Guide to the CCES Candidates Dataset

Jaclyn Kaslovsky[1], Shiro Kuriwaki[2], James M. Snyder, Jr.[2], and Stephen Ansolabehere[2]

[1]Center for the Study of Democratic Politics, Princeton University.
[2]Department of Government, Harvard University.

September 2020

Between 2006-2018 alone, there were 3,460 races for Congress and Governor in the general election, in which 6,684 Democratic and Republican candiates ran for office. The CCES asked about all these choices. Many researchers use the CCES to analyze how a respondent evaluated a candidate in a race for Congress or Governor. The CCES data refers to such candidates by number, where `Cand1` and `Cand2` for a given respondent i represents the two candidates who are contesting i's seat, and the dataset records whether, for example, i reported voting for `Cand1` over `Cand2` even though the respondent sees the actual numbers while taking the survey.

Now it is often tedious to immediately tell which candidate `Cand1` and `Cand2` refer to, without cross-checking with other columns in the data. Usually, candidate 1 is the Democrat and candidate 2 is the Republican, but there are some exceptions, and of course two respondents in different districts have different candidates, with different names, incumbency statuses, or ideology scores. Moreover, some of the reference variables in the Common Content, like `HouseCand1IncumbentNum`, contain errors.[1]

Here we provide two datasets to facilitate and form the basis of such analyses:

1. A **candidate level** dataset: Includes new data such as incumbency, general election vote, party, and the election result, from Jim Snyder's historical candidate level datasets. There is one row for every candidate in each election they ran in.
2. A CCES **respondent level** dataset then provides a linkage between a constituent and the candidates who are running in that district. There is one row per candidate reference, e.g. if there are N voters voting for three offices (House, Senate, Governor), and each race consists of three candidates (Democrat, Republican, Third party), then the dataset will contain $3 \times 3 \times N$ rows.

[1]For example in the 2014 Common Content, `HouseCand1IncumbentNum` for respondents in WA-05 are miscoded. Cathy McMorris Rodgers was the incumbent, not Joseph Pakootas. Additionally, 4 out of 127 respondents in `IL-03` (Lipinski, D) have a missing `HouseCand1IncumbentNum` value.

1 Examples, Unique Identifiers, and Counts

The respondent-level data is formatted as in Table 1.

Table 1 – Example of Respondent Data Format

Respondent-Level Information				Race-Level		Candidate-Level Information				
year	case_id	st	dist	office	totalvotes	cand	name_snyder	party	inc	votes
2016	304099877	MO	5	H	324,270	1	CLEAVER, EMANUEL, II	D	1	190,766
2016	304099877	MO	5	H	324,270	2	TURK, JACOB	R	0	123,771
2016	304099877	MO	5	S	2,802,546	1	KANDER, JASON	D	0	1,300,200
2016	304099877	MO	5	S	2,802,546	2	BLUNT, ROY D.	R	1	1,378,458
2016	304099877	MO	5	G	2,803,018	1	KOSTER, CHRIS	D	0	1,277,360
2016	304099877	MO	5	G	2,803,018	2	GREITENS, ERIC	R	0	1,433,397

Each `year` × `case_id` combination uniquely defines a respondent. Use these two variables to merge back to the cumulative common content.

In addition, use the `office` × `cand` variables to uniquely define a candidate for a given respondent. For example, in Table 1 we see that there are six rows for case ID `304099877` in the 2016 CCES: two candidates for three offices (House, Senate, and Governor). In all three cases, `cand == 1` is the Democrat and `cand == 2` is the Republican.

Candidates run at the level of districts. Therefore, when merging respondents with candidates, it is sufficient to merge on `office` × `state` × `dist` × `name_snyder`. **JK: Should we still say this even though we took out st and dist from the respondent-level dataset?** We use full name instead of party to distinguish candidates because party or last name is not sufficient: in rare occasions, candidates of the same party do contest the same seat and it is sometimes ambiguous what constitutes the last name of a candidate.

The candidate-level dataset shares the same values of the respondent-level data, but is more compact because there is only one row per candidate, as shown in Table 2.

Table 2 – Example of Candidate Data Format

year	st	office	dist_up	party	name_snyder	inc	votes	w_g	totalvotes
2016	MO	H	5	D	CLEAVER, EMANUEL, II	1	190,766	1	324,270
2016	MO	H	5	R	TURK, JACOB	0	123,771	0	324,270
2016	MO	H	5	Lbt	WELBORN, ROY	0	9,733	0	324,270
2016	MO	S	NA	D	KANDER, JASON	0	1,300,200	0	2,802,546
2016	MO	S	NA	R	BLUNT, ROY D.	1	1,378,458	1	2,802,546
2016	MO	S	NA	Const	RYMAN, FRED	0	25,407	0	2,802,546
2016	MO	S	NA	Grn	MCFARLAND, JOHNATHAN	0	30,743	0	2,802,546
2016	MO	S	NA	Lbt	DINE, JONATHAN	0	67,738	0	2,802,546

Please note there are two versions of respondent-level datasets, for pre-election and post-election. The even-year CCES has a pre-election wave, which could start as early as October, and a post-election wave, which can occur as late as mid-November. Starting from 2010, the CCES started distinguishing the candidates between pre and post waves, in case the voter moved districts before and after the election or the candidates changed.

