PA

A Method to Audit the Assignment of Registered Voters to Districts and Precincts

Brian Amos[1] and Michael P. McDonald[2]

[1] *Assistant Professor, Wichita State University, Wichita, KS 67260, USA. Email:* brian.amos@wichita.edu
[2] *Associate Professor, University of Florida, Gainesville, FL 32611, USA. Email:* michael.mcdonald@ufl.edu

Abstract

Electoral boundaries are an integral part of election administration. District boundaries delineate which legislative election voters are eligible to participate in, and precinct boundaries identify, in many localities, where voters cast in-person ballots on Election Day. Election officials are tasked with resolving a tremendously large number of intersections of registered voters with overlapping electoral boundaries. Any large-scale data project is susceptible to errors, and this task is no exception. In two recent close elections, these errors were consequential to the outcome. To address this problem, we describe a method to audit the assignment of registered voters to districts. We apply the methodology to Florida's voter registration file to identify thousands of registered voters assigned to the wrong state House district, many of which local election officials have verified and rectified. We discuss how election officials can best use this technique to detect registered voters assigned to the wrong electoral boundary.

Keywords: geocoding, geographic information systems, data analysis algorithms

1 Introduction

Nearly all United States legislators, at all levels of government, are elected to office by the people residing in a defined geographic area known as a district. Periodically, district boundaries are redrawn through a political process known as redistricting. Precinct boundaries, which determine at which polling place many voters will cast an Election Day ballot, are drawn through an administrative process known as reprecincting, and a change in precinct lines is known to affect individuals' voter turnout (Haspel and Knotts 2005; Amos, Smith and Ste. Claire 2017) and mode of voting (Brady and McNulty 2011).[1] Redistricting politics are intense, and we do not wish to rehash the extensive scholarship. Suffice to say that epic political battles are waged to determine exactly where district boundaries should lie (Bullock 2010), and these choices may affect voting behavior (Hayes and McKee 2009). We are primarily interested here in the nuts and bolts election administration task of assigning voters to electoral boundaries—districts and precincts—so that they can receive a ballot for the offices they are intended to vote in (Herrnson, Hanmer and Niemi 2012).

Scholars have long recognized that ballots are the principal "point of contact between the average voter and h[er] government" (Beard 1909, 590). Ensuring that election officials give voters the correct ballot appears to be an easy task. However, two recent elections demonstrate that consequential administrative errors happen. In June 2018, dozens of voters in Habersham County, Georgia received a letter from their county's Office of Elections and Registration informing them that they had been assigned to the incorrect state House district due to "a past voting precinct redistricting issue" (Purcell 2018). In May, the House district in question had held its primary election for the 2018 cycle, and the Republican primary was decided by just 67 votes. The losing candidate challenged the results, and a judge ordered a re-vote.

Political Analysis (2020)
vol. 28:356–371
DOI: 10.1017/pan.2019.44

Published
17 January 2020

Corresponding author
Brian Amos

Edited by
Jeff Gill

Authors' note: The replication materials for this paper can be found at Amos (2019).

[1] We use the term "precincts" to refer to what the Census Bureau called "Voting Tabulating Districts" or VTDs, which are their generic term for the various names that states call precincts, wards, election districts, and so on.

In the 2017 Virginia general election, at least 384 registered voters in the northeastern part of the state were assigned to incorrect state House of Delegates districts, of whom at least 147 cast a ballot (Virginia State Board of Elections 2017, 3). One hundred and twenty-five of these were voters incorrectly assigned to House District 28, a number greater than the Republican candidate's 82-vote margin of victory. The elections to the chamber as a whole resulted in a near tie, so these errors affected not only who would represent this district but also partisan control of the Virginia House of Delegates. Despite the error, a federal judge declined to overturn the results, and an appeal to the ruling failed (Weiner 2018).

From a naïve viewpoint, election officials should easily determine which district voters' addresses are located in. In practice, election officials use data-driven representations of a jurisdiction's geography to manage the scale of assigning thousands of voters to the many overlapping districts and precincts in their jurisdiction. This approach introduces two types of errors, in addition to mundane clerical error. First, a particular map may be inconsistent with the real world: for instance, the map could place an address at the wrong location or it could be outdated, such as omitting a new residential subdivision. Second, electronic maps used to locate addresses and to delineate district lines may be the product of different sources and processes and, therefore, be inconsistent with each other.

We refer to this phenomenon as *administrative redistricting*, whereby election officials, in effect, assign registered voters into a legislative district different than the one the law assigns them to. These errors tend to receive attention only in extremely close elections where losing candidates have the resources to detect its presence. There is, thus, a good reason to suspect that the problem occurs elsewhere. To shed light on the scope of the problem, we develop a methodology to detect administrative redistricting by overlaying geocoded voter registration addresses onto district boundaries. When we apply the methodology in Colorado and Florida, we find thousands of registered voters assigned to the wrong state legislative district. We communicated our findings to local election officials, who verified and corrected the district assignment errors. In the course of our communications, election officials provided to us insights as to the processes that result in these problems.

As mentioned, we do not wish to dwell on the political process of redistricting; we wish to focus on election administration procedures of assigning registered voters to the districts they are eligible to vote within. Nonetheless, to explain how these errors occur, we review the two ways by which districts are defined in the course of redistricting and how voters are assigned to districts. Our review dwells, at times, in esoteric legal and geographic minutia, but it is precisely these details that can cause administrative redistricting. We conclude with lessons we learned from geocoding voter registration records and recommendations for improving the integrity of elections by auditing the assignment of registered voters to electoral boundaries.

2 Defining Electoral Boundaries

Historically, district boundaries are defined by what are known as *metes and bounds*. Metes and bounds describe boundaries by geographic and political features, such as roads, water, city boundaries, and so on. Consider the 2012 New York State Assembly districts, which are described in typical metes and bounds language:

> 1. First district. In the county of Suffolk, that part of the town of Brookhaven bounded by a line described as follows: Beginning at the intersection of Edwards Avenue South and the Brookhaven/South Hampton town line, thence southwesterly along said line to a line extended easterly from the East Moriches census designated place to the Brookhaven/South Hampton town line, thence westerly along said line . . . (New York Consolidated Laws 2012).

