Validating Self-reported Turnout by Linking Public Opinion Surveys with Administrative Records

Ted Enamorado    Kosuke Imai
Princeton University
Seminar at the Center for the Study of Democratic Politics
Princeton University
March 8, 2018
Bias of Self-reported Turnout

- Where does this gap come from?
- Nonresponse, Misreporting, Mobilization
Turnout Validation Controversy

- The Help America Vote Act of 2002 \(\rightarrow\) Development of systematically collected and regularly updated nationwide voter registration records

- Ansolabehere and Hersh (2012, *Political Analysis*):
  “electronic validation of survey responses with commercial records provides a far more accurate picture of the American electorate than survey responses alone.”

  “Matching errors ... drive down “validated” turnout estimates. As a result, ... the apparent accuracy [of validated turnout estimates] is likely an illusion.”

- Challenge: Find several thousand survey respondents in 180 million registered voters (less than 0.001%) \(\rightarrow\) finding needles in a haystack

- Problems: *false matches and false non-matches*
Methodological Motivation

- In any given project, social scientists often rely on multiple data sets.
- Cutting-edge empirical research often merges large-scale administrative records with other types of data.
- We can easily merge data sets if there is a common unique identifier.  
  \[\Rightarrow\] e.g. Use the `merge` function in R or Stata.
- How should we merge data sets if no unique identifier exists?  
  \[\Rightarrow\] must use variables: names, birthdays, addresses, etc.
- Variables often have measurement error and missing values.  
  \[\Rightarrow\] cannot use exact matching.
- What if we have millions of records?  
  \[\Rightarrow\] cannot merge “by hand.”
- Merging data sets is an uncertain process.  
  \[\Rightarrow\] quantify uncertainty and error rates.
- **Solution:** Probabilistic Model.
Overview of the Talk

1. Turnout validation:
   - 2016 American National Election Study (ANES)
   - 2016 Cooperative Congressional Election Study (CCES)

2. Probabilistic method of record linkage and fastLink (with Ben Fifield)

3. Simulation study to compare fastLink with deterministic methods
   - fastLink effectively handles missing data and measurement error

4. Empirical findings:
   - fastLink recovers the actual turnout
   - Clerical review helps with the ANES but not with the CCES
   - Bias of self-reported turnout appears to be largely driven by misreporting
   - fastLink performs at least as well as a state-of-art proprietary method
The 2016 US Presidential Election

- Donald Trump’s surprising victory \( \Rightarrow \) failure of polling
- Non-response and social desirability biases as possible explanations

- Two validation exercises:
  1. The 2016 American National Election Study (ANES)
  2. The 2016 Cooperative Congressional Election Study (CCES)

- We merge the survey data with a nationwide voter file

- The voter file was obtained in July 2017 from L2, Inc.
  - total of 182 million records
  - 8.6 million “inactive” voters
ANES Sampling Design

Sampling Frame

- Universe of Addresses 48 contiguous states + DC
- Universe of Addresses 50 states + DC

Sample

- 2,880 Addresses
  - 1,946 Maximum eligible
    - 1,454 Eligible
- Multi-stage stratified cluster sampling

Pre-election Survey

- 1,181 respondents
  - Unit non-response 16%

Post-election Survey

- 1,059 respondents
  - Panel attrition 10%

Internet Sample

- 7,800 Addresses
  - 3,569 Eligible
  - Simple random sampling, excluding ‘drop point’ addresses
  - 6,956 Maximum eligible

- 3,090 respondents
  - Unit non-response 13%

- 2,590 respondents
  - Panel attrition 16%

Face-to-Face Sample
CCES Sampling Design

Target Sample

Random sample (within strata) of U.S. citizens profiles drawn from the 2010 and 2012 American Community Surveys (50 states + DC)

Sample

Sample 106,382 opt-in respondents

Target sample is matched to a pool of 95,826 respondents after quality control

Pre-election Survey

Complete interviews: 87,389

Non-informative responses 9%

Common content: 64,600 respondents

Panel attrition 18%

Post-election Survey

52,899 respondents
### Bias of Self-reported Turnout and Registration Rates

<table>
<thead>
<tr>
<th></th>
<th>ANES</th>
<th>CCES</th>
<th>Election project</th>
<th>Voter files all</th>
<th>Voter files active</th>
<th>CPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turnout rate</td>
<td>75.96</td>
<td>83.79</td>
<td>58.83</td>
<td>57.55</td>
<td></td>
<td>61.38</td>
</tr>
<tr>
<td></td>
<td>(0.92)</td>
<td>(0.27)</td>
<td></td>
<td></td>
<td></td>
<td>(1.49)</td>
</tr>
<tr>
<td>Registration rate</td>
<td>89.18</td>
<td>91.93</td>
<td>80.37</td>
<td>76.57</td>
<td></td>
<td>70.34</td>
</tr>
<tr>
<td></td>
<td>(0.71)</td>
<td>(0.21)</td>
<td></td>
<td></td>
<td></td>
<td>(1.40)</td>
</tr>
</tbody>
</table>

