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

Ted Enamorado    Kosuke Imai
Princeton        Harvard
Seminar at Yokohama City University
July 9, 2018
Where does this gap come from?

- Nonresponse, Misreporting, Mobilization
The Help America Vote Act of 2002 ➔ 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.”

Berent, Krosnick, and Lupia (2016, *Public Opinion Quarterly*): “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%) ➔ 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.
  - e.g. Use the `merge` function in \texttt{R} or Stata.
- How should we merge data sets if no unique identifier exists?
  - must use variables: names, birthdays, addresses, etc.
- Variables often have measurement error and missing values.
  - cannot use exact matching.
- What if we have millions of records?
  - cannot merge “by hand”.
- Merging data sets is an uncertain process.
  - 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 (with Ben Fifield)
   - Details in “Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records”
   - Open-source software package: fastLink

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 is 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
  - Sampling Frame
  - Multi-stage stratified cluster sampling

Sample
- 2,880 Addresses
  - 1,946 Maximum eligible
    - 1,454 Eligible
  - Unit non-response 16%
- 7,800 Addresses
  - 3,569 Eligible
    - 6,956 Maximum eligible
  - Unit non-response 13%

Pre-election Survey
- 1,181 respondents
  - Panel attrition 10%
- 3,090 respondents
  - Panel attrition 16%

Post-election Survey
- 1,059 respondents
  - Face-to-Face Sample
- 2,590 respondents
  - Internet 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

Post-election Survey
- Panel attrition 18%
  - 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></td>
<td>80.37</td>
<td>76.57</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>
<tr>
<td>Pop. size (millions)</td>
<td>224.10</td>
<td>224.10</td>
<td>232.40</td>
<td>227.60</td>
<td>227.60</td>
<td>224.10</td>
</tr>
</tbody>
</table>

- 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
Preprocessing

- 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 (1.6%), Use of initials (0.4%), Complete (98.1%)
     - 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>
<th>Data set $A$</th>
<th>Data set $B$</th>
<th>Agreement patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>$A.1 - B.1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>James V</td>
<td>Michael F</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Smith</td>
<td>Martinez</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>780 House</td>
<td>4 House</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Devereux St.</td>
<td>16th St.</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>780 Street</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Enamorado and Imai (Princeton/Harvard)  Validating Self-reported Turnout  Yokohama (July 9, 2018)
• Independence assumptions for computational efficiency:
  1. Independence across pairs
  2. Independence across variables: \( \gamma_k(i,j) \perp \perp \gamma_{k'}(i,j) \mid M_{ij} \)
    \(\rightsquigarrow\) two ways to relax this assumption
  3. Missing at random: \( \delta_k(i,j) \perp \perp \gamma_k(i,j) \mid M_{ij} \)

• Mixture model:

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

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

• 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:
  - Unequal size: 1:100, 10:100, and 50:100, larger data 100k records
  - Equal size (100k records each): 20%, 50%, and 80% matched
- 3 missing data mechanisms:
  - Missing completely at random (MCAR)
  - Missing at random (MAR)
  - 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