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press

The transition from governments using metes and bounds to using maps started in 1890 when the Census Bureau began delineating and reporting decennial population statistics for selected small geographic areas in New York City (United States Census Bureau n.d.). Over time, the Census Bureau extended their coverage by mapping smaller geographies, across a greater range of the country. By 1990, the Census Bureau had mapped the entire country down to what are commonly known as census blocks, which are geographic units similar to city street blocks in urban areas, but may follow other geographic features, such as road medians, streams, and railroads. Census cartographers may also use political boundaries to delineate census block boundaries, working with states and localities to collect and verify the location of these political boundaries. Census blocks can be quite small, and a typical medium-sized state has hundreds of thousands of them.

As the Census Bureau canvassed the nation to create localized cartography, some states abandoned metes and bounds in favor of defining their electoral boundaries by *census geography*. Doing so ensures that legislative districts meet legal equal population requirements since the Census Bureau reports decennial census total population statistics within their defined geographic units. A district defined in terms of census geography will list the census units assigned to each district, such as counties, census places, tracts, block groups, and blocks. These legal definitions typically take the form of a list of unique identifiers identifying each piece of census geography assigned to a district, what is sometimes referred to as a block equivalency file.

Metes and bounds may coincide with census geography, even if census geography is not explicitly mentioned. The New York Assembly metes and bounds uses some census units, such as the “East Moriches census designated place” (census places are a component of the census geographic hierarchy). Indeed, it is likely that the roads, town boundaries, and other features used in New York’s metes and bounds coincide with census geography since Census Bureau cartographers often use physical features and political boundaries to define census geography. Some geographic information systems (GIS) enable consultants to draw maps using census geography as the base mapping layer and export district descriptions in metes and bounds language.

In theory, census geographic boundaries could align with physical and political features. In practice, errors exist when initial census maps are not aligned perfectly with the real world or when real world features change. The Census periodically releases updates to its geography to reflect the addition of newly built roadways, diversion of existing streets, changes to political boundaries, and to correct the virtual locations of streets and other features. As we shall see, these updates can have occasional substantive implications by effectively moving the virtual representation of district boundaries and registrants’ addresses.

3 Assigning Voters to Districts

When a registered voter votes, election officials must provide a ballot listing of all the elections the individual is eligible to participate in. Election officials create all of the permutations of needed ballots, which are known as ballot styles. The largest localities may have hundreds of ballot styles for the hundreds of thousands of voters residing in a myriad of overlapping congressional, state legislative, judicial, and local government districts. Managing this workflow is labor intensive, which is why election officials rely on data-driven solutions.

To manage the assignment of registered voters to districts, so that they can be given the correct ballot style, most election officials append district identifiers to each individual voter registration record. Election officials employ two types of procedures to make these assignments. In the first procedure, election officials create a master address database that lists the valid street address ranges associated with each district. For example, in Figure 1, we present a hypothetical street segment assignment as a map where the dark houses are assigned to District 1 and the light houses are assigned to District 2. The district boundary line is represented by a dotted line. A typical

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press

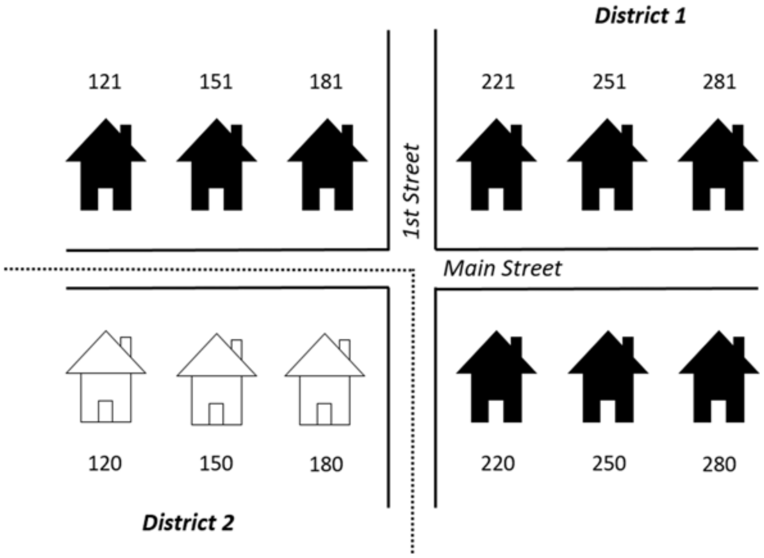

Figure 1. Depiction of street segment district assignments. *Notes:* The dotted line represents the boundary between District 1 and District 2. Homes are color-coded by which district they are assigned to.

master address database lists all known street segments falling within numbered block segments, with separate records for the even and odd numbered sides of streets. Identifying the even and odd sides of streets is necessary since streets often coincide with district boundaries. For example, the even 100–198 Main Street households are assigned to District 2, while the odd numbered 101–199 Main Street households are assigned to District 1. The even and odd addresses in the next 200–299 Main Street block are similarly designated to District 1, and so on. In the event a district line bisects the middle of a street block, the street ranges are segmented according to where the district boundary crosses the street, although where a district line lies exactly can be a source of district assignment error.

The other method to assign registrants to districts involves geocoding. Election officials may use geocoding procedures to pinpoint the latitude and longitude of addresses, and then overlay district boundaries onto these pinpoints to assign registered voters to districts. Geocoding algorithms generally work by corresponding an address with a database that contains latitude and longitude coordinates for addresses, roads, and political and administrative units. Census Bureau cartographers create an extensive system of geographic data in the course of their work commonly known as TIGER, which stands for Topologically Integrated Geographic Encoding and Referencing. Some Census Bureau databases are specifically designed for geocoding, and commercial vendors may supplement these data with other databases, such as local tax assessor data. When a geocoding algorithm encounters an address it cannot match with its pinpoint database, it uses heuristics to infer an approximate latitude and longitude coordinate. For example, if a geocoder does not recognize 150 Main Street, it may infer that the address lies halfway between the even side of the end point coordinates of 100 and 198 Main Street, and it will assume that the building's setback is a default distance perpendicular to the street. This attempt will fail if the street name is not in the geocoding database or a street number does not fall within existing valid street ranges. A geocoder heuristic will then attempt to infer what it can from higher geographic levels, such as zip codes and city names.

A geocoding heuristic may make a good guess of a building's location or it may make an error (Krieger *et al.* 2001; Cayo and Talbot 2003). The highest level of geocoding confidence is "rooftop" accuracy, where the address is contained in the geocoder's (hopefully correct) database of associated latitude and longitude coordinates. Two sources of error may arise if a geocoder must

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press

infer latitude and longitude of an address by interpolating between start and end coordinates of a street or any available street midpoints. First, in our example, 150 Main Street may not fall directly in the middle of the 100 and 198 Main Street line segment, despite what the numbers may suggest. Second, the geocoder must guess at the setback distance from the street to the building. Most geocoders, by default, apply a standard, nonzero setback to all addresses so that buildings will not fall directly on roads. If a driveway is long or does not run perpendicular to a road, this inferential method may locate an address at the wrong location. If the geocoder cannot match the street address at all, the algorithms typically rely solely on the center point of the city or zip code geography, with a potentially high error on the proper location of the address.

