12 October 2022

This note concerns the following article:

Enos, R.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: 123-142. https://doi.org/10.1111/ajps.12156

I was recently asked by the editors of the *American Journal of Political Science*, in response to a query from another scholar, to add a note to the article Dataverse on the use of commercial voter files for political science research. I have added this note here: https://doi.org/10.7910/DVN/26612.

This article used voter file and precinct-level vote returns to measure how voting changed in response to demographic changes caused by the large-scale removal of public housing in Chicago between 2000 and 2004.

The query to which I was responding was questioning why the replication files for the article included fewer registered voters and a lower participation rate than what is found in official records from the Chicago Board of Election Commissioners. This is a worthwhile question and something that I could have been explained more clearly in the original article, so I welcome the opportunity to add the addendum. Below, in section 1 of this memo, I will provide a selection from my complete response to the editors. I add this here because it gives more detail than what was posted on Dataverse.

There is also another section in the article where the replication data does not match official records: this is in precinct data used to measure the change in vote choice pre and post removal of the public housing. In section 2 of this memo, I give more details on this data and why counts of the precincts in the replication data do not match official records.

In both cases, the reasons for the discrepancies between the data provided for the article and official reports are probably not surprising for scholars familiar with electoral data in the United States but some of the information on the size of some variation in voter files may be of interest to those who use this type of data. I also hope that these more detailed explanations will clear up any confusion about the article.

I can be reached at renos@gov.harvard.edu for questions or comments.

SECTION 1: VOTER FILE DATA

Below is an excerpt from the response I provided to *AJPS* editors on August 3, 2022. It may be useful for understanding the use of voter files generally and in the particular case of the replication data from the *AJPS* article.

> I was asked to respond to the following:
>
> > *From the official election data in 2004 in Chicago, there are totals of 1,416,101 registered voters and 1,056,830 ballots cast. This provides a total of 359,271 non-voters in the 2004 election. From Enos (2015)'s replication file, there are 1,132,646 observations (perhaps understandably, as the manuscript notes there were issues with*

> *geocoding certain voters). From these observations, 463,531 have a value of "0" on Enos (2015)'s vote variable, suggesting they are non-voters. This gives us over 100,000 more non-voters in the Enos (2015) dataset than in the official Chicago records, despite the Enos dataset having nearly 300,000 fewer observations. The turnout rate reported by the official Chicago website was 74.63%, while the turnout rate in the Enos (2015) data was 59.08%, a rather large discrepancy.*

In response, I have prepared this brief memo. It has four sections:

1. An explanation of the processing of the individual voter data used in the article.
2. An overview of voter files generally and why the noted sort of discrepancy in turnout numbers is not unusual.
3. A brief argument for why any measurement error created by such discrepancies should not be expected to create bias using the research design in the article.
4. Some reflection on potential points of confusion with the article.

I will cite numbers from quite a few datasets below. If the editors would like to verify my numbers, please let me know and we can arrange a way for the data to be shared.

Processing the data:

The data I used in my article was taken from an Illinois voter file produced by the commercial firm VCS (Voter Contact Services). Based on the creation date of the files on my server, I probably obtained this file sometime in late 2005, most likely December. Based on when I obtained it and the last recorded election in that file, which is the November 2004 General Election, the file was probably created sometime between November 2004 and December 2005. I will refer to this file as the *2005 IL VCS file.* Note that the firm VCS appears to have been sold to Labels and Lists in 2011.[1] Labels and Lists was later rebranded as L2. I'll say more about L2 data below.

The 2005 IL VCS file, is intended to be a snapshot of all voters in Illinois at the time it was created, but it contains only 7,062,615 voter records, which is less than the 7,499,488 registered voters in the state for the 2004 general election according to the Illinois State Board of Elections.[2] I will explain more about why such discrepancies are not unusual below, but I mention it now to demonstrate that the difference in registered voters in Chicago between my replication data and in the official counts is probably not a consequence of my processing of the data because it can be found in the unprocessed statewide file as well.

