Behavioral Applications of Voter Files

Eitan Hersh

Tufts University
Department of Political Science

June 14, 2018
Overview

1. Voter files as is
2. Linked geographically
3. Linked individually
4. Some lessons
Overview

1. Voter files as is
Overview

1. Voter files as is
2. Linked geographically
Overview

1. Voter files as is
2. Linked geographically
3. Linked individually
Overview

1. Voter files as is
2. Linked geographically
3. Linked individually
4. Some lessons
Voters files, as is
Voters files, as is

Voters files, as is

Voters files, as is


**Basic Advantages:** Sample size, official turnout, demographics, location data, daily updates
Voter files, linked to geographic data
Voter files, linked to geographic data

Voter files, linked to geographic data

Voter files, linked to geographic data

Voter files, linked to geographic data


![Graphs showing the relationship between percent Republican and Census Block Group Income, in $10K for different Black Distributions in the North and South regions.](image-url)
Voter files, linked to geographic data


Advantages: nested geographies; households and neighborhoods, custom defined; vendors do the hard work.
Voter files, individuals linked to external data
Voter files, individuals linked to external data

1. Completely outsource to vendor
Voter files, individuals linked to external data

1. Completely outsource to vendor
2. Completely in-house linkage
Voter files, individuals linked to external data

1. Completely outsource to vendor
2. Completely in-house linkage
3. Hybrid outsourced/custom linkage
Voter files, individuals linked to external data

1. Completely outsource to vendor
   - Example 4: Validation of CCES with Catalist voter file (Ansolabehere and Hersh, "Validation," *Political Analysis*, 2012.)
   - Example 5: Survey of 2016 max-out donors (Hersh and Schaffner, in draft)
2. Completely in-house linkage
Voter files, individuals linked to external data

1. Completely outsource to vendor
Voter files, individuals linked to external data

1. Completely outsource to vendor
   - Example 4: Validation of CCES with Catalist voter file
     (Ansolabehere and Hersh, “Validation,” *Political Analysis*, 2012.)
Voter files, individuals linked to external data

1. Completely outsource to vendor
   - Example 4: Validation of CCES with Catalist voter file (Ansolabehere and Hersh, “Validation,” Political Analysis, 2012.)
   - Example 5: Survey of 2016 max-out donors (Hersh and Schaffner, in draft)
Voter files, individuals linked to external data

1. Completely outsource to vendor
   - Example 4: Validation of CCES with Catalist voter file (Ansolabehere and Hersh, “Validation,” *Political Analysis*, 2012.)
   - Example 5: Survey of 2016 max-out donors (Hersh and Schaffner, in draft)

2. Completely in-house linkage

Research question: Did 9/11 change the politics of victims’ families and neighbors?

Research question: Did 9/11 change the politics of victims’ families and neighbors?

Data

- Obituary data from NYT
- Pre-9/11 NY voter file (L2)
- Year 2013 NY voter file (Catalist)
In-House Individual Link

Victims' Families Relative to Control

September 11th Attack

-0.04 -0.02 0 0.02 0.04 0.06 0.08
Diff. in Turnout Rate, Family-Control
Year of General Election

Victims' Neighbors Relative to Control

September 11th Attack

-0.01 0 0.01 0.02 0.03 0.04
Difference in Donation Rate
Year

Advantages: long time frame, precise geography, pseudo control group, no misreporting, highly customized
Advantages: long time frame, precise geography, pseudo control group, no misreporting, highly customized
In-House Individual Link

Example 7: Stephen Ansolabehere and Eitan Hersh, “ADGN: An Algorithm for Record Linkage Using Address, Date of Birth, Gender, and Name,” *Statistics and Public Policy*, Forthcoming.
Example 7: Stephen Ansolabehere and Eitan Hersh, “ADGN: An Algorithm for Record Linkage Using Address, Date of Birth, Gender, and Name,” *Statistics and Public Policy*, Forthcoming.

Research question: Are protected racial groups less likely to possess valid photo identification in Texas?
Example 7: Stephen Ansolabehere and Eitan Hersh, “ADGN: An Algorithm for Record Linkage Using Address, Date of Birth, Gender, and Name,” *Statistics and Public Policy*, Forthcoming.