Neither master address databases nor geocoding is a fool-proof method of assigning registered voters to districts. When election officials encounter a problematic registration application, they may place an application into a suspense status while they conduct additional research. Most commonly with respect to address errors, applications are designated with a suspense status because election officials do not recognize a street address, as may occur with new construction, a spelling error by an applicant, or illegible handwriting. Some addresses might not commonly be recognized as a valid street address, such as a rural address that is merely a description of the property or an urban address that identifies a park or alley where a homeless person resides. Native-American reservations can be problematic when residents' homes are associated with their mailbox, located miles from the physical location of their home, an issue that affected North Dakotan Native Americans who were required to present voter identification cards with a valid street address in order to vote in the 2018 election (Ortiz 2018). When election officials encounter these issues, they may contact the applicants to resolve indeterminate addresses.

4 Methodology

We audit the assignment of registered voters to state legislative districts in Colorado and Florida, using a December 2017 snapshot of the Colorado voter file and a November 2017 snapshot of the Florida voter file. We focus primarily on these two states because we worked with election officials to verify our methodology. Additionally, to illustrate geocoder performance, we report geocoding statistics for a June 2017 Louisiana voter file, an August 2017 North Carolina voter file, a June 2016 New York voter file, and a September 2017 Ohio voter file.

Our methodology to detect when registered voters are assigned to the wrong district is similar to the geocoding approach used by some election officials to assign their registered voters to districts, described above. We geocode voter registration addresses, overlay district boundaries, and identify instances where a voter registration district identifier is not the same as the overlaid district. An innovation that distinguishes our approach from election officials' approach is that we use multiple geocoding databases in sequence, two based on direct address matching, two based on interpolating street addresses, and several at less precise levels of detail. As described in the accompanying replication materials, we perform our detection methodology using ERSI's ArcGIS software for our geocoding and Python scripting for our analyses, which makes it possible for us to implement the procedure on any voter file with minimal programming effort, although the task of geocoding a large state's voter registration file is computationally intensive and can take several days.

Our approach works well as a check on election officials' street segment approach to assign voters to districts since we can independently verify district assignments using an alternative methodology. Furthermore, our methodology has detected voters assigned to the incorrect district even in localities where election officials use a geocoding system to assign registrants to districts. This may happen when a locality uses a geocoding database different from the ones we use, a situation we confirmed during a discussion with a local election official who verified that

Table 1. Percentage of addresses matched by our sequential geocoding approach.

Database	CO (%)	FL (%)	LA (%)	NY (%)	NC (%)	OH (%)
2016 point address	75.30	85.74	68.83	81.76	81.36	85.72
2015 point address	7.07	2.05	1.97	1.91	2.75	3.70
Highest precision subtotal	*82.38*	*87.79*	*70.80*	*83.67*	*84.11*	*89.41*
ESRI street address	14.42	9.90	25.29	14.38	12.76	8.51
Census street address	1.04	0.76	1.69	0.45	1.12	0.82
High precision subtotal	*97.84*	*98.46*	*97.78*	*98.51*	*97.99*	*98.74*
Street name	0.92	0.34	0.67	0.26	0.59	0.33
ZIP code	1.24	0.70	1.55	1.23	1.42	0.93
Places	0.00	0.01	0.00	0.00	0.00	0.00
No match	0.00	0.49	0.00	0.00	0.00	0.00
Total	*100.00*	*100.00*	*100.00*	*100.00*	*100.00*	*100.00*

their geocoder had misplaced an apartment complex. The specific geocoders we use, in order, are the following:[2]

- Point address database from the 2016 ESRI Business Analyst data.
- Point address database from the 2015 ESRI Business Analyst data.
- Street address interpolation database from the 2016 ESRI Business Analyst data.
- Street address interpolation database built from the 2016 Census Address Feature shapefile.
- Street name database from the 2016 ESRI Business Analyst data.
- ZIP code database from the 2016 ESRI Business Analyst data.
- Place name database from the 2016 ESRI Business Analyst data.

Using ESRI's ArcMap software, our geocoding process attempts to find a match in the first listed database. If a match is found, the process halts and reports the latitude and longitude given by that database; else, the next database on the list is attempted. The process continues down the list until a match is found.

We expect different performance from different companies' geocoding databases, simply because these are generally proprietary data collections that companies spend considerable effort to develop and sell for profit. We infer from documentation that vendors typically begin geocoding database development with the Census Bureau's TIGER products. Companies supplement the TIGER databases with other data, such as local tax assessor and surveyor maps, the Post Office's National Change of Address files, and information from consumer credit monitoring agencies. Companies may use different databases, different algorithms to merge records, and have different timing as to when they collect data. This results in companies producing different geocoding databases.

To demonstrate the performance of our approach, we run the geocoding process, described above, on Colorado, Florida, Louisiana, New York, North Carolina, and Ohio voter registration file snapshots. The percentage matched by each geocoder is given in Table 1. We consider the first two geocoders, which have latitudes and longitudes for specific addresses, to be of the highest precision and consider street address interpolation to be less precise, though still of generally high quality. We achieve matches in the point address databases in 80%–90% of cases in every state except Louisiana and high precision across all states in over 97% of cases. We disregard the few cases that fall to the low-precision geocoders because of their questionable accuracy.

[2] None of the databases we use provides rooftop accuracy since access to these types of services at the scale of a statewide voter registration file—much less multiple statewide files—was prohibitively expensive for this project. We crosscheck for potential false positives using Google's Geocoder API, which does provide rooftop accuracy, later in this paper.

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press

Table 2. Comparison of geocoding performance on the Colorado voter registration file.

	Highest precision		**High precision**	
	2016 ESRI point	**2015 ESRI point**	**ESRI street address**	**Census street address**
2016 ESRI point	*2,811,823*	150,453	56,673	109,520
2015 ESRI point	263,272	*2,924,642*	59,290	103,363
ESRI street address	787,097	676,895	*3,542,247*	181,033
Census street address	757,342	638,366	98,431	*3,459,645*

Florida is the only state with more than a handful of entries that produced no match at all; the vast majority of these cases are voters who qualified for removal of certain data (including address data) from public information requests, such as some public employees and persons in confidentiality programs due to domestic violence or stalking cases.