[1] https://campaignsandelections.com/industry-news/labels-lists-buys-vcs/

[2] https://www.elections.il.gov/ElectionOperations/VoterTurnout.aspx

I extracted all voters in Chicago by mailing address, so I simply searched for voters with "Chicago" in the city variable included in the file.[3] I'll refer to this sample as the *2005 Chicago VCS sample.* This sample had 1,366,357 records. This is fewer than the 1,416,101 registered voters in 2004 according to the Chicago Board of Election Commissioners,[4] but again this appears to be consistent with the discrepancies between the 2005 IL VCS file and the official state counts. It's also notable that the official number for registered voters in Chicago at the time of 2000 the general election was 1,472,534, so the official number of registered voters actually decreased between 2000 and 2004. This might be a result of purging or moving (see below) and the number of registrants found in the 2005 Chicago VCS sample might be a reflection of these decreases.

The 2005 Chicago VCS sample contains a variable for voting in 2004, which was coded 1 = voting in the 2004 statewide primary only, 2 = voting in the 2004 general election only, and 3 = voting in the 2004 statewide primary and general election, with missing values for all others. These missing values represent non-voting, either because the voter didn't vote when eligible or other reasons, including not being part of the electorate during the election (see below). I created a binary variable of 1 for voters who had a 2 or 3 in the original file and 0 otherwise. Note that, as with the missing values in the original 2005 IL VCS file, the 0 value in the 2004 turnout variable that I created collapses voters who did not vote for any reason, including not voting when eligible or not being part of the electorate.

The 2005 Chicago VCS sample contained 808,952 voters in Chicago who voted in the 2004 general election, which is fewer than the 1,056,830 on official records. The turnout percentage of 59% in the 2005 Chicago VCS sample was also less than the 75% turnout reported in official records. Note too that the 2005 IL VCS file contained 4,424,978 voters who voted in the 2004 election, which is fewer than the 5,350,493 reported in official records. This is notable because while the 2005 IL VCS file contained 94% of the registered voter as can be found in the official counts (see above), it only contained 83% of the turnout total in 2004, so, similar to the 2005 Chicago VCS sample, the entire 2005 IL VCS file contains more non-voters than would be expected given the discrepancy in registration numbers from official counts – although not as great of a discrepancy as Chicago, which contains 96% of the registered voters (1,366,357/1,416,101) and 76% (808,952/1,056,830) of the turnout. As I explain below, record keeping of voter files is often decentralized to different localities and the population of Chicago is much different than Illinois as a whole – including, crucially, probably being more fluid – so these differences in the discrepancies in turnout rates might be expected.

To create the file contained in the replication material, which I will call the *replication file*, I removed voters who could not be geocoded, both with distances from the housing projects and for matching to Census data and also voters for whom race could not be imputed. The selection of the sample, including the removal of other voters, is described in the Appendix of the article. As noted, this yielded a data set of 1,132,646 voters, of which 669,115 are coded as having

[3] In some municipalities, where common mailing address names do not necessarily align with municipal boundaries, e.g., "Brooklyn" rather than "New York City" this method of finding voters by text might miss some people, but I don't believe this is an issue in Chicago, where everywhere uses "Chicago" in the address.
[4] https://chicagoelections.gov/en/election-results-specifics.asp

voted in 2004. This is a turnout rate of 59%, which is consistent with the turnout rate in the 2005 Chicago VCS file, so the turnout discrepancy between the replication file and the official counts does not appear to be based on my processing of the data. That the turnout rate after processing remains the same is also notable because it indicates that my removal of voters was not correlated with voter participation, i.e., I was no more likely to remove a participating voter than a non-participating voter. This is important because turnout is the dependent variable in the analysis and so this makes it less likely that my processing of the data created bias in the inferences.

Voter files and error in voter files

My apologies if anybody reading this already knows the following about voter files, but it is important for framing the discussion. For a useful background on these files and their use in politics, I suggest Hersh (2015). Some of the description in that book (which is excellent) is a bit dated but is especially relevant to the files I used in the article (see below). Igielnik, et al (2018)[5] also did an important study comparing commercial voter files that deals with many of the same issues that I will describe below, including sources of error in the files.