Research question: Are protected racial groups less likely to possess valid photo identification in Texas?

Data

- Texas voter file
- 10 ID databases (e.g. drivers, passport holders)
In-House Individual Link

A=Zip5+Street Num; G=Gender; D=MDY of Birth; N=First+Last Name
Y=Year of Birth; Z=Zip5; F=First Name; m=Middle Initial; L=Last Name
### Table: Number and Percent of instances of NO-MATCH and MATCH by Racial Group, Using Catalist Racial Classification

<table>
<thead>
<tr>
<th>Race</th>
<th>NO-MATCH</th>
<th>MATCH</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anglo</td>
<td>296,156(3.6%)</td>
<td>7,949,860(96.4%)</td>
<td>8,246,016</td>
</tr>
<tr>
<td>Black</td>
<td>127,908(7.5%)</td>
<td>1,569,861(92.5%)</td>
<td>1,707,769</td>
</tr>
<tr>
<td>Hispanic</td>
<td>174,715(5.7%)</td>
<td>2,867,782(94.2%)</td>
<td>3,042,497</td>
</tr>
<tr>
<td>Other</td>
<td>9,691(2.0%)</td>
<td>481,621(98.0%)</td>
<td>491,312</td>
</tr>
<tr>
<td>All</td>
<td>608,470(4.5%)</td>
<td>12,879,124(95.5%)</td>
<td>13,487,594</td>
</tr>
</tbody>
</table>
Table: Number and Percent of instances of NO-MATCH and MATCH by Racial Group, Using Catalist Racial Classification

<table>
<thead>
<tr>
<th>Race</th>
<th>NO-MATCH</th>
<th>MATCH</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anglo</td>
<td>296,156</td>
<td>7,949,860</td>
<td>8,246,016</td>
</tr>
<tr>
<td></td>
<td>(3.6%)</td>
<td>(96.4%)</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>127,908</td>
<td>1,569,861</td>
<td>1,707,769</td>
</tr>
<tr>
<td></td>
<td>(7.5%)</td>
<td>(92.5%)</td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>174,715</td>
<td>2,867,782</td>
<td>3,042,497</td>
</tr>
<tr>
<td></td>
<td>(5.7%)</td>
<td>(94.2%)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>9,691</td>
<td>481,621</td>
<td>491,312</td>
</tr>
<tr>
<td></td>
<td>(2.0%)</td>
<td>(98.0%)</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>608,470</td>
<td>12,879,124</td>
<td>13,487,594</td>
</tr>
<tr>
<td></td>
<td>(4.5%)</td>
<td>(95.5%)</td>
<td></td>
</tr>
</tbody>
</table>

Advantages: direct answer to question, customized algorithm for legal purpose
Voter files, individuals linked to external data

1. Completely outsource to vendor
2. Completely in-house linkage
Voter files, individuals linked to external data

1. Completely outsource to vendor
2. Completely in-house linkage
3. Hybrid outsourced/custom linkage

Research question: Do Democratic and Republican physicians treat patients differently?

Research question: Do Democratic and Republican physicians treat patients differently?

Data

- National Provider Identification (NPI) File, CMS
- Voter file in 29 party-registration states
- Survey of primary care physicians
Targeting the Population

Step 1 The National Provider Identification File
1. 560,896 U.S.-based physicians
2. 149,936 in primary care specialties (internal, family, adult, general practice)
3. 85,722 in 29 party-registration states
4. 42,861 in 50% simple random sample

Step 2 Match to Voter File (Catalist)
1. 161,553 plausible matches sent by Catalist
2. 18,430 records with unique matches, plus 5,820 confident matches (57% match rate)
3. Of 24,250 confident matches, 20,296 mailable addresses

Step 3 Target sample
Of 20,296, 13,678 when restricted to registered Ds and Rs
Oversample doctors in mixed-partisan practices
1. Classify practices by size and partisan composition
2. Draw 100% physicians from mid-size bipartisan stratum (754) and 6% sample of all other strata (775)
3. Survey 1,529 physicians
Targeting the Population

Step 1  The National Provider Identification File

- 560,896 U.S.-based physicians
- 149,936 in primary care specialties (internal, family, adult, general practice)
- 85,722 in 29 party-registration states
- 42,861 in 50% simple random sample