Table 2 illustrates for Colorado's 3,670,396 registered voters how different geocoding approaches produce different results. The diagonal number represents the number of addresses geocoded by a geocoding database. The off-diagonal entries represent the number of addresses the row database geocoded that the column database did not. For example, the 2016 ESRI point address database geocoded 2.81 million addresses, while the 2015 database was able to geocode 2.92 million. The 2016 database successfully geocoded in 150,453 addresses that the 2015 database could not, and the 2015 data successfully geocoded 263,272 addresses that the 2016 could not. In sum, the 2015 database successfully geocoded a net 112,819 more addresses than the 2016 database. The two highest precision geocoding procedures, which contain point address entries, produce fewer successful geocodes than the high precision procedures. This is to be expected since the high precision algorithms will guess at the location of street addresses that do not appear in the highest precision point databases.

We observe in Table 2 that we geocode more records by layering legacy geocoding databases. Even different vintages of geocoding databases from the same company result in more successfully geocoded addresses. We attribute this phenomenon to occasional changes to road networks, such as when a street is renamed or when street addresses are renumbered. Voter registration addresses may become out of sync with commercial vendors' geocoding databases when voter registration databases are not updated as quickly as commercial vendors' geocoding databases. A legacy geocoding database may thus geocode addresses where the most current database will not. We have also observed instances where a voter registration database has more recent addresses than a geocoding database, which may occur when occupants of a newly built home register to vote but the Census Bureau or commercial vendors have not yet incorporated the new home into their geocoding databases.

One potential concern in using different geocoding databases is the order in which we apply them. In our specific example, there is an *a priori* logic in using point address databases before interpolation-based databases and the 2016 point address database before the 2015. However, were multiple point address databases sourced from different vendors available, the question of which should be applied first is less obvious. One approach would be to apply all addresses to all databases and to investigate those addresses that were flagged as a misassignment in any of them. However, in comparing the geocodes in Colorado across databases, we find a relatively small percentage of differences where addresses were assigned. Unsurprisingly, well over 99% of geocodes matched by both the 2015 and 2016 ESRI point databases were placed within one-tenth of a mile of each other. When we select a random sample of 500 geocodes made by the 2016 ESRI point address database and feed them through the Google Geocoder API, 96% of those Google identifies with rooftop accuracy fall within one-tenth of a mile of the ESRI placement and 99% fall within one-quarter of a mile. Similarly, the Census and ESRI street address databases fall in the same range: 95% and 97% of addresses fall within one-tenth of a mile of the 2016 ESRI point address database assignment, respectively.[3] Thus, it appears that the larger difference between databases is in the addresses included, rather than where included addresses are geocoded.

Once addresses are represented as latitude and longitude points on a map, we geo-spatially merge these data with state legislative district boundaries, obtained from the Census Bureau. This simple operation reports where two different geographies coincide. We then identify registered voters potentially assigned to the wrong district by comparing the district assignment found in the voter registration file with the assignment inferred from the geo-spatial merge. Not all of the addresses we identify by this method may be true assignment errors. For example, when a rural address is located using the street interpolation method, the actual setback location from a road can vary greatly from where a geocoder's default setback algorithm places it. We thus have greater confidence when we identify clusters of addresses with potential assignment errors than with isolated cases.

5 Administrative Redistricting

We presented election officials with some of the suspected errors identified by our methodology, and they validated and rectified them. In the course of our work, we identify a number of ways by which administrative redistricting happens. Some involve human error, some involve overreliance on geocoding algorithms, some involve obscure updates to census geography, and some involve the district lines. Our list may not be exhaustive since there may be other causal factors leading to district assignments we have not encountered. We wish to emphasize that we have not detected situations that appear to be malfeasance to intentionally override the legally specified district boundaries. We therefore obscure the sources of the following examples since election officials have been generously responsive to our communications, and we mutually seek to improve election integrity.

5.1 Data Entry Error

This mundane error occurs when human operators make data entry errors into election management databases. A frequent example of this error involves databases of street address segments. A data entry error in a street segment database creates district assignment errors for an entire street segment, which is easily observed when affected residences are overlaid on satellite imagery maps. Figure 2 depicts such a scenario. The thick black line represents the boundary between District A and District B. The cross symbols represent geocoded addresses assigned to District B in the voter file, while the dashed gray line represents one particular street that spans across District B into District A. As can be seen, a number of voters who are geographically in District A are assigned to District B. Furthermore, all of these incorrect assignments fall on one side of the street—the odd side, in this case—while those falling on the even side were correctly assigned, signaling a likely street segment error.

Figure 3 depicts a more substantial data entry error resulting in over 800 registered voters assigned to the wrong state House district. These assignment errors occurred during a local redistricting, whereby registered voters within a precinct split by new local districts had their state House district changed to match the new city commission district lines. The map on the left in Figure 3 shows the assignments in a snapshot of the voter registration file before the city commission redistricting and the one on the right is after. The black line represents a state legislative boundary, and X's and O's represent geocoded addresses based on which district they are assigned to in the voter registration file. Though the state House district did not change, the assignments did; the changes follow those made in the city commission districts. The assignment

[3] Replication data for this paper are available on the *Political Analysis* Dataverse (Amos 2019).

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press

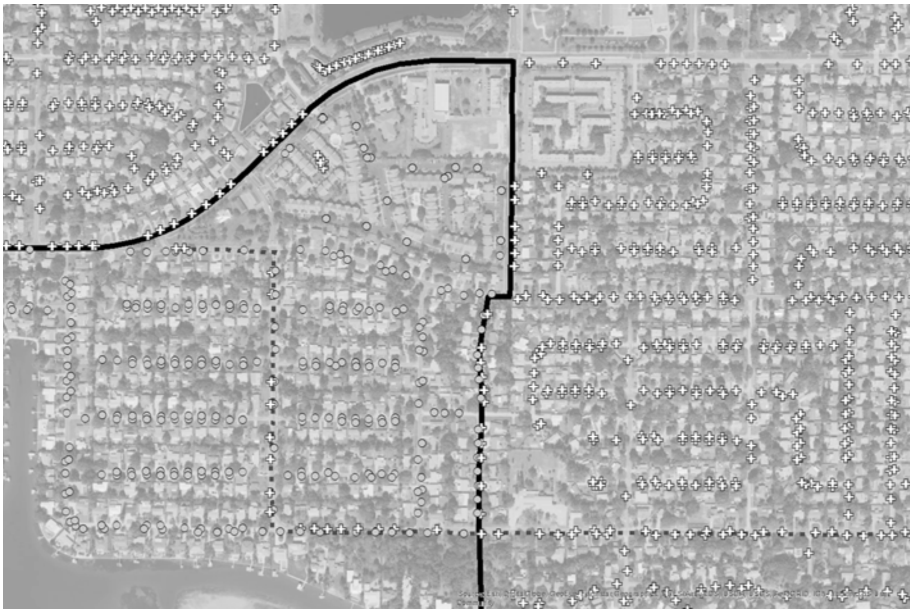

Figure 2. Example of street segment error.