Enamorado and Imai (Princeton/Harvard) Validating Self-reported Turnout Yokohama (July 9, 2018) 17 / 28
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</td>
<td>68.79</td>
<td>77.15</td>
</tr>
<tr>
<td></td>
<td>(0.63)</td>
<td>(0.71)</td>
<td>(0.67)</td>
</tr>
<tr>
<td></td>
<td>80.37</td>
<td>70.34</td>
<td></td>
</tr>
<tr>
<td>CCES</td>
<td>66.60</td>
<td>58.59</td>
<td>70.52</td>
</tr>
<tr>
<td></td>
<td>(0.18)</td>
<td>(0.19)</td>
<td>(0.19)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></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></th>
<th>Post-election</th>
<th></th>
<th>Actual turnout</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fastLink</td>
<td>clerical</td>
<td>review</td>
<td>fastLink</td>
<td>clerical</td>
<td>review</td>
</tr>
<tr>
<td>ANES</td>
<td>63.59</td>
<td>58.09</td>
<td>(0.91)</td>
<td>64.97</td>
<td>59.78</td>
<td>(1.00)</td>
</tr>
<tr>
<td>CCES</td>
<td>54.11</td>
<td>48.50</td>
<td>(0.31)</td>
<td>55.67</td>
<td>50.25</td>
<td>(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>ANES</th>
<th>CCES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Registered</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not registered</td>
<td>Did not Vote</td>
</tr>
<tr>
<td>fastLink</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8.11 (1.58)</td>
<td>14.45 (1.74)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clerical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>review</td>
<td>0.90 (0.78)</td>
<td>5.97 (1.21)</td>
</tr>
<tr>
<td>fastLink</td>
<td>16.37 (0.84)</td>
<td>10.15 (0.73)</td>
</tr>
<tr>
<td>Clerical</td>
<td>8.04 (0.73)</td>
<td>4.67 (0.59)</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
Do Voters Misreport Turnout?

- 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

<table>
<thead>
<tr>
<th>ANES</th>
<th>CCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>High school or less</td>
<td>High school or less</td>
</tr>
<tr>
<td>Some college</td>
<td>Some college</td>
</tr>
<tr>
<td>College</td>
<td>College</td>
</tr>
<tr>
<td>Post-graduate</td>
<td>Post-graduate</td>
</tr>
</tbody>
</table>

### Income (in thousands)

<table>
<thead>
<tr>
<th>ANES</th>
<th>CCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 30</td>
<td>Less than 30</td>
</tr>
<tr>
<td>Between 30 and 60</td>
<td>Between 30 and 60</td>
</tr>
<tr>
<td>Between 60 and 100</td>
<td>Between 60 and 100</td>
</tr>
<tr>
<td>More than 100</td>
<td>More than 100</td>
</tr>
</tbody>
</table>

### Interest in Politics

<table>
<thead>
<tr>
<th>ANES</th>
<th>CCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not at all</td>
<td>Not at all</td>
</tr>
<tr>
<td>Not much</td>
<td>Not much</td>
</tr>
<tr>
<td>Some</td>
<td>Some</td>
</tr>
<tr>
<td>A lot</td>
<td>A lot</td>
</tr>
</tbody>
</table>

### Race

<table>
<thead>
<tr>
<th>ANES</th>
<th>CCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Others</td>
<td>Others</td>
</tr>
<tr>
<td>Whites</td>
<td>Whites</td>
</tr>
<tr>
<td>Blacks</td>
<td>Blacks</td>
</tr>
<tr>
<td>Hispanics</td>
<td>Hispanics</td>
</tr>
</tbody>
</table>
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 (0.34)</td>
<td>52.85 (0.34)</td>
</tr>
</tbody>
</table>

Number of respondents: 34,344 (L2) 8,773 (CCES) 6,678 (fastLink) 64,600 (Overall)
State-level Comparison

Proprietary Method

- Bias = 4.18
- RMSE = 7.11
- Correlation = 0.51

fastLink

- Bias = 4.32
- RMSE = 7.32
- Correlation = 0.60

Turnout rate based on the voter file (%)

Validated turnout rate (%)

Correlation = 0.51
RMSE = 7.11
Bias = 4.18

Enamorado and Imai (Princeton/Harvard) Validating Self-reported Turnout Yokohama (July 9, 2018)
Predicting Match Type

Race

Proprietary

fastLink

Others

Hispanics

Blacks

Whites

Common

Non-matches

Common Matches

Others

Hispanics

Blacks

Whites

Interest in Politics

Proprietary

fastLink

Not at all

Not much

Some

A lot

Common

Non-matches

Common Matches

Not at all

Not much

Some

A lot

A lot

Some

Not much

Not at all
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
  - Our validated turnout variables available as part of ANES and CCES