# User's Guide and Codebook for the CCES 2016 Voter Validation Supplemental Data

Ted Enamorado* Kosuke Imai†

May 15, 2018

*Ph.D. Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Email: tede@princeton.edu

†Professor, Department of Politics and Center for Statistics and Machine Learning, Princeton University. Professor of Visiting Status, Graduate Schools of Law and Politics, The University of Tokyo. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: http://imai.princeton.edu

**Suggested citation:** Enamorado, Ted, and Kosuke Imai. 2018. "User's Guide and Codebook for the CCES 2016 Voter Validation Supplemental Data." Technical report, Princeton University.

**Acknowledgments:** This report was prepared by Ted Enamorado and Kosuke Imai. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors. The authors thank Bruce Willsie of L2, Inc for making the national voter file available and Steffen Weiss of YouGov for technical assistance.

See the CCES Common Content Guide 2016 for more information on the CCES 2016 Study.

# Contents

**1 Overview of Voter Validation Data and Matching to CCES Respondents** **4**

1.1 The National Voter File . . . . . 4

1.2 The Merging Process . . . . . 4

1.2.1 The Within-state Merge . . . . . 5

1.2.2 The Across-state Merge . . . . . 6

1.2.3 Clerical Review . . . . . 8

**2 Codebook** **8**

**3 How to use the Voter Validation Dataset** **15**

**CCES 2016 Voter Validation Study at a Glance**

- **Title:** CCES 2016 Validation Supplemental Data.
- **Purpose:** To validate self-reported turnout and registration from the CCES 2016 Common Content Study.
- **How to use with CCES 2016:** The `csv` file `cces2016voterval.csv` can be merged with the CCES 2016 via the case ID variable `V101`.
- **Data source:** The turnout variables in `cces2016voterval.csv` come from the nationwide voter file facilitated for academic purposes to the Princeton Team by L2, Inc.
- **Number of records in this dataset:** 64600 observations
- **Number of CCES respondents:** 64600 observations
- **Data:** Data are available free of charge from: `https://cces.gov.harvard.edu/`

# 1 Overview of Voter Validation Data and Matching to CCES Respondents

## 1.1 The National Voter File

For this validation project, we have established an academic data use agreement with L2, Inc., a leading national non-partisan firm and the oldest organization in the United States that supplies voter data and related technology to candidates, political parties, pollsters and consultants for use in campaigns. L2 provided us with a copy of the nationwide voter file in July 2017. That file includes over 180 million records of registered voters, of which 131 million are recorded as voting in the 2016 General Election. We note that this nationwide voter file contains approximately 4 million fewer registered voters who cast a ballot in the 2016 presidential election than the total number of actual votes recorded in the United States Election project (`http://www.electproject.org/`). As a result, the turnout rate among the voting eligible population based on the L2 data is about 1 percentage point lower than the actual turnout. This discrepancy arises from the fact that L2 had already removed those who had deceased and moved out of state (but had not registered in another state yet) between the election day and the time when we received the voter file.

## 1.2 The Merging Process

The lack of a unique identifier that unambiguously links records is one the most challenging problems to overcome when linking records from two datasets. For example, not all datasets include a unique social security number that allows for a perfect merge. Using fields that are common between datasets seems to be the next reasonable step; however, these variables are sometimes imperfectly coded due to misspellings in names, the use of nicknames, missing observations, duplicates, etc. This makes any merging process based on such fields an uncertain one — e.g., deterministic decision rules based on noisy fields do not work well in practice. Moreover, it is difficult to judge the precision of proprietary methods used by many vendors due to the secrecy on the algorithms used.

To address this uncertainty, we apply a canonical probabilistic record linkage model (Fellegi and Sunter, 1969), improved and implemented via the open-source software package fastLink (Enamorado, Fifield and Imai, 2017). In addition to accounting for the uncertainty process which can

be adjusted for in subsequent empirical analyses, fastLink is transparent, open-source, and makes use of a scalable algorithm that allows for the merging of large datasets. We invite the interested reader to learn more about fastLink at `https://cran.r-project.org/package=fastLink` and through Enamorado, Fifield and Imai (2017). To merge the nationwide voter file to the 2016 ANES involves first conducting the within-state merge followed by the across-state merge. Finally, we conduct a clerical review. We detail each step below. The detailed description of validation results for the CCES and the American National Election Study (ANES) are described in Enamorado and Imai (2018).

