Data Codebook for 'Ticket Splitting in a Nationalized Era'

Shiro Kuriwaki

Here is an overview of the datasets and their variable names:

Parquet files

.parquet files are large datasets and represent the main cast vote record data. You will need the `arrow` package to read these files (see the overall README on that).

```
## ├── by-votechoice
## │   └── part-0.parquet
## ├── by-person-ID
## │   └── part-0.parquet
## ├── maryland
## │   └── part-0.parquet
```

`by-votechoice` is stored in long form where one row is a vote choice for a particular office.

- `elec` : South Carolina Election
- `voter_id` : Voter ID I assigned within election
- `Dvoter` : Proportion of votes for Democrat statewide, defined as D in paper
- `top_party2` : The top of the ticket vote (President / US-Senate), defined in paper
- `top_party2_alt` : An alternative measure defined in paper Fig. B5
- `office` : Down-ballot office
- `party` : Party choice for down-ballot office. -1: Democrat, 1: Republican, 0.5: Third party, 0 for other. Only available if contest is contested by more than one major party.
- `ncand` : Number of major party candidates running in that contest
- `jID` : Contest ID
- `open` : Whether the contest is open, i.e. there is no incumbent
- `copar` : The interaction of `party` and `top_party2` , i.e. a straight vote
- `inc_copar` : The interaction of `copar` and incumbency of the candidate.

`by-person-ID` indicates contest IDs for each voters.

- `voter_id` , `elec` : Same as above
- `precinct_id` : An identifier for precinct
- `county` : County
- `USH_jID` , `HOU_jID` , `SEN_jID` , `SHF_jID` , `CCD_jID` , `JPR_jID` : Contest IDs for six offices.

`maricopa`, `maryland`, and `palmbeach` are cast vote records from the three other states I examine in the paper. Only `maryland` is a parquet; the two others are csvs. These are organized in wide format, so each row represents a voter instead of a choice on the long ballot. These are analyzed in scripts starting with `10_`.

CSV files

.csv files are typically small datasets with metadata and summary statistics specific to a particular script.

```
## ├── by-HOU-USH-dist_split.csv
## ├── by-contest_cand-metadata.csv
## ├── deluca_quality.csv
## ├── hist-elecs_by-office.csv
```

`by-contest_cand-metadata.csv` gets used the most frequently for candidate-level metadata in South Carolina. Sources: originates from South Carolina Election Commission, my own data collection, and DIME (version 3.1). Limited to D vs. R contests. Used in almost all scripts.

- `elec`, `office`, `jID`, `county`, `open`: Same as above but for candidates
- `dist`: District number if applicable.
- `n_dr`: Number of D and R candidates.
- `row_id_R`, `row_id_D`: Row ID for candidates
- `cand_R`, `cand_D`: Names of candidates
- `incumbency_R`, `incumbency_D`: Incumbency indicator for candidates
- `hits_R`, `hits_D`: Number of newspaper mentions in term, described in main text.
- `money_R`, `money_D`: Dollars raised, described in main text
- `Rmoneyadv`, `Rnewsadv`: Log ratios of hits and money, described in section A5.

`deluca_quality.csv` is a small dataset from DeLuca, described in main text. Used in Figure 5.

- `state`, `elec`, `office`, `dist`: Same as above
- `top_office`: Reference office used to define ticket splitting
- `split_for_D`, `split_for_R`: split ticket rates from CVR
- `straight_for_D`, `straight_for_R`: straight ticket from CVR
- `quality_differential`: DeLuca's main measure of quality differential

`by-HOU-USH-dist_split.csv` is a small file estimating the split ticket rate for State House.

- `elec`, `year`: Election year
- `USH_dist`, `HOU_dist`: Combination of US House and State House district IDs
- `hstraight`: Proportion of voters voting the same party in the two pairs of contests
- `n`: Number of voters

`hist-elecs_by-office.csv` is a dataset of historical elections. Appendix A6. Data collected by David Lublin, Carl Klarner, and supplemented by myself. The table bins the contest-level dataset

into bins.

- `office` : Office examined
- `yr_bin` : The year range covered
- `n` : Number of contests
- `pct_R` , `pct_D` : Percent won by Republican, Democrat
- `mar_R` : Win margin of Republican

Stata dta files

.dta files are also for one-off uses of survey data. Sources: ANES and CCES.

```
## ├── hist-svy_anes.dta
## ├── hist-svy_cces.dta
## ├── hist-svy_cd-2020.dta
## ├── hist-svy_cd.dta
```

`hist-svy_anes.dta` is an extract from the ANES cumulative file, coded similar to Jacobson (2015, JOP)'s replication data. Figure 1.

- `year` : Year, VCF0004 in ANES
- `weight` : Survey weight, VCF0009x in ANES
- `id` : Respondent ID, VCF0006a in ANES
- `state` : state, VCF0901b in ANES
- `cd` : congressional district, combination of state and VCF0900c
- `hvote` : House vote, 1 for Democrat and 0 for Republican, VCF0704 in ANES
- `pvote` : Presidential vote, coded the same way
- `straight` : Interaction of `hvote` and `pvote` , straight ticket
- `VCF0704` : Original House vote coding
- `VCF0707` : Original Presidential vote coding.

`hist-svy_cces.dta` is an extract from the CCES cumulative file, https://doi.org/10.7910/DVN/II2DB6. Figure 1.

- `year` , `case_id` , `weight` , `cd` : Year, respondent ID, weight, and district, as described in CCES cumulative codebook.
- `pres_v` , `pres_i` : Presidential vote for post-election vote (v) and pre-election intent (i)
- `rep_v` , `rep_i` : Same but for US House
- `straight` , `split` : Interaction between `pres_v` and `pres_i`

`hist-svy_cd.dta` is CD-level information about the contestedness of each district in each election. We limit our analysis to these districts. Figure 1.

- `year` , `st` , `cd` : District identifiers
- `contes` contested by D and R candidate. My coding and Jacobson's coding for earlier years.

`hist-svy_cd-2020.dta` is a single-year file indicate the Presidential vote in each district. Used in Figure B7 as a robustness check. `pct_trump20` is the voteshare for Trump in 2020, and `pct_trump16` is the voteshare for Trump in 2016.

Rds files

.rds files are cluster objects estimated from the clusterCVR described in the main text.

```
## ├── clusters
## │   ├── by-K
## │   │   ├── D12_list.rds
## │   │   ├── D16_list.rds
## │   │   ├── R12_list.rds
## │   │   ├── R16_list.rds
## │   │   ├── p12_list.rds
## │   │   └── p16_list.rds
## │   ├── p12_D-subset_k4.rds
## │   ├── p12_R-subset_k4.rds
## │   ├── p16_D-subset_k4.rds
## │   └── p16_R-subset_k4.rds
```

Each file with subset `k4` is an output of the clusterCVR package using the same 2012 and 2016 parquet data as above. Used in Figure 3. See the documentation of the package at https://github.com/kuriwaki/clustercvr for details on the output. Overall, the main parameters are

- `pi` : estimates of bloc sizes
- `mu` : array of vote choice for each office in each bloc (cluster)
- `loglik` : The log likelihood fit of the final iterations

`D` indicates Clinton/Obama voters in 2012 or 2016, and `R` indicates Trump/Romney voters. These are subsets of the data.

Each file in `_list` follows the convention of the rest of the files but show summary statistics of the log likelihood fit for a series of clusters, 2 to 10. Each item in the list represents a choice of the cluster value K. Used in Appendix C, Figure C1.