| Pop. size (millions) | 224.10 | 224.10 | 232.40 | 227.60 | 227.60 | 224.10 |

- Based on the ANES sampling and CCES pre-validation weights
- **Target population**
  - ANES (face-to-face): US citizens of voting age in 48 states + DC
  - ANES (internet) / CCES: US citizens of voting age in 50 states + DC
  - Election project: cannot adjust for overseas population
  - Voter file: the deceased and out-of-state movers (after the election) are removed
Election Project vs. Voter File

Correlation = 0.98
We merge with the nationwide voter file using name, age, gender, and address:

1. 4,271 ANES respondents
2. 64,600 CCES respondents

Standardization:

1. Name: first, middle, and last name
   - ANES: Missing (1.5%), Use of initials (0%), Complete (0.4%)
   - CCES: Missing (2.7%), Use of initials (5.9%), Complete (91.4%)

2. Address: house number, street name, zip code, and apartment number
   - ANES: Complete (100%)
   - CCES: Missing (11.6%), P.O. Box (2.6%), Complete (85.9%)

Blocking:

- Direct comparison \(\sim 18\) trillion pairs
- Blocking by gender and state \(\sim 102\) blocks
  1. ANES: from 48k (HI/Female) to 108 million pairs (CA/Female)
  2. CCES: from 3 million (WY/Male) to 25 billion pairs (CA/Male)

- Apply fastLink within each block
Probabilistic Model of Record Linkage

- Many social scientists use deterministic methods:
  - match “similar” observations (e.g., Ansolabehere and Hersh, 2016; Berent, Krosnick, and Lupia, 2016)
  - proprietary methods (e.g., Catalist, YouGov)

- Problems:
  1. not robust to measurement error and missing data
  2. no principled way of deciding how similar is similar enough
  3. lack of transparency

- Probabilistic model of record linkage:
  - originally proposed by Fellegi and Sunter (1969, *JASA*)
  - enables the control of error rates

- Problems:
  1. current implementations do not scale
  2. missing data treated in ad-hoc ways
  3. does not incorporate auxiliary information
The Fellegi-Sunter Model

- Two data sets: $A$ and $B$ with $N_A$ and $N_B$ observations
- $K$ variables in common
- We need to compare all $N_A \times N_B$ pairs
- Agreement vector for a pair $(i, j)$: $\gamma(i, j)$

\[
\gamma_k(i, j) = \begin{cases} 
0 & \text{different} \\
1 & \text{similar} \\
L_k - 2 & \text{identical} \\
\end{cases}
\]

- Latent variable:

\[
M_{i,j} = \begin{cases} 
0 & \text{non-match} \\
1 & \text{match} \\
\end{cases}
\]

- Missingness indicator: $\delta_k(i, j) = 1$ if $\gamma_k(i, j)$ is missing
How to Construct Agreement Patterns

- Jaro-Winkler distance with default thresholds for string variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td>Middle</td>
</tr>
<tr>
<td>Data set $A$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>James</td>
</tr>
<tr>
<td>2</td>
<td>John</td>
</tr>
<tr>
<td>Data set $B$</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Michael</td>
</tr>
<tr>
<td>2</td>
<td>James</td>
</tr>
</tbody>
</table>

Agreement patterns

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$A.1 - B.1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$A.1 - B.2$</td>
<td>2</td>
<td>NA</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$A.2 - B.1$</td>
<td>0</td>
<td>NA</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$A.2 - B.2$</td>
<td>0</td>
<td>NA</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Independence assumptions for computational efficiency:

1. Independence across pairs
2. Independence across variables: $\gamma_k(i, j) \perp \perp \gamma_{k'}(i, j) \mid M_{ij}$
3. Missing at random: $\delta_k(i, j) \perp \perp \gamma_k(i, j) \mid M_{ij}$