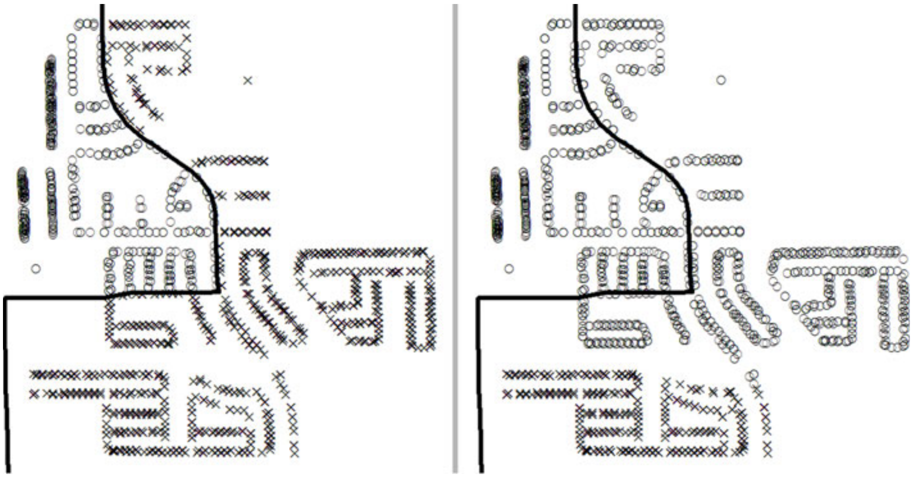

Figure 3. Example of human operator error. Left: district assignment prior to a local redistricting. Right: district assignment after a local redistricting.

error persisted across several elections following the local redistricting until we alerted local election officials of the error.

5.2 Geocoding Error

Some election officials use geocoding as a component of their process to assign voters to districts, similar to our methodology. Geocoders can fail to correctly locate an address's longitude and latitude. While mundane data entry error in a geocoding database can cause such errors, a more prevalent issue occurs when a geocoder makes an algorithmic guess as to where a home is located.

Figure 4 depicts setback issues that are prevalent in rural areas. The white line represents a district boundary that follows a small waterway alongside a road. Local election officials located the circled home's address along the road, to the right of the district boundary, whereas the true

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press

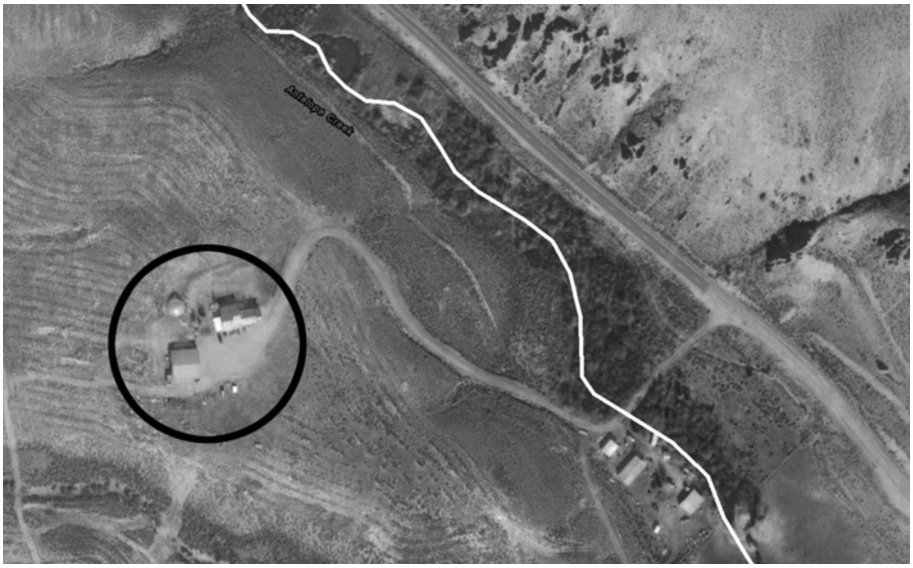

Figure 4. Example of a setback gap error.

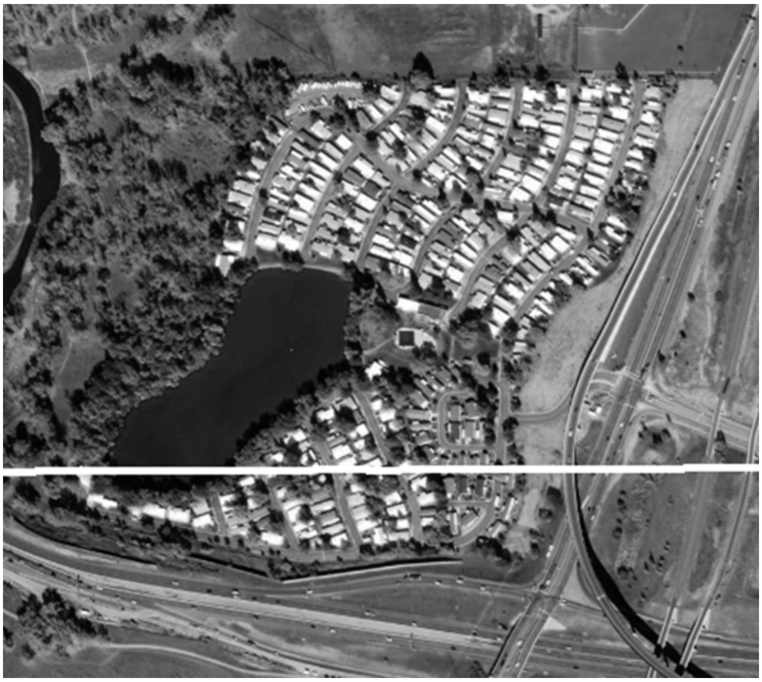

Figure 5. Example of a geocoding error.

location of the home is set back far from the road along a winding driveway. One of our geocoding databases with rooftop accuracy is able to place this household correctly.

