User's Guide and Codebook for ANES 2016 Time Series Address-Level Supplemental Data on Voters [*]
American National Election Studies[**]
Stanford University & the University of Michigan
September 18, 2018

1. Data description

The *ANES 2016 Time Series Address-Level Supplemental Data on Voters* gives the voter turnout status for adults living at each address that was sampled for the ANES 2016 Time Series Study. Turnout is given for the general elections in 2012, 2014, and 2016. Additionally, the file includes information about the quality of the address match.

Two supplemental voter data files

This is the second file of supplemental voter data released for the ANES 2016 Time Series study. The two files serve different purposes. The first file, 2016 Time Series Vote Validation, provides supplemental turnout data for the ANES respondents as an alternative to survey self-reports of turnout. The second file, documented here, provides data about all identified voters at all ANES sampled addresses, for survey respondents and non-respondents alike.

Address matching methods

There were 10,680 addresses in the ANES sample (of which 2,880 were for the face-to-face sample and 7,800 were for the Internet sample). These addresses were merged to a nationwide voter file based on four linkage fields: ZIP code, street name, house number, and, if applicable, apartment number. Matching attempts were limited by ZIP code, so only addresses located in the same ZIP code on the voter file were compared. Matching was implemented in the R package `fastLink`[1] using methods proposed by a team at Princeton University.[2]

To make a comparison between the values of each linkage field across datasets, two levels of agreement were used for the string-valued variable (street name) based on the Jaro-Winkler distance with 0.94 as the threshold. For the remaining variables – house number, ZIP code, and apartment number – a binary comparison based on exact matching was used. A one-to-one matching was enforced based on the probability of being a match. Thus, for each residential address sampled by the ANES, the best address we could find in the voter file was kept. If many individuals in the voter file reside in such a household, all of them are matched to the same ANES sampled address.

[*] Suggested citation: ANES. 2018. *User's Guide and Codebook for ANES 2016 Time Series Address-Level Supplemental Data on Voters*. Palo Alto, CA and Ann Arbor, MI: Stanford University and the University of Michigan.
[**] Acknowledgments: The supplemental data on voters were prepared by Ted Enamorado, Benjamin Fifield, and Kosuke Imai of Princeton University in collaboration with ANES, using data from L2, Inc. ANES dissemination of the data was funded by the National Science Foundation by grants to the University of Michigan and Stanford University (grant nos. SES-1444721 and SES-1444910 respectively).
[1] Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2018. "fastLink: Fast Probabilistic Record Linkage." Available through The Comprehensive R Archive Network: https: //CRAN.R-project.org/package=fastLink
[2] Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2018. Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records. Technical Report. Department of Politics, Princeton University.

For more information about the matching methods, sample code, and vote validation for this dataset, see the *User's Guide and Codebook for the ANES 2016 Time Series Voter Validation Supplemental Data*.[3]

Address matching results

The file contains one record per individual at a matched address and one record per address where no match was found. There were 10,680 addresses in the ANES sample. There are 39,400 records on the data file, describing 287 addresses that were not matched to an address on the voter file and 39,113 adults associated with addresses on the voter file that were considered to have been matched to the ANES sample addresses.

For 2.7% of ANES sample addresses (n = 287), no address match was found on the voter file. About 45 percent of addresses had data for 1 person, 28 percent produced data for 2 people, 10 percent produced 3, and five percent produced 4. The 98th percentile had data for 20 people, the 99th percentile had data for 50 people, and the maximum number of people at one address was 425. Addresses with large numbers of person records were usually apartment buildings where the apartment number of the sampled address did not match the record on the voter file. Addresses with a large number of indicated residents, such as 10 or more, may not be limited to a single dwelling unit but may, for example, include residents from multiple dwelling units in a multi-unit building.

Linkage to other ANES 2016 data

The data can be merged with other ANES 2016 Time Series datasets for analysis of household-member voter turnout. Note that linkages between household members on this file and specific household members enumerated in the household roster are not possible. Also note that the "long" layout of this file, with multiple records per case ID, should, for most purposes, be restructured to a "wide" layout (one record per case ID, with additional records converted to variables) before merging. This can be done with the "restructure" command or "casestovars" syntax in SPSS or the "reshape" command in Stata.

2. Variables on the file (Codebook)

The data are available in three formats: Stata (.dta), SPSS (.sav), and a comma-separated values file (.csv). The file contains 13 variables as follows:

`version`. Dataset version. This variable identifies the dataset and is "ANES2016TS_AddressVoteSupp_20180918" for all cases in the initial release. A later date (in the yyyymmdd format) indicates a later re-release.