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 many states, the data is collected by counties and then passed to the state governments. However, in the past decade or so, in many states, this process has become more centralized so that registration is maintained by the state 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 who to target for mobilization efforts. These efforts make use of variables such as address to find voters who live in certain districts and turnout history to know where to target resources most efficiently.

The ideal for the state governments is to collect data that perfectly captures the electorate during a campaign, this 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. There are two key things to keep in mind about this: first, a voter file created at any given moment is unlikely to match official records of the last election 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 they assume to be inactive voters.[6] Second, records are

[5] https://www.pewresearch.org/methods/2018/02/15/commercial-voter-files-and-the-study-of-u-s-politics/

[6] The reasons for purging vary by state and there are sometimes accusations that purging is done to shape the electorate in ways that are favorable to a particular party. Other reasons for purging appear to be idiosyncratic: In Massachusetts, voter registration is handled at the town or city level. In Cambridge, where I live, I receive a yearly

further distorted because of limitations in administrative capacity, such as mistakes in entry or failure to match across records as individuals move residency. 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. For example, in examining a statewide California file from 2005 created by the secretary of state, which I'll call the *2005 CA SOS file* (see below), I personally appear in that file twice – once from when I was in college and then a newer address in graduate school.[7] Records can be bad for other reasons: often, even when people die, the records are not shared across government agencies so that 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 greatly improved the overall records, but issues remain.

These processes, in addition to distorting counts of overall registration – e.g., I was listed as registered in Alameda County, California, even though I didn't live there, can also distort counts and rates 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 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 in my case of moving from Alameda County to Los Angeles County, if the records were linked and not flagged as having moved (which states do not always effectively do), then somebody looking at the voter file might conclude that my votes cast in the election in Alameda County in November 2000 had been cast in Los Angeles County, thus inflating voter turnout in Los Angeles County for that election.

This distortion of turnout rates is worth dwelling on, so I will give another example to illustrate how it could systematically drive down the overall turnout rate in a voter file. In the 2005 Chicago VCS sample, I do not appear at all – even though I had lived in the city from 2001 to 2004 and was registered to vote. I moved away in the summer of 2004, so it could be that the state correctly removed me from the voter file after I left, however that strikes me as an unusually quick purging of a voter. It is also possible that VCS correctly identified me as having moved and removed me. In either case, it's a good example of how turnout counts be distorted because my voter turnout in 2002, an election in which I voted, would have also been purged. Say another person moved into my apartment after I left – if this person came from another state, county, or even possibly another address in Chicago and their vote histories were not carried with them – a likely event given the quality of record linking in 2005 – then the turnout rate in the file would be driven down. In fact, this can happen even if the same person moves within the same city: for example, examining a 2008 statewide voter file from the California Secretary of State, *2008 CA SOS file*, I appear on the file in the second address at which I lived in Los Angeles, but my previous registrations in Alameda County and my first address in Los Angeles have been deleted. These deletions are good record-keeping because I no longer lived

city-wide census form with a stark warning that failure to complete the census for two consecutive years will result in removal of my voter registration.

[7] This also speaks to the idiosyncrasies in record keeping because I had also been registered when I was in high school, but that record has been removed from the 2005 CA SOS file, while my college registration had not.

at either of those locations – but my record in 2008 had lost my vote history from the 2004 general election. Because my record from 2004 was lost, even though I had cast a vote and was the same person, the file would downwardly bias the voter turnout for that election, simply because somebody had moved within the same city.[8]

This sort of movement alone could explain the sort of discrepancies noted in my replication file and the 2005 Chicago VCS sample. Other than the 2005 IL VCS file, the oldest Illinois file in my possession is a 2012 file from the commercial firm Target Smart (I will discuss Target Smart in more detail below). I extracted the Chicago sample using the address field and will call this the *2012 Chicago TS sample*. In the 2012 Chicago TS sample, the firm flags 99,190 voters as having "recently moved." These represent voters who had recently changed addresses. It is not clear if the vote history of these movers would be preserved by Target Smart, but 99,190 movers, if this represents a typical year, would be approximately enough movers to explain the differences between the 2005 Chicago VCS sample and the official counts if their voting records were not maintained.

