13 September 2022

Note About Vote File Data

The data used in Enos (2016)[1] was taken from an Illinois voter file produced by the commercial firm VCS (Voter Contact Services). The file was produced sometime between November 2004 and December 2005.[2] The purpose of this note is to provide background information on voter files as a data source, how they relate to other official records, and potential sources of error in the data when measuring behaviors such as voter registration and voter participation.

In the United States, a person must register in order to vote. The rules and processes for registering vary from state to state. Voter files are lists of registered voters collected by states and local governments. Each row consists of a single voter, typically with some demographic and administrative variables, and typically including indicators of turnout in past elections. How the data are collected and maintained varies from state to state. In some states, the data are collected by counties and then passed to the state governments. In the last decade or so, this process has become more centralized in many states so that registration is maintained by the state government only. The data are made available for use by politicians to aid in their campaigns. The information contained in the file is useful for campaigning because it tells politicians whom to target for mobilization efforts, using variables such as address to find voters who live in certain districts and turnout history to know where to target resources most efficiently.[3]

The ideal for state governments is to collect data that perfectly capture the electorate during a campaign, which means having an accurate representation of who lives where and their past vote history. For this reason, in theory, states will add and remove voters from the rolls as they move, die, or have other changes that should affect their relevance to campaigns. This creates two potential sources of error when using these files for research: First, a voter file created at any given moment is unlikely to match the official records of turnout and registration from recent elections because of population changes, such as movement, death, and other removals of people after the election. Other removals include states purging voter files of records that the state assumes to be inactive. This size of the error is likely to increase with the time between the creation of the voter file and the election in question. Second, records are further distorted because of limitations in administrative capacity, such as mistakes in data entry or failure to match across records as individuals move residences. This was especially true in past decades when record-keeping technology and coordination across agencies was relatively poor. As voters moved, say across, or even within counties and states, official records would not keep up with these voters. Records can be bad for other reasons: often, even when people die, the records are not shared across government agencies, so voters continue to appear as registered voters, even though they had died years before. Modern improvements, including the centralization of records by states and even commercial firms, has improved the overall accuracy, but issues remain.

---

[1] Enos, Ryan D. 2016. "What the Demolition of Public Housing Teaches Us About the Impact of Racial Threat on Political Behavior," *American Journal of Political Science 60*(1):123–142

[2] VCS was sold to Labels and Lists in 2011 (https://www.elections.il.gov/ElectionOperations/VoterTurnout.aspx). Labels and Lists was later rebranded as L2.

[3] See: Hersh, Eitan D. *Hacking the electorate: How campaigns perceive voters.* Cambridge University Press, 2015.

These processes, in addition to distorting counts of overall registration, can also distort counts of voter turnout. A file has a record of turnout attached to each individual voter. If a voter moves and the record is not linked across addresses, then the voter may be counted as not voting in past elections because the previous vote history is lost. This creates the potential for an undercount of participation. On the other hand, if a voter moves and vote history is kept, this can also create a misleading view that the voter participated in a locality in which they did not participate, say if the records were linked and not flagged as having moved (which states do not always effectively do).

The important take away from this discussion is that these files should not be expected to recover the official count of votes in any given election. Depending on the source of the error, this may create measurement error that inflates or deflates counts of registration and/or turnout. This error may vary across localities within a state: For example, a locality with a fast-growing population that has many new potential voters moving in may be adding new voters, but if it has poor record keeping, then it might not track the voting history of these new voters and so it might undercount participation. On the other hand, some states might do a good job keeping vote histories intact as individuals move, which can be useful for campaigns in identifying likely voters, but this can inflate vote counts in a particular locality for past elections.

And there are other sources of administrative error too, most of which probably contribute to turnout undercounts. For example, the simple process of a poll worker putting a check mark next to the name of the voter after handing them a ballot: the voter may get the ballot, cast it, and have the vote added to the totals, but may not be recorded as voting on the file because the check missed their name at the polling place. Provisional and absentee ballots can create similar problems with recording who voted.

Commercial firms that collect and sell voter file data are also important and increasingly common sources of data. In theory, these firms specialize in keeping accurate records, even more accurate than the states themselves, so that they have a valuable product to sell to campaigns. Part of their added value is that they attempt to track voters who move, die, or otherwise leave the electorate, so that campaigns will not waste resources on mistargeting. The tracking these firms employ is an inexact science because there is no ground truth: in the United States, there is no central repository of people, citizens, or voters.[4] So, these firms do educated guessing, based on matching names and other characteristics to track individuals across place and time.

This educated guessing creates large discrepancies between the files sold by different firms and these discrepancies are illustrative of the error contained in this type of data. For example, I have access to the nationwide voter files from three reputable firms: L2, Catalist, and Target Smart. The number of voters in the three datasets in 2018, all of which are supposed to represent the complete set of voters in the United States in that year, varies by tens of millions of voters: L2 lists 180,735,645 voters, Catalist lists 192,224,447, and Target Smart lists 199,794,609.