Another form of geocoding error is depicted in Figure 5. The pictured area is a mobile home community which is cut by a district boundary, represented in white. All homes in the community share the same first-level address, which is located to the south of the district line. The homes are distinguished from one another by something akin to an apartment number in a second-level address. Election officials apparently placed all of the homes at the first-level address, thereby assigning many homes to the wrong district.

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press

5.3 District Boundary Error

We have focused on geocoding errors, but it is possible that the electronic representation of district boundaries are in error. Such was the case in Colorado. When we presented state election officials with suspected assignment errors, local election officials responded that many were caused by inaccurate district boundary lines. Local election officials had won a state court ruling that ordered all state legislative districts that split precincts to follow precinct boundaries. Remarkably, neither state officials nor the Census Bureau were notified of the court ruling, and, thus, the publicly available district boundaries were incorrect and have since been corrected.

5.4 Geography Updates

What perhaps comes closest to intentional administrative redistricting occurs when district boundary lines are adjusted to reflect realignment of roads, political boundaries, or other geographic features. When a district border is described in metes and bounds, changes to features that comprise a district boundary effectively legally change the district's lines. Census Bureau cartographers may update census geography to reflect these changes, too, which can likewise affect district boundaries. State and local governments may take action independent of the Census Bureau. In a personal communication with a state's chief geographic information officer, we learned that the state's policy in these situations is to modify census block boundaries in an internal state database. If such changes encompass residences, people are moved into a new district.

5.5 Asynchronous Data

Census Bureau cartographers update continually the TIGER geography to account for changes to physical and political boundaries. These updates also happen in reverse when geographers discover that prior versions of their virtual maps do not correspond well to the intended physical and political features. These corrections to census geography may not alone cause district assignment errors. However, administrative redistricting may occur if the legal district boundaries and the geocoding databases become temporally out of sync. We present such a scenario in Figure 6. This county uses a geocoding approach to assign voters to districts that accurately located voters' residences using the 2017 TIGER data but defined the district boundaries in circa 2013 TIGER data. We represent the same state legislative boundary as defined by census block geography in the 2013 TIGER data (white) and 2017 TIGER data (black), between which a correction was made. Where these two lines diverge, registered voters were assigned to the wrong district, represented by white dots.

6 Counting Errors and Disparate Effects

Our final step in detecting errors is to join our geocoded addresses to Census district shapefiles and to identify registrants whose listed district in the voter registration file differs from the one we determine through our method. We carry this process out for the Colorado House of Representatives, Colorado Senate, Florida House of Representatives, and Florida Senate (Amos 2019). Table 3 presents summary statistics of voters we identified as being assigned to the wrong district with the highest precision geocoding procedures.

There are two ways to categorize a potential error in assignment. The first is to report the raw count of voters who our method places in a different district than the one listed in the voter registration file; these counts are given in the first column of Table 3, along with the share this represents of the total count of registered voters in the second column. As seen in the previous section, however, a single error can lead to many affected voters. To reframe the error, we cluster probable assignment errors based on whether their addresses are within a tenth of a mile of each other. The cases presented in Figures 2 and 3, therefore, would each be counted as a single cluster.

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press

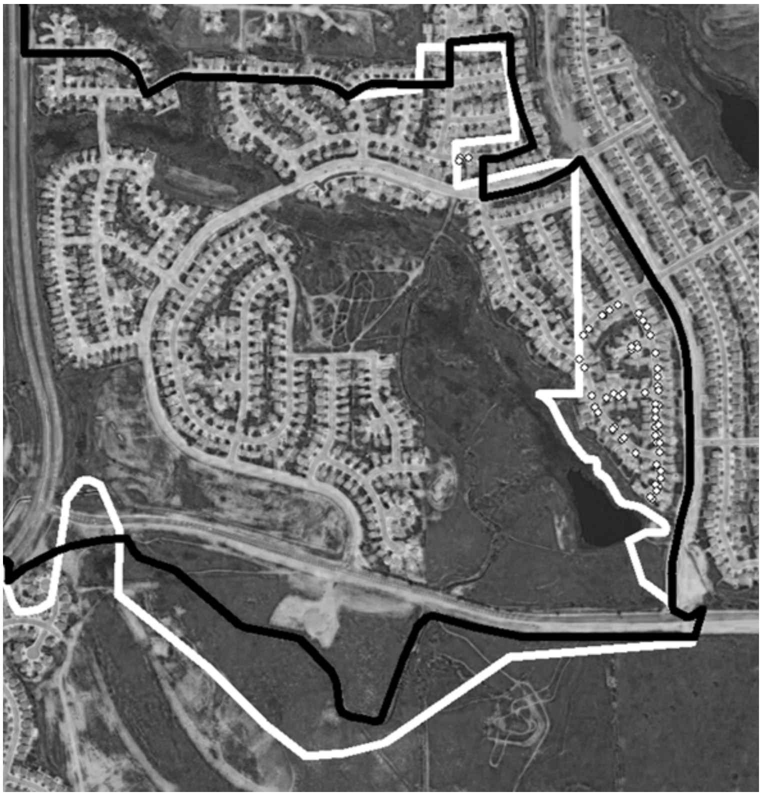

Figure 6. Example of census geography changes resulting in assignment error.

Table 3. Summary statistics for suspected registered voters assigned to wrong district.

	Number of assignment errors	Percentage of registered voters (%)	Number of clusters	Percentage of affected urban registered voters (%)	Largest cluster size	Percentage of clusters with one or two addresses (%)
Colorado House	998	0.027	107	65.03	391	82
Colorado Senate	556	0.015	50	71.04	136	74
Florida House	1,261	0.013	219	85.80	552	84
Florida Senate	2,520	0.027	246	89.37	1,355	84

Figure 6 has two clusters since the lines diverge around two distinct population centers. In each of the four maps we examined, at least one cluster had more than 100 registered voters, and at least two-thirds of clusters in each map represented just one or two unique addresses.