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 particular source of the error, this may create measurement error that inflates or deflates counts of registration and this may vary across locality. For example, if locality has a fast-growing population so that lots of new voters are moving in, this may mean that it adds new voters but if it has poor record keeping, then it might not track the voting history of these new voters and so might undercount participation. On the other hand, some states might do a good job of 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.

And there are other sources of administrative error too, most of which probably contribute to undercounts. For example, the simple process of a poll worker putting a check 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 Voter Files

Commercial firms that collect and sell voter file data are also important players in this landscape. The VCS data that I used is from a commercial firm. 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

[8] In fact, even though the voter file has a field for previous address, my previous address as not recorded with my updated record.

resources on mistargeting.[9] 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.[10] So, these firms do educated guessing, based on matching names and other characteristics to keep track of individuals.

This educated guessing creates large discrepancies between the files sold by different firms and these discrepancies are illustrative of why voter files should not be expected to match official vote counts. For example, I have access to the nationwide voter files from three reputable firms: L2, Catalist, and Target Smart.[11] L2 is very commonly used by academics. Target Smart is rarely used by academics but, as far as I can tell, is considered high quality, and is widely used by Democratic campaigns. Catalist is somewhat more commonly used by academics and is also considered very reputable by practitioners. 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, vary by tens of millions of voters: L2 lists 180,735,645 voters, Catalist has 192,224,447, and Target Smart has 199,794,609.[12]

The discrepancies in the total registered voters from these commercial firms is worth dwelling on because it shows very plainly that counts of voters in voter files should not be expected to match official totals. 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. If this difference is averaged across 50 states (not counting DC for the 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 official counts in each of these states: If we assume the official count falls somewhere in the middle of the difference between the files and if they both had the same amount of error from the official counts, they would, on average be expected to each miss the official count by 200,000 voters in a state – and, of course, in populous states (like Illinois), this difference would be much larger. And, unless we expect that the missing or extra voters in these files are also perfectly aligned with official counts in their turnout, the deviations from official counts of turnout will also likely be large. Of note is that the Igielnik, et al (2018) study 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, importantly for the discussion here, 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).

[9] Notably, these lists are now widely used in political science research and they now seem to be the most common way that researchers obtain voter files – although when I obtained the VCS list in 2005, I think using commercial lists was uncommon.

[10] 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.

[11] Note that I don't have direct access to a Catalist file, but rather collected the numbers below from a colleague.

[12] It might be meaningful that L2, which bought VCS, is lower than the rest because, as noted, the 2005 VCS file has fewer voters than the official counts.

Comparing the 2005 IL VCS file to Other Voter Files

I believe that the fact that these files will generally not accurately represent the official statistics is generally understood by people working regularly with these files, but I wanted to check to see if the 2005 IL VCS file that I used in the article had an especially large amount of error, so I audited some other files in my possession. This is not a systematic survey of voter files because I don't have a systematic collection and I did not look at every file I have, rather I looked at files that had characteristics that may make them good benchmarks to judge the amount of error.

Perhaps a useful benchmark is to consider the number of votes cast in the 2005 Chicago VCS sample as percentage of the official counts. As mentioned above, the 2005 Chicago VCS sample contains 96% of the registered voters (1,366,357/1,416,101) and 76% (808,952/1,056,830) of the official votes cast in 2004.

A good place to start might be to look at files from the same state and, because the technology for maintaining and matching records has improved with time, to look at files as close in time as possible to the 2005 file used in the article. Checking the 2012 Chicago TS sample, discussed above, for turnout in the 2004 general election, it lists 773,987 individuals has having voted, which is 73% of the official total – but given that the file was produced in 2012, and lots of voters moved in and out of the city across those 8 years, this might not be a fair test. I can, however, look at the 2012 Chicago TS sample total for the 2010 election general, which would be an election less than two years before the creation of the 2012 Chicago TS sample – this is a somewhat similar timespan between the creation of the 2005 IL VCS file and the November 2004 election. Target Smart lists 661,781 has having voted in that election, which is 94% of the 705,869 from the Chicago Board of Elections, certainly better than 76% of the total in the 2005 Chicago VCS file, but also informative because commercial voter files are often considered to have improved quite a bit between 2005 and 2012 and Target Smart still is missing the official counts by over 5 percentage points.