### 1.2.1 The Within-state Merge

The aim of the within-state merge is to link the records of individuals who remained in the same state from the time of their CCES interview to the time when the voter files were updated. These individuals include those who remained in the same residence and those who moved within a state and updated their address in the voter file. Since a voter file for each state contains millions of records, we further reduce the scale of each merge process by additionally blocking observations on gender within each state, producing a total of 102 separate merges (50 states + DC $\times$ 2 gender categories). To perform each merge, we use the following linkage fields which are present in both datasets:

- First Name
- Last Name
- Age
- House number
- Street Name
- Postal code

To make a comparison between the values of each linkage field across datasets, we follow the literature (see e.g., Winkler, 1990; Cohen, Ravikumar and Fienberg, 2003), and use three levels of agreement for the string valued variables (first name, last name, and street name) based on the Jaro-Winkler distance with 0.85 and 0.94 as the thresholds. We also use three levels of agreement

for age based on the absolute distance between values, with 1 and 2.5 years as the thresholds for separate agreements, partial agreements, and disagreements, respectively. For the remaining variables (i.e., house number and postal code), we utilize a binary comparison based on exact matching, indicating whether they have an identical value. Specifically, for each one of the 102 state-gender blocks, we used the following code:

```
library("fastLink")
matches.out <- fastLink( dfA = subset.1, dfB = subset.2,
                         varnames = c("first_name", "last_name", "age",
                                      "house_number", "street_name",
                                      "zip_code"),
                         stringdist.match = c("first_name", "last_name",
                                              "street_name"),
                         numeric.match = c("age"),
                         partial.match = c("first_name", "last_name",
                                           "street_name", "age"),
                         cut.a = 0.94,
                         cut.p = 0.85,
                         cut.a.num = 1,
                         cut.p.num = 2.5,
                         threshold.match = 0.0001
                       )
```

Listing 1: Within-state merge via fastLink

In the above code, `subset.1` (`subset.2`) represents the subset of a given state-gender block for the ANES (voter file). The names of the six variables used in the within-state merge are specified in `varnames`, while `stringdist.match` and `numeric.match` contain the list of variables that will be compared using string and numeric distance measures, respectively. The `partial.match` argument contains the list of linkage fields for which we make a comparison using three discrete levels (agreement, partial agreement, and disagreement. Finally, the `cut.a`, `cut.p`, `cut.a.num`, and `cut.p.num` arguments specify the thresholds used for the string and numeric distance comparison. For more details on these options and extra features of fastLink, please see `https://cran.r-project.org/package=fastLink`.

### 1.2.2 The Across-state Merge

The main problem of the within-state merge is the possible failure to match individuals who changed their voter registration address between the day of the CCES interview and the time

when the nationwide voter file was updated. The within-state merge may also miss people who were registered to vote at a different address than the residential address recorded by the CCES if those two addresses belong to different states. In an effort to locate these individuals, we merged the sample of CCES with the within-state matching probability less or equal to 0.75 — there were 23585 such cases. We merged those observations with the whole voter file without any subsetting. The linkage fields used are:

- First Name
- Middle Name
- Last Name
- Age

To compare the values of each linkage field across two datasets, we used the binary agreement variable for the string valued variables (first name and last name) based on the Jaro-Winkler distance with 0.94 as the threshold. We also used the binary agreement variable for age based on the absolute distance between values, with one year as the threshold used to separate agreements and disagreements. After a careful clerical review of all the possible matches, we identified 874 CCES respondents that were determined to have a record in the voter file. The code for the across-state merge is given below.

```
library("fastLink")
matches.out <- fastLink( dfA = not.found.within, dfB = voter.file,
                         varnames = c("first_name", "middle_name", "last_name", "age"),
                         stringdist.match = c("first_name", "last_name"),
                         numeric.match = c("age"),
                         cut.a = 0.94,
                         cut.a.num = 1,
                         threshold.match = 0.9999
                       )
```

Listing 2: Across-state merge via fastLink

In the above code, `not.found.within` represents the subset of the CCES that could not be successfully matched in the within-state match step. The dataset `voter.file` is the full voter file without any subsetting. The remaining options in fastLink can be described in a similar fashion

as to those used for the the within-state merge, with the exception that string and numeric comparisons were made based on two agreement levels, either agree or disagree — in other words, we did not use partial matching.

### 1.2.3 Clerical Review

The final step of our validation procedure involves a careful clerical review as recommended in the literature (Winkler, 1995). The clerical review helps reduce any remaining uncertainty regarding the merging process by examining the suitability of each declared match obtained using fastLink. We discarded 4115 cases that fastLink declared as potential matches. In most instances, the discarded cases were due to individuals being matched to someone who lived in the same household and shared an identical name but with an age difference greater than 10 years.