Nonparametric mixture model:

$$\frac{N_A N_B}{\prod_{i=1}^{N_A} \prod_{j=1}^{N_B}} \left\{ \sum_{m=0}^{1} \lambda^m (1 - \lambda)^{1-m} \prod_{k=1}^{K} \left( \prod_{\ell=0}^{L_k-1} \pi_{k \ell m}^{1\{\gamma_k(i, j) = \ell\}} \right)^{1-\delta_k(i, j)} \right\}$$

where $\lambda = P(M_{ij} = 1)$ is the proportion of true matches and $\pi_{k \ell m} = Pr(\gamma_k(i, j) = \ell \mid M_{ij} = m)$

Fast implementation of the EM algorithm (R package fastLink)

EM algorithm produces the posterior matching probability $\xi_{ij}$

Deduping to enforce one-to-one matching

1. Choose the pairs with $\xi_{ij} > c$ for a threshold $c$
2. Use Jaro’s linear sum assignment algorithm to choose the best matches
Simulation Studies

- 2006 voter files from California (female only; 8 million records)
- Validation data: records with no missing data (340k records)
- Linkage fields: first name, middle name, last name, date of birth, address (house number and street name), and zip code
- 2 scenarios:
  1. Unequal size: 1:100, 10:100, and 50:100, larger data 100k records
  2. Equal size (100k records each): 20%, 50%, and 80% matched
- 3 missing data mechanisms:
  1. Missing completely at random (MCAR)
  2. Missing at random (MAR)
  3. Missing not at random (MNAR)
- 3 levels of missingness: 5%, 10%, 15%
- Noise is added to first name, last name, and address
- Results below are with 10% missingness and no noise
Error Rates and Estimation Error for Turnout

80% Overlap

50% Overlap

20% Overlap

False Negative Rate

Absolute Estimation Error (percentage point)

Enamorado and Imai (Princeton) Validating Self-reported Turnout CSDP (March 8, 2018) 17 / 28
Runtime Comparisons

Equal size

Dataset size (thousands)

Time elapsed (minutes)

RecordLinkage (Python)

RecordLinkage (R)

fastLink (R)

Unequal Size

Largest dataset size (thousands)

Time elapsed (minutes)

RecordLinkage (Python)

RecordLinkage (R)

fastLink (R)

- No blocking, single core (parallelization possible with fastLink)
Merge Procedure and Results

- Use of three agreement levels for string variables and age
- Merge process:
  1. within-block merge
  2. remove within-state matches (posterior match prob. > 0.75)
  3. across-state merge (exact match on gender, names, age)
  4. clerical review (for both matches and non-matches)

- Our analysis uses posterior match probability as well as ANES and CCES (pre-validation) sampling weights
### Match Rate as an Estimate of Registration Rate

<table>
<thead>
<tr>
<th></th>
<th>Pre-election</th>
<th>Post-election</th>
<th>Registration rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fastLink</td>
<td>clerical</td>
<td>fastLink</td>
</tr>
<tr>
<td>ANES</td>
<td>76.54 (0.63)</td>
<td>68.79 (0.71)</td>
<td>77.15 (0.67)</td>
</tr>
<tr>
<td>CCES</td>
<td>66.60 (0.18)</td>
<td>58.59 (0.19)</td>
<td>70.52 (0.19)</td>
</tr>
</tbody>
</table>

- Registration rate is difficult to compute:
  - only some states classify voters as “active” or “inactive”
  - definition differs by states
- Clerical review
  - appears to work for the ANES
  - may have introduced false negatives for the CCES
### Validated Turnout Rates

<table>
<thead>
<tr>
<th></th>
<th>Pre-election</th>
<th>Post-election</th>
<th>Actual turnout</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fastLink</td>
<td>clerical review</td>
<td>fastLink</td>
</tr>
<tr>
<td>ANES</td>
<td>63.59 (0.91)</td>
<td>58.09 (0.93)</td>
<td>64.97 (0.96)</td>
</tr>
<tr>
<td>CCES</td>
<td>54.11 (0.31)</td>
<td>48.50 (0.31)</td>
<td>55.67 (0.37)</td>
</tr>
</tbody>
</table>

- fastLink plus clerical review works well for the ANES
- fastLink alone works better for the CCES
## Validated Turnout by Response Category