Users should pick the dataset that is relevant for the question of interest. For example, pre-election vote intent questions are asked among pre-election respondents and post-election vote choice questions are asked among post-election respondents. A few (about a less than 1 percent) of respondents change their state district sometime between the pre and post waves, so the candidates can differ in such instances.

In the subsequent tabulations, we only show the pre-election wave numbers for simplicity.

Table 3 summarizes the number of unique respondents and the number of rows.

Table 3 – Summary of Counts

Pre-Election Wave				Post-Election Wave			
		Respondent × Office × Candidate Pairs				Respondent × Office × Candidate Pairs	
Year	Respondents	Non-Missing	Total	Year	Respondents	Non-Missing	Total
2006	36,403	178,739	180,203	2006	0	0	0
2008	32,747	108,664	109,073	2008	0	0	0
2010	55,400	302,789	307,413	2010	46,684	255,340	259,108
2012	74,023	275,791	278,054	2012	45,017	168,480	169,632
2014	56,200	274,844	283,615	2014	48,888	239,248	246,879
2016	64,600	243,918	247,372	2016	52,899	200,454	203,234
2018	68,217	368,022	371,079	2018	51,808	278,449	280,858

Note: 2006 and 2008 post-waves did not re-ask location, so they would be identical to the pre-waves.

2 Data Sources

The *respondent-level data* is an intermediate output of the creation of the Cumulative CCES Dataset, available on Dataverse https://doi.org/10.7910/DVN/II2DB6.

The *candidate-level data* is a subset of the data collected by James M. Snyder. Vote counts and candidate listings of Congress are entered from the House Clerk's Official Election Reports, at https://history.house.gov/Institution/Election-Statistics/. Election results for the state office of Governor are collected from statements of votes by each state's Secretary of State. Incumbency is collected by a manual inspection of the candidate's biography.

3 Usage Example: Merging to the CCES

Below is example R code of how to use the candidate data in conjunction with other CCES data. This answers the following question: are voters in the racial minority more likely to vote for the losing candidate? Hajnal (2009) showed that this was the case and raised issues with representation.

This code

1. Loads the data: CCES cumulative, candidate data (respondent-level), candidate data (candidate-level)
2. Slims it down to necessary covariates
3. Limits to contested races
4. Merge the candidates information on victory into the Cumulative Common Content
5. Analyze the relationship between voter race and candidate victory

```
library(haven)
library(tidyverse)
library(fs)

rel_dir <- "~/Dropbox/CCES_candidates/Release/"
ccc_dir <- "~/Dropbox/cces_cumulative/data/release"

# Read datasets ------

## Candidate dataset
cand_case <- read_dta(path(rel_dir, "cces_candidates_pre.dta"))
cand_info <- read_dta(path(rel_dir, "candidates_snyder.dta"))

## Cumulative CCES (separate dataset)
ccc_cumulative <- read_rds(path(ccc_dir, "cumulative_2006_2019.rds"))

# Slim down data ----
## Cumulative
ccc_house <- ccc_cumulative %>%
  select(year, case_id, st, dist_up, cd, race,  intent_rep_chosen) %>%
  mutate(cand = as.integer(intent_rep_party),
         race = as_factor(race))

## candidate data by respondent
cand_df <- cand_case %>%
  select(year, case_id, cand, name_snyder, won)
```

```r
# Find contested races with a D and a R ----
race_df <- cand_info %>%
  filter(office == "H") %>%
  select(year, st, dist_up, party, inc, won) %>%
  group_by(year, st, dist_up) %>%
  summarize(n_Ds = sum(party == "D", na.rm = TRUE),
            n_Rs = sum(party == "R", na.rm = TRUE),
            .groups = "drop")
contested <- race_df %>%
  filter(n_Ds == 1, n_Rs == 1) %>%
  select(year, st, dist_up)

# Merge candidate dataset to cumulative
# subset to contested candidates and add whether the candidate won
house_df <- ccc_house %>%
  filter(race %in% c("White", "Black", "Hispanic")) %>%  # subset to three races
  inner_join(contested, by = c("year","st", "dist_up")) %>% # subset to contested
  left_join(cand_df, by = c("year", "case_id", "cand"))

# Results -----
# summary statistics - are racial minorities likely to vote for losing candidates?
results_long <- house_df %>%
  group_by(year, race) %>%
  summarize(vote_for_winning = mean(won, na.rm = TRUE),
            n = n())

## present in table
pivot_wider(results_long,
            id_cols = year,
            names_from = race,
            values_from = vote_for_winning)
```