As a check for potential false positives, we geocoded the addresses identified by ESRI's point address database described in Table 3 using Google's Geocoding API. Google's API uses a similar approach as our method, in that it first attempts to find a match at the highest level of precision and then attempts an interpolation heuristic. Google has an extensive rooftop-accurate database. While ESRI point address databases do not have as broad a scope as Google's, we identified in our investigations some addresses where Google's API used interpolation but where ESRI did not. We assume that when Google's results align with our ESRI-based results, the district assignment error is real. We investigate further when Google does not validate ESRI's results. In total, we find that 89.4% of the registered voters identified as district assignment errors using ESRI's database

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press

are verified with Google's API, with a tendency toward Google verifying larger clusters with greater frequency than smaller clusters.[4]

Extrapolating these numbers out to the rest of the country is difficult for two reasons. First, we cannot be certain that the two states' legislative chambers we examined are representative of the United States. Second, we do not know to what degree a registered voter is assigned incorrectly to districts at different levels of government. In the two state legislative chambers within the two states we examined, we find in Colorado that 334 voters were assigned to both incorrect state House and Senate districts, and in Florida, the number is 243. Taking this into account, the total error rate is 0.033% in Colorado and 0.037% in Florida. Assuming the U.S. Census Bureau's 2016 Current Population Survey estimate of 157.6 million registered voters is accurate, we estimate that nationally there are roughly 50,000–60,000 registered voters assigned to the wrong state legislative district. Taking into account U.S. House districts and local districts, as well as addresses matched with street address interpolation geocoders,[5] the total number of voters with any assignment error is likely more than 100,000. Since some larger errors tend to cluster, there is a reasonable chance that election outcomes have been affected by these assignment errors, in addition to those reported by the media in Georgia and Virginia. Indeed, the worst case we detected and election officials verified was just shy of potentially affecting the outcome of a primary election.

We can examine who appears most likely to be affected by administrative redistricting, though we again caution about generalizing our findings. Referring back to our previous results, there are competing factors that may contribute to errors disproportionately affecting different populations. First, the density of district lines scales with population density. A rural county may completely reside in a single district at the congressional and the state legislative levels, while an urban county may contain a dozen or more districts across levels of government. Thus, an urban county has many more opportunities for district assignment errors than a rural county. Second, the greater population density of urban areas means that a single error may affect more registered voters than in rural areas, such as an error that affects an entire apartment complex. Conversely, rural areas are less likely to follow neat street patterns of cities and suburbs, leading to a greater possibility that election officials place a residence in the wrong location and wrong district.

In Table 3, we report the percentage of registered voters for each chamber that fall into census-designated urban areas. Colorado's share of affected urban registered voters is 65% for the state House and 71% for the state Senate and is considerably more rural in character than the statewide percentage of urban population of 86%. Florida's share of affected urban registered voters is 86% for the state House and 89% for the state Senate, which is slightly less than the statewide share of the urban voting-age population of 91%. In both states, there is thus a tendency toward rural registered voters to be more often assigned to the wrong district. The differences between the states appear to be due partly to the distribution of cluster sizes: a larger share of Florida's errors fall in the small number of urban clusters that affect many voters. Colorado, on the other hand, has more that fall into the long "tail" of errors that affect few voters, which are disproportionately in rural areas.

Colorado and Florida are states with party registration, which permits us to examine partisan effects. We observe mixed effects. In Colorado, our snapshot of the statewide voter registration file

[4] The tendency toward larger clusters being more often independently verified by ESRI and Google is evident if the percentage of address clusters identified is smaller than the percentage of registered voters identified. For the Colorado House, 62% of clusters and 79% of voters were confirmed. For the Colorado Senate, these numbers were 70% and 78%, respectively; for the Florida House, 86% and 91%; and for the Florida Senate, 90% and 95%.

[5] Looking at assignment error counts for the four geocoders with high precision—both point address and address interpolation—we find 3,789 for the Colorado House, 3,118 for the Colorado Senate, 5,196 for the Florida House, and 4,946 for the Florida Senate. These counts represent a large increase over the numbers in Table 3, but we suspect they produce a higher rate of false positives.

is 30% registered Republican and 31% registered Democrat, while those we identify are assigned to the wrong district are 36% Republican and 28% Democrat for the state House and 31% Republican and 35% Democrat for the Senate. In Florida, the statewide party registration is 39% registered Republican and 38% registered Democrat, while those assigned to the wrong House district are 32% Republican and 45% Democrat and those assigned to the wrong Senate district are 40% Republican and 33% Democrat. We can also examine disparities by race in Florida since this information is recorded in the voter registration file. Statewide, registered voters are 67% non-Hispanic White, 13% non-Hispanic Black, and 15% Hispanic. The House assignment errors are 54% non-Hispanic White, 22% non-Hispanic Black, and 16% Hispanic; and the Senate assignment errors are 81% non-Hispanic White, 5% non-Hispanic Black, and 7% Hispanic. Thus, the Florida state Senate errors tend to affect Republicans (and more non-Hispanic White), like Colorado in the state House, but the Florida state House errors tend in the opposite direction, like Colorado's state Senate.

7 Discussion

Registered voters are occasionally assigned to the wrong districts, and these errors have led to real consequences in close elections, most recently in Georgia and Virginia. The problem is much more extensive than those two examples, though; we and the *Washington Post* (Mellnik, Fischer-Baum and Soffen 2018) have detected thousands more registered voters assigned to the wrong districts in just Colorado, Florida, and Virginia. Fortunately, none of these additional assignment errors occurred in elections close enough to affect election outcomes, but there is no reason to believe this is true for all other states and other levels of government.

Our investigations illuminate practical recommendations for scholars and election administrators who geocode addresses, particularly voter registration addresses. As geocoding technology has become more accessible, scholars are using geocoding in applications such as determining voter registrants' race (Imai and Khanna 2016), the turnout effects of distance to polling locations (Dyck and Gimpel 2005; Haspel and Knotts 2005; Gimpel, Dyck and Shaw 2006; Amos, Smith and Ste. Claire 2017), as well as numerous applications in areas as diverse as medicine (e.g., Krieger *et al.* 2003) and gambling addiction (Welte *et al.* 2004). Statistical estimations can be dependent on statistical software applications and platforms (Altman and McDonald 2003). Geocoding errors are not necessarily random. We observe this in the high frequency of addresses that geocoding algorithms fail to locate, particularly school dorms and military barracks. We cannot know the magnitude of induced biases, if any, without replication and forensics analyses of prior research. Given inaccuracies and inconsistences we have observed across the Census Bureau's and commercial vendors' geocoding databases, we believe these errors could threaten the validity of existing or future research. Geocoding algorithms typically produce reports for each processed address, which is how we generate the Table 1 statistics. These reports inform users how many addresses are geocoded successfully at each level of precision and permit users to investigate failure patterns. We recommend that at a minimum, scholars append these reports to their replication databases and include all records that fail to be geocoded and may otherwise be omitted from an analysis. Geocoding databases are fundamentally time-bound to reflect new development and changes to existing road networks. We thus agree with Whitsel and coauthors (2004) that to ensure full replicability of studies employing geocoding, scholars must note which vintage of geocoding databases they use. Ideally, to troubleshoot potential biases, replication archives would be created that include geocoding algorithms that parse addresses, the databases that provide rooftop precision coordinates, and the heuristic algorithms that estimate coordinates when rooftop precision is unavailable. We recognize that this is a significant challenge since commercial geocoding databases are often proprietary. The fundamental issue as to whether

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press

academic research should be dependent on proprietary data and methods is beyond the scope of our essay, but it is a discussion worth having.

