survey data
- `02_prep-acs.R` : Start making the poststrat table from ACS data, assisted by survey data. [1hr]
- `03_model-turnout.R` : Apply a turnout mode to the poststrat table made in 02 [1hr]
- `04_fit_mr.R` : Fit the multilevel (hierarchical) regression [5hr]
- `05_poststrat.R` : Predict the regression on the post-strat cells
- `06a_calib-oneway.R` : Calibrate estimates to geographic voteshare [1hr]
- `06b_summ-to-race-cd.R` : Add up cells to cd-race cells
- `06c_calib-twway-national.R` : Calibrate estimates again to national race voteshare
- `07a_summ-statelevel.R` : On the side, summarize SEs to state level
- `07b_summ-sims_export.R` : Final summary to CD-level and export dataset [2hr]
- `08_FL-validation-model.R` : Standalone Florida validation
- `09_prior-pd.R` : Prior predictive distribution (Figure A1)

Data in `build/data/output`

- `ccc_2016-2020_voted_2pty.rds` : A cleaned version of the CCES. Script `01_prep-cces.R` constructs this data from the Cumulative CCES. This file has standardized race into four categories as used in the paper, and limited the set of variables used. It also subsets to validated voters and those who expressed reporting for the Democratic or Republican Presidential party, as the script documents. The dataset `ccc_2016-2020.rds` is a larger set of the data with the same variable formatting but without limiting to the validated voters and two-party voters – this file is used for the construction of the post-stratification table.

- `brm16_fit.rds` and `brm20_fit.rds` are the fitted brms regression objects that are saved from `05_fit_mr.R` . They were last computed

in July 2022.

- `dir_fit_race` is a dataframe of direct (i.e. non-partially pooled) estimates from the survey. They are easily reproduced in `04`

- The post-stratification process in scripts `06a - 06c` save their output in intermediate steps due to the computational time it takes. `fs::dir_create()` make those files as necessary. The main output of the process is `mrp-full-posterior.csv.gz` , which is saved in the "analyze" project.

- `by-cd_pct-race_elec.csv` : For each CD, what is the proportion of a given race? These estimates are computed from the post-stratification table.

- The **post-stratification table** is the final output of scripts 03 and 04. This is referred to as `acs_synth-turnout.rds` in the code. The scripts can take several hours to run with the current implementation of the synthetic estimator. (A copy of this table is also made available in a separate Dataverse repository, https://doi.org/10.7910/DVN/MAZNJ6 under the `poststrat_synth.csv` table) To replicate the results in the APSR paper, use `acs_synth-turnout.rds` instead.

Data in `build/data/input`

1. `ccc_2016-2020_dataverse-original.rds` : The CCES original dataset, only subsetted to 2016 and 2020. The original sources of this data is the Dataverse repository on the cumulative CCES https://doi.org/10.7910/DVN/II2DB6, version 7.

2. `cd_presvote_2016-2020.csv` : Presidential vote in the district, taken from Daily Kos. This is essentially a combination of `ccesMRPprep::cd_info_2016` and `ccesMRPprep::cd_info_2020` .

3. `cd_turnout-vap.csv` : The total turnout and VAP for each CD Year. The turnout is reported by the Presidential vote recorded in Daily Kos, and the VAP is reported in the ACS.

4. `cd-race_turnout-CPS-estimate.csv` : The estimated turnout (as a fraction of VAP) by race, in the whole country. We start with the CPS microdata (November supplement) estimating turnout. We then propose the following: First, estimate the composition of each race in the self-reported turnout population by the CPS, using Voter Supplement weights. Then, multiply the composition by the total amount of votes for President in the nation as reported by McDonald. That serves as the numerator of the turnout. Finally, divide this by the ACS count of the total number of VAP for each race in the country. This serves as the denominator.
    - `frac_VAP` : Composition of race among turnout reported by CPS
    - `totalvotes` : votes for President (McDonald)
    - `turnout_CPS` is the final estimate of the turnout rate
5. `exitpoll_target-Rvote.csv` : The target of Trump voteshare by race we use for two-way calibration. This comes from the NEP Exit Poll estimates, which we report on in Appendix B.
6. `states-groupings.csv` : A table of 50 states with groupings used in synthetic table estimation