Year	White	Black	Hispanic
2006	0.58	0.71	0.58
2008	0.58	0.64	0.63
2010	0.60	0.52	0.59
2012	0.58	0.70	0.62
2014	0.60	0.52	0.61
2016	0.59	0.57	0.61
2018	0.60	0.63	0.62

4 Variable Descriptions

4.1 Respondent Level Variables

- `case_id`: Case (i.e. respondent) identifier

year	case ID	
	Rows	Unique Cases
2006	180,203	36,403
2008	109,073	32,747
2010	307,413	55,400
2012	278,054	74,023
2014	283,615	56,200
2016	247,372	64,600
2018	371,079	68,217

- `year`: CCES year
- `dataset`: The source of CCES data. Most of the time this will be the Common Content, which we denote with simply the year, so `year == dataset`. There are two exceptions. `2012p` refers to the 2012 Panel Study. `2018c` is the Competitive District Study of 2018, separate from the Common Content.
- `st`: State Abbreviation
- `dist`: Congressional district number for current Congress.
- `dist_up`: Congressional district number for upcoming Congress. This variable will differ from `dist` when the respondent has been redistricted. As a result, this variable is most commonly different from `dist` in 2012 after the new districts from the decennial census went into effect.

4.2 Candidate Variables

- `office`: The office the candidate is running for. Following the Snyder data, we use `H` for the US House of Representatives, `S` for the US Senate, and `G` for Governor.

Respondents

year	office		
	G	H	S
2006	57,690	68,889	53,624
2008	10,643	65,338	33,092
2010	100,633	110,140	96,640
2012	19,564	145,251	113,239
2014	88,178	124,647	70,790
2016	21,192	127,372	98,808
2018	117,962	142,130	110,987

Candidates

year	office		
	G	H	S
2006	173	1,230	144
2008	38	1,260	116
2010	186	1,365	189
2012	39	1,294	141
2014	137	1,147	142
2016	45	1,114	148
2018	160	1,122	144

- `cand`: Candidate number for respondent. This variable tells the user whether the candidate is candidate number 1, 2, or 3 in order to match to other variables included in the CCES, such as `HouseCand1_Gender`, for example.

year	cand		
	1	2	3
2006	91,615	88,588	0
2008	53,127	51,127	4,819
2010	138,398	141,274	27,741
2012	134,745	136,805	6,504
2014	125,411	126,346	31,858
2016	119,681	119,249	8,442
2018	173,076	169,940	28,063

- `name_snyder`: Standardized candidate name from James Snyder. The syntax is `[Last name], [First Name] [Middle name] ([Nickname]), [Jr/Sr/I/II/III]`. Some examples of names are below, to give a sense of the syntax.

  `SEWELL, TERRYCINA ANDREA (TERRI)`: commonly known as Terri Sewell (AL)

  `GRASSLEY, CHARLES ERNEST (CHUCK)`: commonly known as Chuck Grassley (IA)

  `CORNYN, JOHN, III`: commonly known as John Cornyn (TX)

  `KENNEDY, JOSEPH P. (JOE), III`: commonly known as Joe Kennedy (MA)

  `WASSERMAN SCHULTZ, DEBBIE`: note the last name is not hyphenated and is two words

  In order to make names comparable across years, Snyder uses the full name, i.e. spelling out middle names, as much as possible and goes beyond what is printed on the ballot or the House Clerk document.

- `party`: The party affiliation of the candidate. We use the "short" or colloquial party name. For example, the Democrat-Farmer-Labor Party in Minnesota is given a `D` instead

of `DFL`. Candidates who ran on third party tickets in Connecticut and New York are simply given the major party name.