Step 2  Match to Voter File (Catalist)

- 161,553 plausible matches sent by Catalist
- 18,430 records with unique matches, plus 5,820 confident matches (57% match rate)
- Of 24,250 confident matches, 20,296 mailable addresses

Step 3  Target sample

- Of 20,296, 13,678 when restricted to registered Ds and Rs
- Oversample doctors in mixed-partisan practices
  - Classify practices by size and partisan composition
  - Draw 100% physicians from mid-size bipartisan stratum (754) and 6% sample of all other strata (775)
  - Survey 1,529 physicians
Targeting the Population

Step 1  The National Provider Identification File

1. 560,896 U.S.-based physicians
Targeting the Population

Step 1  The National Provider Identification File

1. 560,896 U.S.-based physicians
2. 149,936 in primary care specialties (internal, family, adult, general practice)
Targeting the Population

Step 1  The National Provider Identification File

1. 560,896 U.S.-based physicians
2. 149,936 in primary care specialties (internal, family, adult, general practice)
3. 85,722 in 29 party-registration states
Targeting the Population

Step 1  The National Provider Identification File

1. 560,896 U.S.-based physicians
2. 149,936 in primary care specialties (internal, family, adult, general practice)
3. 85,722 in 29 party-registration states
4. 42,861 in 50% simple random sample
Targeting the Population

Step 1  The National Provider Identification File
1. 560,896 U.S.-based physicians
2. 149,936 in primary care specialties (internal, family, adult, general practice)
3. 85,722 in 29 party-registration states
4. 42,861 in 50% simple random sample

Step 2  Match to Voter File (Catalist)
Targeting the Population

Step 1  The National Provider Identification File

1. **560,896** U.S.-based physicians
2. **149,936** in primary care specialties (internal, family, adult, general practice)
3. **85,722** in 29 party-registration states
4. **42,861** in 50% simple random sample

Step 2  Match to Voter File (Catalist)

1. **161,553** plausible matches sent by Catalist
Targeting the Population

Step 1  The National Provider Identification File
1. 560,896 U.S.-based physicians
2. 149,936 in primary care specialties (internal, family, adult, general practice)
3. 85,722 in 29 party-registration states
4. 42,861 in 50% simple random sample

Step 2  Match to Voter File (Catalist)
1. 161,553 plausible matches sent by Catalist
2. 18,430 records with unique matches,

Of 24,250 confident matches, 20,296 mailable addresses

Step 3  Target sample
Of 20,296, 13,678 when restricted to registered Ds and Rs
Oversample doctors in mixed-partisan practices
1. Classify practices by size and partisan composition
2. Draw 100% physicians from mid-size bipartisan stratum (754) and 6% sample of all other strata (775)
3. Survey 1,529 physicians
Targeting the Population

**Step 1** The National Provider Identification File

1. **560,896** U.S.-based physicians
2. **149,936** in primary care specialties (internal, family, adult, general practice)
3. **85,722** in 29 party-registration states
4. **42,861** in 50% simple random sample

**Step 2** Match to Voter File (Catalist)

1. **161,553** plausible matches sent by Catalist
2. **18,430** records with unique matches, plus **5,820** confident matches (57% match rate)
Targeting the Population

Step 1  The National Provider Identification File
1. 560,896 U.S.-based physicians
2. 149,936 in primary care specialties (internal, family, adult, general practice)
3. 85,722 in 29 party-registration states
4. 42,861 in 50% simple random sample

Step 2  Match to Voter File (Catalist)
1. 161,553 plausible matches sent by Catalist
2. 18,430 records with unique matches, plus 5,820 confident matches (57% match rate)
3. Of 24,250 confident matches, 20,296 mailable addresses
Targeting the Population

Step 1  The National Provider Identification File

1. 560,896 U.S.-based physicians
2. 149,936 in primary care specialties (internal, family, adult, general practice)
3. 85,722 in 29 party-registration states
4. 42,861 in 50% simple random sample

Step 2  Match to Voter File (Catalist)

1. 161,553 plausible matches sent by Catalist
2. 18,430 records with unique matches, plus 5,820 confident matches (57% match rate)
3. Of 24,250 confident matches, 20,296 mailable addresses