I also examined a file for Illinois from the commercial firm L2 from 2018, I'll call this the *2018 IL L2 file*. This file is obviously quite a bit newer than the 2005 IL VCS file so record-keeping has probably improved considerably, but it might also be informative because it is from the same state and, as noted above, L2 purchased VCS so the methodology used to create the two files may have some similarities. This 2018 IL L2 file has 7,742,183 registered voters, which is 95% of the 8,099,372 in official counts, but also lists 5,231,408 has having voted, which is 112% of the official voter turnout count of 4,635,541. This means the turnout rate in the 2018 IL L2 file was 68% compared to 57% in official counts, so this is a good example of a commercial file distorting voter turnout rates in a similar fashion, but opposite direction, to the 2005 IL VCS file used in the article.[13]

Another perspective is to look at files that were produced at approximately the same time as 2005 IL VCS file. For this, I looked at the file from September 2005 by the California Secretary of State, the *2005 CA SOS file*. This was produced around the exact same time as the 2005 VCS file, although a crucial difference is that the 2005 CA SOS file was not processed by a commercial

[13] I did not check the turnout rates for Chicago specifically using the 2018 IL L2 file, but I can do so if it would be of interest.

firm – this is the raw data collected by the state from the counties. The 2005 CA S0S file lists 16,271,340 registrants. According to the CA Secretary of State[14], 16,557,273 Californians were registered, so there is an undercount in the file of almost 300,000 voters. The 2005 CA SOS file lists 11,993,608 as having voted in the November 2004 election, while the official count is 12,589,683 votes, so the 2005 CA SOS has 95% of the official count. The discrepancy of about 500,000 voters is actually similar in absolute size, although smaller in percent, to the discrepancy found in the 2005 IL VCS data. But this discrepancy is also very telling because the 2005 CA SOS file and the official counts were created by the exact same government agency and, yet there is a difference of 500,000 voters. Counting voters appears to just by a very inexact process.

To examine similar geographies to Chicago, I checked three large counties in the 2005 CA SOS file. I extracted the voters for these counties using the locality code provided in the file. In the 2005 CA SOS file, Los Angeles County has 2,999,089 ballots cast in November 2004, while the LA County Registrar[15] lists 3,085,582 ballots cast. This is a discrepancy of over 86,000 voters but is fairly small in percentage terms because it comes within 97% of the total. What is remarkable about the LA totals though is that the 2005 CA SOS file contains 4,103,765 registered voters, compared to 3,972,738 in official counts, so the 2005 CA SOS file contains over 200,000 additional non-voters compared to the official counts in Los Angeles County.[16]

So, the 2005 CA SOS file overreports registered voters in Los Angeles County, but it underreports registered in San Diego County: San Diego County contains 1,371,359 registered voters, which is over 100,000 fewer and represents 90% of the 1,513,300 listed by the San Diego Registrar of Voters.[17] The file also contains 1,090,708 voters turning out in November 2004, which is 95% of the 1,145,035 on official counts. So, while voter turnout rates are underreported in Los Angeles County, they are overreported in San Diego County

San Francisco County (which is conterminous with the City of San Francisco) has 433,959 voters in the 2005 CA SOS, while the Department of Elections[18] lists 486,937, so again an undercount. The file lists 343,817 as having voted, while the official count lists 361,822, again about 95% of the total.