Respondents

year	party					
	D	R	I	Lbt	Grn	Other
2006	90,029	88,150	445	0	0	115
2008	53,023	51,373	1,004	2,106	401	757
2010	138,244	140,789	11,601	6,960	2,654	2,541
2012	133,980	136,047	2,154	2,035	182	1,393
2014	122,719	123,603	5,033	18,681	2,187	2,621
2016	123,894	109,841	1,261	6,710	1,085	1,127
2018	179,027	162,617	2,750	16,397	4,180	3,051

Candidates

year	party					
	D	R	I	Lbt	Grn	Other
2006	527	479	107	138	62	234
2008	499	462	92	145	56	160
2010	496	532	192	192	73	255
2012	492	487	112	150	63	170
2014	504	489	123	136	42	132
2016	461	448	76	145	61	116
2018	514	470	103	151	43	145

- `party_formal` The formal party names, for example DFL in Minnesota. Third party candidate names follow what is given in the House Clerk Document.
- `inc`: Candidate incumbency status. A 0 means the candidate is not an incumbent, a 1 means the candidate is an incumbent, and a 2 means the candidate is an incumbent that was elected in a special election.
  - This may differ from the CCES incumbency variable in some redistricting cases when two incumbents were forced to run against each other. For example, in 2012 Betty Sutton (OH-13) and Jim Renacci (OH-16), both House incumbents, were forced to run in OH-16. While the CCES incumbency variable (HouseCandIncumbent) lists only Sutton as the incumbent, this dataset will list both Sutton and Renacci as incumbents. This is why we include the variable `current_inc` below.

Respondents

year	inc			
	0	1	2	3
2006	99,551	73,804	3,948	1,747
2008	61,174	46,645	506	339
2010	215,981	81,673	1,029	4,106
2012	168,916	98,920	7,437	0
2014	163,239	105,661	2,576	0
2016	146,116	99,056	1,159	0
2018	236,043	126,893	3,178	1,908

Candidates

year	inc			
	0	1	2	3
2006	1,019	520	5	3
2008	911	493	8	2
2010	1,299	424	9	8
2012	1,000	459	13	0
2014	888	502	8	1
2016	878	418	5	0
2018	1,003	408	11	4

- `current_inc`: Candidate incumbency status for that specific respondent. A 0 means the candidate is not the respondent's current representative, a 1 means the candidate is the respondent's current representative. A candidate could be an incumbent in the district but the not the respondent's current representative if the respondent was redistricted. This variable is only included for respondents with `inc=1`.

- `candidatevotes`: The number of total votes the candidate received.

  - For candidates running on multiple party tickets, this will be the *sum* of all of their votes. For example, in 2016, Rep. Rosa L. DeLauro (CT-03) ran as a Democrat and also ran as a Working Families Party candidate. She won 192,274 votes in the former and 21,298 votes in the latter, so her `candidatevotes` is the total, 213,572.
  - Florida does not report the vote count for a House candidate if she is unopposed. In these case, we have the vote count as `NA` but have the candidate winning (`w_g == 1`).

- `totalvotes`: Total votes for all candidates in the general election included in Jim Snyder's data

- `won`: Candidate won the general election (0/1)

- `data_note`: 145 candidates were not included in the election dataset we used to collect the candidate-level variables. We still include these candidates in the supplemental data in case researchers are interested in specific districts or races. The reason for these non-merges can be broken down into four categories, listed below.

  1. Incorrect Election: Candidates with this code did not run in the respondent's district or state. This likely occurs because the respondent's information was incorrectly entered.
  2. Not on General Election Ballot: Candidates with this code either withdrew from the race, were disqualified, were write-in candidates, did not make it to the included runoff election, or did not receive enough votes to appear on the ballot.
  3. Unopposed in Oklahoma: In Oklahoma, unopposed candidates do not appear on the general election ballot.

4. Missing from election data: Candidates with this code were not included in the data we used to merge in the election information. Often times this category is made up of candidates in races with jungle primaries, such as Louisiana. This also includes all candidates from Washington D.C..

5 Extensions to Candidate Gender and Race

Our dataset does not contain information about the candidate's gender or race. We may add to the dataset in future versions, but in the meantime there are several related data sources users can rely on.

Numerous years in the CCES have data available on candidate race and gender for interested researchers. Please see the table below for further information regarding the availability of such information by year and where it can be located.

Table 4 – The Availability of Candidate Race and Gender Data by Year

CCES	Candidate Race	Candidate Gender
2006		
2008	Variables for H, S, G	Variables for H, S, G
2010	Supplemental Data for H[1]	Variables for H, S, G
2012	Supplemental Data for H[2]	Variables for H and S
2014	Supplemental Data for H[3]	
2016	Supplemental Data for H, S[4]	
2018		Variables for current H post

Note that the 2013 common content also includes the gender of House members.

1. https://doi.org/10.7910/DVN/KC9EQR
2. https://doi.org/10.7910/DVN/NI3BDE
3. https://doi.org/10.7910/DVN/D1N0GO
4. https://doi.org/10.7910/DVN/IAOZOU

Note that some of this data has been aggregated in the Cumulative Contextual File, located here: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26451&version=5.0&widget=dataverse@cces.

6 Merging to Other Datasets

ICPSR

7 Version History

- Dataverse 1.0: Initial Release, **SK: Enter date of dataverse upload here**.