Step 3  Target sample
Targeting the Population

Step 1  The National Provider Identification File

1. 560,896 U.S.-based physicians
2. 149,936 in primary care specialties (internal, family, adult, general practice)
3. 85,722 in 29 party-registration states
4. 42,861 in 50% simple random sample

Step 2  Match to Voter File (Catalist)

1. 161,553 plausible matches sent by Catalist
2. 18,430 records with unique matches, plus 5,820 confident matches (57% match rate)
3. Of 24,250 confident matches, 20,296 mailable addresses

Step 3  Target sample

Of 20,296, 13,678 when restricted to registered Ds and Rs
Targeting the Population

Step 1  The National Provider Identification File

1. 560,896 U.S.-based physicians
2. 149,936 in primary care specialties (internal, family, adult, general practice)
3. 85,722 in 29 party-registration states
4. 42,861 in 50% simple random sample

Step 2  Match to Voter File (Catalist)

1. 161,553 plausible matches sent by Catalist
2. 18,430 records with unique matches, plus 5,820 confident matches (57% match rate)
3. Of 24,250 confident matches, 20,296 mailable addresses

Step 3  Target sample

Of 20,296, 13,678 when restricted to registered Ds and Rs

Oversample doctors in mixed-partisan practices

1. Classify practices by size and partisan composition
2. Draw 100% physicians from mid-size bipartisan stratum (754) and 6% sample of all other strata (775)
Targeting the Population

Step 1  The National Provider Identification File

1. 560,896 U.S.-based physicians
2. 149,936 in primary care specialties (internal, family, adult, general practice)
3. 85,722 in 29 party-registration states
4. 42,861 in 50% simple random sample

Step 2  Match to Voter File (Catalist)

1. 161,553 plausible matches sent by Catalist
2. 18,430 records with unique matches, plus 5,820 confident matches (57% match rate)
3. Of 24,250 confident matches, 20,296 mailable addresses

Step 3  Target sample

Of 20,296, 13,678 when restricted to registered Ds and Rs
Oversample doctors in mixed-partisan practices

1. Classify practices by size and partisan composition
2. Draw 100% physicians from mid-size bipartisan stratum (754) and 6% sample of all other strata (775)
3. Survey 1,529 physicians
...We are trying to better understand how doctors take a patient’s social history and what factors may impact differences in the ways physicians approach a patient’s social history...
Nine Vignettes

A healthy-appearing, 38-year old, male patient comes to your office for a physical. This is his first appointment with you. He does not have any known prior chronic medical issues. During the patient interview, the patient...
The Survey

Nine Vignettes

A healthy-appearing, 38-year old, male patient comes to your office for a physical. This is his first appointment with you. He does not have any known prior chronic medical issues. During the patient interview, the patient...

- ... acknowledges consuming about 20 alcoholic beverages in a typical week but denies any related physical concerns.
The Survey

Nine Vignettes

A healthy-appearing, 38-year old, male patient comes to your office for a physical. This is his first appointment with you. He does not have any known prior chronic medical issues. During the patient interview, the patient...
The Survey

Nine Vignettes

A healthy-appearing, 38-year old, male patient comes to your office for a physical. This is his first appointment with you. He does not have any known prior chronic medical issues. During the patient interview, the patient...

- ... acknowledges using recreational marijuana approximately three times per week but denies any related physical concerns.
Results

Democrats More Concerned

Republicans More Concerned

Firearms

Sex Worker

Depression

Alcohol

Obesity

Cigarette

Helmet

Marijuana

Abortion

Effect of Republican on Perception of Seriousness
Results

Marijuana

- Refer to counseling
- Ask motivation to stop
- Urge patient to cut down
- Discuss health risks
- Discuss legal risks

How Likely to Include in Treatment Plan, Republican Effect

Advantages: over-sample mixed practices, target survey to homes, avoid political questions
Example 9: Eitan Hersh and Gabrielle Malina, “Partisan Pastor” In Progress

Research questions: Do pastors reflect the political views of their congregants? Do Democratic and Republican pastors lead differently?
Example 9: Eitan Hersh and Gabrielle Malina, “Partisan Pastor” In Progress

Research questions: Do pastors reflect the political views of their congregants? Do Democratic and Republican pastors lead differently?