I also looked for other files from my collection (I have idiosyncratically collected these since graduate school) from around the same time period. I have two other files from 2005 or earlier, neither of which includes records of participation, but can be useful in checking for discrepancies in voter registration: I have a statewide file from Texas in 2005, but I do not know the exact source. The file contains 6,832,553 voters, which is only 56% of the 12,308,372 listed by the Texas Secretary of State.[19] I also have a New York State file from 2001 from L2. The 2001

[14] https://www.sos.ca.gov/elections

[15] https://www.lavote.gov/home/voting-elections

[16] Also remarkable is that the file lists only 1,068,522 in November 2000, less than half of the 2,769,927 in official counts.

[17] https://www.sdvote.com/content/rov/en/past-election-info.html

[18] https://sfelections.sfgov.org/results-summary-nov-2004

[19] https://www.sos.state.tx.us/elections/historical/jan05.shtml

NY L2 file contains 9,995,513 voters, which means the file is missing over 1,000,000 voters and is only 91% of the 11,033,578 listed by the New York State Board of Elections.[20] For the five counties making up New York City, the file contains 3,626,785 registered voters, which is only 88% of the 4,104,923 according to official counts. This discrepancy of 12 percentage points in registration in New York City is arguably meaningful because it is similar to the 2004 Chicago VCS sample in that it comes from a commercial voter file from a large city in the early 2000s.

Implications for the published article:

What I've offered is in no way a systematic accounting of discrepancies between voter files and official counts, but I think is enough to demonstrate that the differences between the 2005 IL VCS file I used in my article, both before and after it was processed, and official counts are not surprising. The discrepancies are larger than most others I checked in preparing this response, but it would take a systematic review of voter files to better understand if the size of the discrepancies represents an outlier. Chicago in early 2000s was not known for great administrative capacity, so one could imagine that it produces more error than the typical city. It might also be that VCS was more aggressive in changing voter files than other current firms, but there is no obvious way to verify that. It's also notable that files from Los Angeles and New York City from a similar time-period seem to have similar discrepancies. It should also be noted that such discrepancies will exist for nearly any research conducted with voter files, of which measuring turnout is a common use. However, whether common or not, the discrepancies still might have implications for the veracity of the findings in my article because the discrepancy represents measurement error (if we take the official counts as the ground truth).

Of course, no measure of behavior in the social sciences is without error – take for example survey data, a backbone of data in behavioral science, that is simply accepted to contain measurement error in every variable. So, having established that the measurement error is not unusual, the question is not whether there is measurement error in the data used in this article, but whether the measurement error biases the inferences. Recall that the design in the article was a differences-in-differences comparing changes in turnout between 2000 and 2004 for those close to demolished housing projects to those further away and further making this comparison between white and Black voters.[21] White voters living near the projects were found to have their turnout decrease between 2000 and 2004 compared to white voters living further away, while Black voters saw no change. So, for the errors in the voter file to bias the findings, the errors would have to be correlated with both the changes in turnout and distance from the projects, in addition to race.

As a general statement, it seems unlikely that errors in the file would be correlated with all of these other variables. It would be difficult to test for these correlations directly because we don't know for which voters the errors exist. A possible strategy would be to see if there are systematic errors in certain precincts, but even then one would not know for which individual voters the errors existed. This being said, one could tell a plausible story that voters incorrectly coded as non-voters are disproportionately likely to be white residents who moved to the area

[20] https://www.elections.ny.gov/EnrollmentCounty.html

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

near the demolished projects after they were demolished and saw their vote histories lost with the move. Given the gentrification that happened in the areas of the demolished housing projects, this is a plausible story, but not a threat to the inferences made in the article 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 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. In summary, the measurement error seems to pose little threat of biasing the results in the direction of a false positive.

SECTION 2: PRECINCT-LEVEL DATA

In a second analysis in the article, I examined vote for Republican candidates using precinct-level election results in the 1996, 2000, 2004, and 2008 Presidential general elections and votes for Barack Obama in the 2004 Democratic United States Senate primary election, 2004 US Senate general election, and the 2008 Presidential primary election. These analyses estimated support for certain candidates by Black and white voters using ecological inference and then examined differences between vote choice for those voters close to the demolished public housing projects and farther away from the demolished housing projects. The precincts closer and farther away were matched on income, separately for Black and white voters, before the differences were estimated. There are three key comparisons on which the inferences are based: 1) the differences between Black and white voters, 2) the differences between voters close and far from the demolished housing projects, and 3) the differences between these differences before and after the housing projects were demolished.