<table>
<thead>
<tr>
<th></th>
<th>Registered</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not registered</td>
<td>Did not Vote</td>
<td>Voted</td>
<td>Attrition</td>
</tr>
<tr>
<td><strong>ANES</strong></td>
<td>8.11 (1.58)</td>
<td>14.45 (1.74)</td>
<td>81.74 (0.86)</td>
<td>55.66 (2.41)</td>
</tr>
<tr>
<td>fastLink</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clerical review</td>
<td>0.90 (0.78)</td>
<td>5.97 (1.21)</td>
<td>77.44 (0.99)</td>
<td>48.27 (2.41)</td>
</tr>
<tr>
<td><strong>CCES</strong></td>
<td>16.37 (0.84)</td>
<td>10.15 (0.73)</td>
<td>73.05 (0.28)</td>
<td>24.02 (0.60)</td>
</tr>
<tr>
<td>fastLink</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clerical review</td>
<td>8.04 (0.73)</td>
<td>4.67 (0.59)</td>
<td>68.66 (0.30)</td>
<td>16.44 (0.51)</td>
</tr>
</tbody>
</table>

- Over-reporting is important: many are in the “Voted” category
- Attrition is a problem for the CCES, but not for the ANES
Berent, Krosnick, and Lupia (2016) argue that voters don’t misreport:

1. Poor quality of voter files and difficulty of merging
2. Failure to match survey respondents who actually voted
3. Results in a lower validated turnout rate

As evidence, BKL show:

1. the match rate is lower than the registration rate
2. matched voters do not lie

Our match rate is lower than the registration rate based on voter file

However, we find that matched non-voters do lie at a high rate:

1. matched respondents who voted:
   - ANES: 95.68% (s.e.=0.50, N=3,436)
   - CCES: 92.70% (s.e.=0.36, N=33,329)
2. matched respondents who did not vote:
   - ANES: 33.66% (s.e.=3.01, N=378)
   - CCES: 43.49% (s.e.=1.50, N=3,901)
Who Misreports?

### Education
- **ANES**
  - High school or less
  - Some college
  - College
  - Post-graduate
- **CCES**
  - High school or less
  - Some college
  - College
  - Post-graduate

### Income (in thousands)
- **ANES**
  - Less than 30
  - Between 30 and 60
  - Between 60 and 100
  - More than 100
- **CCES**
  - Less than 27.5
  - Between 27.5 and 60
  - Between 60 and 100
  - More than 100

### Interest in Politics
- **ANES**
  - Not at all
  - Not much
  - Some
  - A lot
- **CCES**
  - Not at all
  - Not much
  - Some
  - A lot

### Race
- **ANES**
  - Whites
  - Hispanics
  - Others
- **CCES**
  - Whites
  - Hispanics
  - Others

Enamorado and Imai (Princeton)

Validating Self-reported Turnout

CSDP (March 8, 2018)
Comparison with CCES Turnout Validation

- Benchmark: 58.83 (election project) and 57.55 (voter file)

<table>
<thead>
<tr>
<th>Validated Turnout</th>
<th>Common matches</th>
<th>CCES only</th>
<th>fastLink only</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2</td>
<td>70.34 (0.35)</td>
<td>8.63 (0.21)</td>
<td>23.16 (0.43)</td>
<td>54.11 (0.31)</td>
</tr>
<tr>
<td>CCES</td>
<td>68.48 (0.35)</td>
<td>10.14 (0.23)</td>
<td>0.00</td>
<td>52.85 (0.34)</td>
</tr>
<tr>
<td>Number of respondents</td>
<td>34,344</td>
<td>8,773</td>
<td>6,678</td>
<td>64,600</td>
</tr>
</tbody>
</table>
State-level Comparison

**Proprietary Method**
- Bias = 4.18
- RMSE = 7.11
- Correlation = 0.51

**fastLink**
- Bias = 4.32
- RMSE = 7.32
- Correlation = 0.60
Predicting Match Type

Race

Proprietary

fastLink

Interest in Politics

Proprietary

fastLink

Enamorado and Imai (Princeton)  Validating Self-reported Turnout  CSDP (March 8, 2018)  27 / 28
Concluding Remarks

- Merging data sets is critical part of social science research
  - merging can be difficult when no unique identifier exists
  - large data sets make merging even more challenging
  - yet merging can be consequential

- We offer a fast, principled, and scalable probabilistic merging method

- Open-source software fastLink available at CRAN

- Application: controversy regarding bias in self-reported turnout
  - Previous turnout validations relied upon proprietary algorithms
  - We merge ANES/CCES with a nationwide voter file using fastLink
  - fastLink yields high-quality matches and recovers actual turnout rate
  - Bias appears to be driven by misreporting rather than nonresponse
  - Probabilistic merge is robust to missing and invalid entries
  - Clerical review may introduce false negatives for messy data
  - fastLink performs as well as a state-of-art proprietary method