# 2 Codebook

1. `V101`: 2016 CCES ID. Unique identifier that links each observation of the CCES 2016 to an observation in the CCES 2016 Voter Validation file.

   ```
   ---------------------------------------------------------------------------
   type:  numeric
   range:  [222168628, 307210331]
   unique values:  64600
   missing:  0/64600
   ---------------------------------------------------------------------------
   ```

2. `merge_type`: Merge Type. It equals to 1 if the match or non-match observation comes from the within-state merge and 2 if the match observation comes from the across-state merge.

   ```
   ---------------------------------------------------------------------------
   type:  numeric
   range:  { 1, 2 }
   unique values:  2
   missing:  0/64600
   ```

| Frequency | Value | Label |
|---|---|---|
| 63726 | 1 | Within-state Merge |
| 874 | 2 | Across-state Merge |

==========

3. agreement_pattern: Agreement Pattern. This is a string that summarizes the level of agreement across linkage fields.

==========

type: string

unique values: 323

missing: 0/64600

| Freq. | Value | Label |
|---|---|---|
| 28892 | FN: A \| LN: A \| AG: A \| HN: A \| SN: A \| ZC: A | Agree on: first name, last name, age, house number, street name, zip code. |
| ⋮ | ⋮ | ⋮ |
| 973 | FN: A \| LN: A \| AG: D \| HN: A \| SN: A \| ZC: A | Agree on: first name, last name, house number, street name, zip code; Disagree on: age. |
| ⋮ | ⋮ | ⋮ |
| 41 | FN: D \| LN: P \| AG: A \| HN: A \| SN: A \| ZC: A | Agree on: age, house number street name, zip code; Disagree on: first name; Partially agree on: last name. |

Key:

A: agree, P: partially agree, D: disagree, NA: missing value. FN: first name, LN: last name, AG: age, HN: house number, SN: street name, ZC: zip code

```
==========================================================================
```

4. `prob_match`: Probability of being a match. Probability that an ANES respondent is a match with an individual in the nationwide voter file conditional on their agreement pattern.

```
==========================================================================
type:  numeric
range:  [ 0, 1]
unique values:  6776
missing:  0/64600


mean:  0.6659
std.  dev:  0.4547

 percentiles:    10%     25%     50%     75%     90%
               0.0002  0.0173  1.0000  1.0000  1.0000
==========================================================================
```

5. `clerical_review`: Clerical Review. It equals to 1 if a pair of records between the ANES is declared a match after an extensive revision of each one of the pairings obtained from fastLink. This variable equals to zero if the pair of records is deemed to be a non-match, the latter includes instances where fastLink attach a high probability of being a match.

```
==========================================================================
type:  numeric
range:  { 0, 1 }
unique values:  2
missing:  0/64600

 Frequency  Value  Label
 26748      0      Non-match after clerical review
 37852      1      Match after clerical review
==========================================================================
```

6. vote2016: Unweighted turnout in the 2016 General Election. It equals to 1 if the best match observation from the voter file voted in the 2016 General Election and equals 0 otherwise. Note that when using the turnout variables for analysis, these need to be weighted by either the probability of being a match (prob_match) or by the clerical review indicator (clerical_review)

```
================================================================================
type:  numeric
range:  { 0, 1 }
unique values:  2
missing:  0/64600

 Frequency  Value  Label
 18332      0      Did not vote
 46268      1      Vote
================================================================================
```

7. vote2014: Unweighted turnout in the 2014 General Election. It equals to 1 if the best match observation from the voter file voted in the 2014 General Election and equals 0 otherwise. Note that when using the turnout variables for analysis, these need to be weighted by either the probability of being a match (prob_match) or by the clerical review indicator (clerical_review)

```
================================================================================
type:  numeric
range:  { 0, 1 }
unique values:  2
missing:  0/64600

 Frequency  Value  Label
 33219      0      Did not vote
 31381      1      Vote
================================================================================
```

8. vote2012: Unweighted turnout in the 2012 General Election. It equals to 1 if the best match observation from the voter file voted in the 2012 General Election and equals 0 otherwise. Note that when using the turnout variables for analysis, these need to be weighted by either the probability of being a match (prob_match) or by the clerical review indicator (clerical_review)

```
================================================================================
type:  numeric
range:  { 0, 1 }
unique values:  2
missing:  0/64600

 Frequency  Value  Label
 26091      0      Did not vote
 38509      1      Vote
================================================================================
```