For these inferences, measuring distances and other spatial calculations were necessary. To do so, I used a Geographic Information System (GIS) implemented in the spatial relational database software PostGIS.[22] I measured the geodesic distance between each precinct and each of the housing projects. I also merged the precincts with geospatial boundaries of Census Block Groups to estimate income by race in each precinct. Income by race is available in the Decennial Census at the Census Block Group level and Block Groups are usually not coterminous with precincts, which are usually formed from a collection of Census Blocks. Thus, income was estimated through a process of spatial interpolation. This interpolation also involved the removal of the spatial footprint of geographic features with no population, such as parks and bodies of water, so that these features would not be counted in the interpolation.

[22] The spatial analysis in this article was all done in SQL in order to take advantage of PostGIS, a spatial database extender for PostgreSQL. At the time I started this project in 2005, a relational database was preferred, if not necessary, for doing spatial computation on the amount of data used in the article. Given current computational power, one could now probably perform the spatial computation using the software R.

The voting analysis reported in the article did not make use of the complete set of precincts that cast votes in these elections and so the replication files also do not include the complete set of precincts.[23] It is also the case that in the replication data, each election in 2004 and 2008 contains the same number of precincts (N = 1,866) and the elections in 1996 and 2000 also contain the same number of precincts (N = 2,402). Also in the replication data, the reported vote totals for some precincts do not match the totals in official counts from the Chicago Board of Election Commissioners. The counts of precincts and votes by precincts in official results can be see here: https://chicagoelections.gov/en/election-results.html.

The reason the analysis did not use the complete set of precincts is because the geographic boundaries of these precincts change across elections and only precincts with constant geographic boundaries or boundaries which could be reconstructed through aggregation (see below) were used. This is necessary for accuracy in the geographic measurements described above. If a researcher did not account for these boundary changes, then the geographic measurements would be wrong. The election in 2004 and 2008 and in 1996 and 2000 contain the same number of precincts, because the precincts change across elections (see below), it is analytically desirable to only use precincts that stay constant across all elections in the analysis, thus keeping a balanced panel, meaning that each year of the analysis uses the same data. If a researcher did not do this, the analysis would not be comparing the same places across time and this would make the over-time comparisons unsound.

To understand this, consider that geographic measurements require GIS shapefiles of both the precincts and housing projects. Consider the shapefile of Chicago precincts available at the Federal Open Data portal: https://catalog.data.gov/dataset/precincts-2010. According to the portal website, this data was used in 2010 and contains 2,576 precincts. But, for example, the data from the Chicago Board of Election Commissioners for the 2004 General election lists 2,709 precincts. The 2004 Senate Democratic primary lists 2,706 precincts. The 2008 Presidential Democratic Primary lists 2,579 precincts. The 2000 General election lists 2,542 precincts. The number of precincts also varies across other elections. These numbers vary because the geographic boundaries of certain precincts change from election to election – so, a precinct labeled as, say, Ward 1, Precinct 1 in 2010 does not have the same geographic boundary as the precinct labeled as Ward 1, Precinct 1 in 2004. For this reason, if a researcher were to naively download shapefiles, say by going to this website: https://catalog.data.gov/dataset/precincts-2010 and merge those with election results from an earlier election, say the November 2004 General Election, a map to display vote outcomes by precinct that was produced this merge would be inaccurate and potentially misleading, not only because the number of precincts in the tabular data from the Election Commissioners and the shapefiles would not be the same, thus precincts would be missing after the merge, but many of the precinct definitions would have changed and so the map would have many precincts in the wrong location. If a researcher were to use those precinct locations for measures of distance from the housing projects, those distance measures would, in many cases, be wrong. An interested party could examine this further by comparing the precinct listings across elections on the Chicago Board of Election Commissioners website.