`V160001a`. 2016 Case ID. Unique identifier that can be used to merge the data with other ANES 2016 Time Series data files. Note that different versions or formats of case IDs may exist. In this version, the first case is 300001 and the last is 407800.

```
V160001a. 2016 Case ID
```

[3] Enamorado, Ted, Benjamin Fifield, Kosuke Imai. 2018. *User's Guide and Codebook for the ANES 2016 Time Series Voter Validation Supplemental Data*. Available at https://www.electionstudies.org/wp-content/uploads/2018/03/anes_timeseries_2016voteval_userguidecodebook.pdf

```
                 type:  numeric (long)
                range:  [300001,407800]
        unique values:  10,680                   missing .:  0/39,400
```

`prob_match`. Match probability. Estimated probability that an ANES address is a match with an address in the nationwide voter file conditional on their agreement pattern. Missing values indicate no match was made.

```
prob_match. Match probability
                 type:  numeric (double)
                range:  [.00006391,.99999998]
        unique values:  491                      missing .:  287/39,400
                 mean:   .976156
             std. dev:   .117536
          percentiles:        10%       25%       50%       75%       90%
                          .993739   .998781   .999837   .999975   .999993
```

`hn_agreement`. House number agreement level. Equals: A if the pair of observations agree on house number, D if they disagree, and NA if the comparison involved a missing value.

```
hn_agreement. House number agreement level
                 type:  string (str2)
        unique values:  3                        missing "":  0/39,400
          tabulation:  Freq.  Value
                      36,515  "A"
                       2,516  "D"
                         369  "NA"
```

`sn_agreement`. Street name agreement level. Equals: A if the pair of observations agree on street name, D if they disagree, and NA if the comparison involved a missing value.

```
sn_agreement. Street name agreement level
                 type:  string (str2)
        unique values:  3                        missing "":  0/39,400
          tabulation:  Freq.  Value
                      38,724  "A"
                         389  "D"
                         287  "NA"
```

`zc_agreement`. ZIP code agreement level. Equals: A if the pair of observations agree on ZIP code and NA if the comparison involved a missing value.

```
zc_agreement. ZIP code agreement level
                 type:  string (str2)
        unique values:  2                        missing "":  0/39,400
          tabulation:  Freq.  Value
                      39,113  "A"
                         287  "NA"
```

`an_agreement`. Apartment number agreement level. Equals: A if the pair of observations agree on apartment number, D if they disagree, and NA if the comparison involved a missing value.

```
an_agreement. Apartment number agreement level
                 type:  string (str2)
        unique values:  3                        missing "":  0/39,400
          tabulation:  Freq.  Value
```

```
 3,039  "A"
18,277  "D"
18,084  "NA"
```

`vote2016`. Voter turnout in 2016, unweighted. It equals 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 the probability of being a match. Missing values indicate no match was made.

```
vote2016. Voter turnout in 2016
                  type:  numeric
                 label:  vote2016
                 range:  [0,1]
         unique values:  2                        missing .:  287/39,400
           tabulation:  Freq.   Numeric  Label
                       13,098         0  Did not vote
                       26,015         1  Voted
                          287         .
```

`vote2014`. Voter turnout in 2014, unweighted. It equals 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 the probability of being a match. Missing values indicate no match was made.

```
vote2014. Voter turnout in 2014
                  type:  numeric
                 label:  vote2016
                 range:  [0,1]
         unique values:  2                        missing .:  287/39,400
           tabulation:  Freq.   Numeric  Label
                       25,757         0  Did not vote
                       13,356         1  Voted
                          287         .
```

`vote2012`. Voter turnout in 2012, unweighted. It equals 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 the probability of being a match. Missing values indicate no match was made.

```
vote2012. Voter turnout in 2012
                  type:  numeric
                 label:  vote2016
                 range:  [0,1]
         unique values:  2                        missing .:  287/39,400
           tabulation:  Freq.   Numeric  Label
                       18,392         0  Did not vote
                       20,721         1  Voted
                          287         .
```

`vote2016_prob`. Probabilistic voter turnout in 2016. Turnout in the 2016 general election weighted by the probability of being a match. It is equal to the product of vote2016 and prob_match.

```
vote2016_prob. Probabilistic voter turnout in 2016
                  type:  numeric
                 range:  [0,.99999998]
         unique values:  466                      missing .:  287/39,400
```

```
               mean:   .648933
           std. dev:   .470608
        percentiles:        10%       25%       50%       75%       90%
                              0         0     .9993   .999932   .999986
```

`vote2014_prob`. Probabilistic voter turnout in 2014. Turnout in the 2014 general election weighted by the probability of being a match. It is equal to the product of vote2014 and prob_match.

```
vote2014_prob. Probabilistic voter turnout in 2014
                  type:  numeric
                 range:  [0,.99999998]
         unique values:  425                     missing .:  287/39,400
                  mean:   .332205
              std. dev:   .467193
           percentiles:        10%       25%       50%       75%       90%
                                 0         0         0   .999364   .999968
```

`vote2012_prob`. Probabilistic voter turnout in 2012. Turnout in the 2012 general election weighted by the probability of being a match. It is equal to the product of vote2012 and prob_match.

```
vote2012_prob. Probabilistic voter turnout in 2012
                  type:  numeric
                 range:  [0,.99999998]
         unique values:  453                     missing .:  287/39,400
                  mean:   .516399
              std. dev:   .494387
           percentiles:        10%       25%       50%       75%       90%
                                 0         0   .981542   .999876   .999985
```