Data

- 41 find-a-church websites, scraped (Upwork)
- Additional pastors added (MTurk)
- Voter file link (Catalist)
- link to survey data
Advantages: comprehensive data on small but important population, within- and cross-denominational variation
Advantages: comprehensive data on small but important population, within- and cross-denominational variation
Some Lessons
Some Lessons

Match rate to registered voters
Some Lessons

<table>
<thead>
<tr>
<th>Match rate to registered voters</th>
<th>Doctors</th>
<th>Pastors</th>
<th>9/11</th>
<th>Donors</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Pass</td>
<td>43%</td>
<td>44</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Some Lessons

<table>
<thead>
<tr>
<th>Match rate to registered voters</th>
<th>Doctors</th>
<th>Pastors</th>
<th>9/11</th>
<th>Donors</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Pass</td>
<td>43%</td>
<td>44</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second Pass</td>
<td>57</td>
<td>63</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Some Lessons

<table>
<thead>
<tr>
<th></th>
<th>Doctors</th>
<th>Pastors</th>
<th>9/11</th>
<th>Donors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Pass</strong></td>
<td>43%</td>
<td>44</td>
<td>68</td>
<td>79*</td>
</tr>
<tr>
<td><strong>Second Pass</strong></td>
<td>57</td>
<td>63</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Only Pass</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Some Lessons

Mail survey response rates

Doctors: 20%
Donors: 17%
Some Lessons

Mail survey response rates

Doctors: 20%

Donors: 17%
Some Lessons

1. 8/1,529 "doctors" (0.5%) said they weren’t the doctor
2. 21/6,500 "donors" (0.3%) said they weren’t the donor
3. False positives in Texas case:

<table>
<thead>
<tr>
<th>SSN Match</th>
<th>No SSN Match</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>135,686</td>
<td>1,343,425</td>
</tr>
<tr>
<td>FP</td>
<td>1,207,739</td>
<td>1,327,340</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6,712,256</td>
</tr>
</tbody>
</table>
Some Lessons

False positives

<table>
<thead>
<tr>
<th>ADGN Match</th>
<th>SSN Match</th>
<th>SSN9 Match</th>
<th>No SSN Match</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No ADGN Match</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,207,739 [TN]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>135,686 [FN]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1,343,425</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADGN Match</td>
<td>119,601 [FP]</td>
<td>5,249,230 [TP]</td>
<td></td>
<td>5,368,831</td>
</tr>
<tr>
<td>Total</td>
<td>1,327,340</td>
<td>5,384,916</td>
<td></td>
<td>6,712,256</td>
</tr>
</tbody>
</table>
Some Lessons

False positives

1. 8/1,529 “doctors” (0.5%) said they weren’t the doctor
Some Lessons

False positives

1. 8/1,529 “doctors” (0.5%) said they weren’t the doctor
2. 21/6,500 “donors” (0.3%) said they weren’t the donor
Some Lessons

False positives

1. 8/1,529 “doctors” (0.5%) said they weren’t the doctor
2. 21/6,500 “donors” (0.3%) said they weren’t the donor
3. False positives in Texas case:
Some Lessons

False positives

1. 8/1,529 “doctors” (0.5%) said they weren’t the doctor
2. 21/6,500 “donors” (0.3%) said they weren’t the donor
3. False positives in Texas case:

<table>
<thead>
<tr>
<th>ADGN Match</th>
<th>No SSN Match</th>
<th>SSN Match</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No ADGN Match</td>
<td>1,207,739 [TN]</td>
<td>135,686 [FN]</td>
<td>1,343,425</td>
</tr>
<tr>
<td>ADGN Match</td>
<td>119,601 [FP]</td>
<td>5,249,230 [TP]</td>
<td>5,368,831</td>
</tr>
<tr>
<td>Total</td>
<td>1,327,340</td>
<td>5,384,916</td>
<td>6,712,256</td>
</tr>
</tbody>
</table>
Behavioral Applications of Voter Files

Eitan Hersh

Tufts University
Department of Political Science

June 14, 2018