That precinct boundaries change is well-understood by scholars of American elections. Precincts are usually constructed by election authorities to create boundaries that define a unique election ballot. So,

[23] Although the replication files do include the complete set of precincts on which the ecological estimation was performed prior to the matching. As described in the original article, because precincts closer to the housing projects were matched with those further away, the inferences are based on a limited set of these precincts.

every voter in a single precinct should be voting on a certain set of offices, say a particular Congressional District, another State Senate District, and another Aldermanic district. A person in another precinct would be voting in a different set of districts. These district boundaries are constructed using Census data, usually at the Census Block level, to balance population across districts. These boundaries may change over time for many reasons, but the most dramatic changes occur after the Decennial Census provides new population counts in the Census Blocks. After a Decennial Census, nearly every election authority in the United States will draw new precinct boundaries to go along with these new district boundaries. So, for example, most election authorities will entirely change precinct definitions between 2020, the last election year under the previous Decennial Census conducted in 2020 and 2022, the first election year after the Census data collected in 2020 is released. In the elections used in my article, Chicago changed nearly all its precinct boundaries in between the 2000 and 2004 elections, after the 2000 Decennial Census. Precinct boundaries may change for other reasons between these Decennial Census data releases, including changes in voter registration not reflected in the Census and other administrative reasons, for example some elections having higher turnout than others. Chicago undertook precinct changes between nearly every election from 1996 to 2008.

For the geospatial measurements in the article, I used the available shapefiles as a starting point, then I undertook a process to account for the boundary changes described above. This process was done separately for the 2004 and 2008 elections and for the 1996 and 2000 elections. In each case, the process used files obtained from the Chicago Board of Election Commissioners listing precinct reassignments (I make these files and the SQL code used to process them available here[24]). Working backwards from 2010 for the 2008 and 2004 elections, and from 2000 for the 2000 and 1996 elections, I used these reassignment files to trace precinct changes – dropping precincts that were reassigned in parts and merging precincts that were reassigned in whole: First, if an entire precinct was reassigned to a new precinct in a subsequent election, the precinct name was transferred across data sets. This allowed for the boundary of the precinct available in GIS shapefiles to be used for geospatial measurements of the reassigned precinct. In some cases, the Election Commission created a new precinct by collapsing multiple precincts (for example, Precincts 6 and 35 in Ward 1 were both collapsed into Precinct 1 in Ward 35 between October 1999 and April 2002). In these cases, for my analysis, the previous precinct votes could be added to create a new precinct that matched the aggregated vote totals of the older precincts. It is for these two reasons that some precinct totals in the official data do not match the totals reported in the official counts. Second, in other cases, precincts were reassigned only in parts, so a section of the geography of a previous precinct was removed from one precinct and reassigned to a new precinct (these are indicated by the "pt" designation in the reassignment files). In this case, because it is not possible to know the vote totals in these partial precincts, the precincts that were reassigned in parts were dropped altogether. It is for this reason that the number of precincts used in the analysis in the article does not match the number in official returns. In order to maintain a balanced panel, meaning that the same precincts were used in each election in the analysis,[25] precincts that were dropped for one election were also dropped for the other elections analyzed.

[24] Note that these files have not been made ready for a replication archive, so they contain my stray code and comments. Also be warned that I would not claim to be an elegant or efficient SQL programmer.

[25] This only refers to balance across the set of elections pre and post redistricting, so the election in 1996 and 2000 have the same number of precincts and the elections in 2004 and 2008 have the same number of precincts.

Does this censoring of partial precincts result in biased inferences? To create bias, consider that censoring would have to be correlated with changes in voter preferences, changes in these differences pre and post demolition, and distance from both the demolished and non-demolished projects, in addition to race. Because the changes to precinct definitions are based on geography, some spatial clustering in the censoring may exist, but that the censoring would simultaneously be correlated with all these other changes and, in particular, with distance from the demolished and non-demolished housing projects seems unlikely. One way to test for bias would be by replicating the analysis using all the precincts from each election, but when doing so, the spatial calculations on a large portion of the precincts would be wrong, so this would not be advisable.