9. vote2016_prob: Turnout in the 2016 General Election weighted by the probability of being a match. It is equal to the product between prob_match and vote2016.

```
================================================================================
type:  numeric
range:  [ 0, 1 ]
unique values:  5792
missing:  0/64600

mean:  0.5824
std.  dev:  0.4828

 percentiles:  10%  25%  50%  75%  90%
                0    0    1    1    1
--------------------------------------------------------------------------------
```

---

10. vote2014_prob: Turnout in the 2014 General Election weighted by the probability of being a match. It is equal to the product between prob_match and vote2014.

```
================================================================================
type:  numeric
range:  [ 0, 1 ]
unique values:  4366
missing:  0/64600

mean:  0.4167
std.  dev:  0.4878

 percentiles:  10%  25%  50%  75%  90%
                0    0    0    1    1
================================================================================
```

11. vote2012_prob: Turnout in the 2012 General Election weighted by the probability of being a match. It is equal to the product between prob_match and vote2012.

```
================================================================================
type:  numeric
range:  [ 0, 1 ]
unique values:  5196
missing:  0/64600

mean:  0.4879
std.  dev:  0.4916

 percentiles:  10%  25%   50%   75%  90%
                0    0   0.16    1    1
================================================================================
```

12. vote2016_clerical: Turnout in the 2016 General Election weighted by the clerical review. It is equal to the product between clerical_review and vote2016.

```
================================================================================
type:  numeric
range:  { 0, 1 }
unique values:  2
missing:  0/64600

 Frequency  Value  Label
 30478      0      Did not vote
 34122      1      Vote
================================================================================
```

13. vote2014_clerical: Turnout in the 2014 General Election weighted by the clerical review. It is equal to the product between clerical.review and vote2014.

```
================================================================================
type:  numeric
range:  { 0, 1 }
unique values:  2
missing:  0/64600

 Frequency  Value  Label
 39585      0      Did not vote
 25015      1      Vote
================================================================================
```

14. vote2012_clerical: Turnout in the 2012 General Election weighted by the clerical review. It is equal to the product between clerical.review and vote2012.

```
================================================================================
type:  numeric
range:  { 0, 1 }
unique values:  2
```

```
missing:  0/64600


 Frequency  Value  Label
 35913      0      Did not vote
 28687      1      Vote
================================================================================
```

# 3 How to use the Voter Validation Dataset

As noted above, any subsequent analysis involving the turnout variables (`vote2016, vote2014, voter2012`) must be weighted by either the probability of being a match (`prob_match`) or by the clerical review indicator (`clerical_review`). These results are, respectively, given in `vote201X_prob` and `vote201X_clerical`. In addition, sampling weights from the CCES should be used to make inferences about a target population. Below we present sample R code to calculate the validated turnout rate, adjusting for the sample design via the R package `survey`.

```r
## Example:
## Calculate Validated Turnout 2016

## Load R package for analysis of survey data
library("survey")
## Load R package to read the .tab file that stores the CCES 2016 data
library("data.table")

## Read CCES 2016
cces16study <- fread("./CCES16_Common_OUTPUT_Feb2018_VV.tab", sep = "\t")

## Read Vote Validation data
cces16vv <- read.csv("./cces2016voterval.csv")

## Merge both datasets
cces16final <- merge(cces16vv, cces16study, by = "V101")

## Add the design:
##  Post-election weight (before validation): commonweight_post
design <- svydesign(ids = ~1,
                    data = cces16final,
                    weights = cces16final$commonweight_post)

## Turnout rate 2016 weighted by:
##  the probability of being a match
```

```
svymean(cces16final$vote2016_prob, design)

## Turnout rate 2016 weighted by:
##  the clerical review indicator
svymean(cces16final$vote2016_clerical, design)
```

Listing 3: Computing the turnout rate using the validated data

# References


Cohen, W. W., P. Ravikumar and S. Fienberg. 2003. "A Comparison of String Distance Metrics for Name-Matching Tasks." In International Joint Conference on Artificial Intelligence (IJCAI) 18.

Enamorado, Ted, Benjamin Fifield and Kosuke Imai. 2017. Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records. Technical Report. Department of Politics, Princeton University.

Enamorado, Ted and Kosuke Imai. 2018. Validating Self-reported Turnout by Linking Public Opinion Surveys with Administrative Records. Technical Report. Department of Politics, Princeton University.

Fellegi, Ivan P. and Alan B. Sunter. 1969. "A Theory of Record Linkage." *Journal of the American Statistical Association* 64:1183–1210.

Winkler, William E. 1990. "String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage." Proceedings of the Section on Survey Research Methods. American Statistical Association.

Winkler, William E. 1995. *Business Survey Methods*. New York: J. Wiley Chapter Matching and Record Linkage, pp. 355–84.