The discrepancies in the total registered voters illustrates the size of error that may sometimes be found in these files for individual states and localities. Consider the nearly 20,000,000 record difference between Target Smart and L2. This means that Target Smart believes there are 20 million more voters in the United States than L2 does. If this difference is averaged across 50 states (not counting DC for the

[4] The closest dataset to a central repository of people is the Census, which is only accurate for a moment every 10 years, not publicly available, and not linked to voter registration.

purposes of this discussion), that would mean that, on average, the files would vary by 400,000 voters in any given state. Consider what this means for error from the true number of registered voters in that state at any given time: If we assume that the true count falls in the middle of the difference between the files, they would, on average be expected to each miss the true count by 200,000 voters in a state – and, of course, in populous states, this difference would be much larger. We should also expect that the deviations from true counts of turnout may be large. Of note is that a study by Igielnik et al. (2018)[5] compared five commercial voter files (presumably including the three that I cover here) and found that when attempting to match to a high-quality sample of survey panelists, only 42% of panelists could be matched to all five files and the turnout rate in 2016 varied across these five files by 15 percentage points, from 71% to 86% (even dropping the largest outlier, the turnout rate varied by 7 percentage points).

In summary, if a researcher is using a voter file to recover a snapshot of the electorate at a particular time, the researcher should consider possible sources of error created by administrative processes and the amount of time between when the file was created and the behavior they are attempting to measure. And, of course, the researcher should consider whether this error may be systematically related to other variables that may affect inference.

As an example of such measurement error and the consideration of potential errors created by the measurement error, consider the discrepancy between the voter file data in this article and the official counts of voters provided by the Chicago Board of Election Commissioners.[6] The VCS file used in the article has 1,366,357 voters in Chicago,[7] which is fewer than the 1,416,101 in official counts for the November 2004 General Election (one of the elections used in the article). The VCS voter file records 808,952 voters as having participated in the 2004 election, while the official counts have 1,056,830. So, the VCS file contains only 96% of the count of registered voters and only 76% of the count of voter turnout. The data from Chicago was extracted from a VCS file for all of Illinois that contains 7,062,615 registered voters, which is 94% of the 7,499,488 registered voters as of the 2004 General Election, according to the Illinois State Board of Elections.[8] Turnout for the 2004 election on the VCS file is 4,424,978, which is 83% of the 5,350,493 in the official count of turnout. Notably, this discrepancy in registration of just over 400,000 voters is similar to the 400,000 average state discrepancy estimate based on differences across commercial files that I noted above. The disproportionate discrepancy in Chicago, compared to the state at large, would be consistent with the population in the largest and most diverse city in the state being the most fluid.

Because the voter files do not perfectly capture the official counts for an election used to make inferences in the article, if we assume the official counts represent the ground truth of registration and turnout, the files contain measurement error. This error does not seem unusual given the differences observed in other commercial voter files, but it still means that nearly 1 in 4 voters who participated, according to official counts, are misreported, so it is important to consider how this error might bias inference. Recall that the design in the article for estimating changes in voter turnout as a function of a demographic change was a difference-in-differences estimation comparing changes in turnout between

[5] https://www.pewresearch.org/methods/2018/02/15/commercial-voter-files-and-the-study-of-u-s-politics/
[6] https://chicagoelections.gov/en/election-results-specifics.asp
[7] Note that the replication file contains 1,132,646 voters because of losses due to an inability to geocode and other reasons reported in the Appendix.
[8] https://www.elections.il.gov/ElectionOperations/VoterTurnout.aspx

2000 and 2004 for voters close to demolished housing projects to those farther away and then making this comparison between white and Black voters.[9] White voters living near the projects were found to have their turnout decrease between 2000 and 2004 compared to white voters living farther away, while Black voters saw little or no change. While a mismeasurement of turnout on the voter file is likely correlated with other variables of interest, such as race and even location, the difference-in-differences design guards against such errors biasing the findings, because the errors would have to be correlated with both the changes in turnout and distance from the projects, in addition to race.

Because it is not possible to observe errors in turnout and registration for individuals in this historical data, it is not possible to check for such correlations directly. While, on the face of it, it seems unlikely that errors would be correlated with location, especially the distance from demolished housing projects in particular, in addition to race and changes in turnout, Chicago was experiencing demographic changes at the time that raise a particular concern: namely that after the demolition of the housing projects, the white population of the areas around these projects increased. These white residents who moved to the area near the demolished projects after the projects were demolished may have been likely to have their vote histories lost with the move and, thus, the measurement error may be correlated with race and distance from the housing projects. This lost vote history for a person moving between 2000 and 2004 is guarded against in the design because voters were not used in the estimation if they were not registered before the demolition of the housing projects. However, even if exclusion based on registration date did not guard against this sort of error, this demographic change is not likely to pose a threat to inference because the bias would most likely be conservative – running in the opposite direction of the findings in the article – and, thus, biasing the result toward zero. To see this, consider that if vote histories prior to the move were lost, voters who had actually voted in 2000 would be coded as having not voted. So, in 2004, a voter with a lost vote history who does not vote would count as no change in turnout and a voter who does vote in 2004 would be seen as having increased turnout, going in the opposite direction of the decrease in turnout observed in those living near the demolished projects.

[9] I also controlled for covariates in some specifications.