We believe that election administrators, scholars, and campaign consultants will have the most success in geocoding voter registration addresses when they layer multiple geocoding databases, rather than use a single database (Ward *et al.* 2005). Occasionally, local governments change street names. Commercial vendors may assiduously update their geocoding databases to reflect these changes, but election officials may not do so with the same promptness. We thus discover that we successfully geocode more addresses when we use legacy geocoding databases rather than relying on the most current version. Scholars should consider employing more than one geocoding solution, at least as a diagnostic check on a single geocoding approach. Increasing the number of geocoding databases may be done with minimal costs, as the Census Bureau freely distributes their TIGER geocoding databases. However, some addresses do not appear in geocoding databases, such as campus dorms, military barracks, and Native-American reservations. A simple way to increase the coverage of a geocoding exercise is to check the frequency of addresses that are not geocoded and create an exceptions file to handle such cases. Three of our examples depicted in Figures 4–6 are illustrated through satellite imagery. We find that generating satellite imagery maps of suspected assignment errors is a powerful medium for troubleshooting.

We conducted our district assignment audits from an academic perspective and have engaged with election officials as white-hat outsiders who wish to improve election integrity. Election officials have been graciously responsive to our communications and have rectified the errors we drew their attention to. Assigning voters to their correct district is important from a normative standpoint, and with this in mind, we recommend that election administrators build district assignment auditing features into their election management systems and that they integrate audits into their election preparations. We have observed that election officials with the capability of auditing district assignments fail to use the tools available to them. Routinizing and automating district assignment audits can mitigate avoidable problems. Even where election administrators have internal auditing systems, we recommend they periodically externally audit their district assignments so as to not overly rely on one geocoding solution.

The good news is that we can audit district and precinct assignments. Technological innovations have progressed such that it is possible to develop and deploy these recommendations, and some vendors have already deployed such systems. We can develop mapping applications that will centralize election boundary collection for districts and precincts. The result of these efforts will be better election data integrity, which will improve voters' experiences, reduce election costs, and improve voters' confidence in the electoral system.

References

Altman, M., and M. P. McDonald. 2003. "Replication with Attention to Numerical Accuracy." *Political Analysis* 11(3):302–307.

Amos, B. 2019. "Replication Data for: A Method to Audit the Assignment of Registered Voters to Districts and Precincts." https://doi.org/10.7910/DVN/Y18MK5, Harvard Dataverse, V1.

Amos, B., D. Smith, and C. Ste. Claire. 2017. "Reprecincting and Voting Behavior." *Political Behavior* 39(1):133–156.

Beard, C. A. 1909. "The Ballot's Burden." *Political Science Quarterly* 24(4):589–614.

Brady, H., and J. E. McNulty. 2011. "Turning Out to Vote: The Costs of Finding and Getting to the Polling Place." *American Political Science Review* 105(1):115–134.

Bullock, C. S. 2010. *Redistricting: The Most Political Activity in America*. Lanham, MD: Rowman & Littlefield Publishers.

Cayo, M. R., and T. O. Talbot. 2003. "Positional Error in Automated Geocoding of Residential Addresses." *International Journal of Health Geographics* 2:10.

Dyck, J. J., and J. G. Gimpel. 2005. "Distance, Turnout, and the Convenience of Voting." *Social Science Quarterly* 86(3):531–548.

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press

Gimpel, J. G., J. J. Dyck, and D. R. Shaw. 2006. "Location, Knowledge and Time Pressures in the Spatial Structure of Convenience Voting." *Electoral Studies* 25(1):35–58.

Haspel, M. H., and G. Knotts. 2005. "Location, Location, Location: Precinct Placement and the Costs of Voting." *The Journal of Politics* 67(2):560–573.

Hayes, D., and S. C. McKee. 2009. "The Participatory Effects of Redistricting." *American Journal of Political Science* 53(4):1006–1023.

Herrnson, P. S., M. J. Hanmer, and R. G. Niemi. 2012. "The Impact of Ballot Type on Voter Errors." *American Journal of Political Science* 56(3):716–730.

Imai, K., and K. Khanna. 2016. "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records." *Political Analysis* 24(2):263–272.

Krieger, N., P. Waterman, K. Lemieux, S. Zieler, and J. Hogan. 2001. "Wrong Side of the Tracks? Evaluating the Accuracy of Geocoding in Public Health Research." *American Journal of Public Health* 91(7):1114–1116.

Krieger, N., P. D. Waterman, J. Y. Chen, M.-J. Soobader, and S. V. Subramanian. 2003. "Monitoring Socioeconomic Inequalities in Sexually Transmitted Infections, Tuberculosis, and Violence: Geocoding and Choice of Area-Based Socioeconomic Measures—The Public Health Disparities Geocoding Project (US)." *Public Health Reports* 118(May–June):240–260.

Mellnik, T., R. Fischer-Baum, and K. Soffen. 2018. "Thousands of Virginians May Have Voted in the Wrong State House Districts." *Washington Post*, January 9, 2018. https://www.washingtonpost.com/graphics/2018/local/virginia-wrong-districts/.

New York Consolidated Laws. 2012. Section 121.

Ortiz, E. 2018. "Native Americans Fighting Back Against North Dakota Voter ID Law." NBC News. Oct. 30. 2018. https://www.nbcnews.com/politics/politics-news/native-americans-fighting-back-against-north-dakota-voter-id-law-n926326.

Purcell, J. 2018. "Some Habersham County Voters Were Assigned to Wrong State House Districts." *Now Habersham*, June 23, 2018. https://nowhabersham.com/some-habersham-county-voters-were-assigned-to-wrong-state-house-districts/.

United States Census Bureau. 2017. 2017 TIGER/Line Shapefiles. Accessed July 12, 2018. https://www.census.gov/cgi-bin/geo/shapefiles/index.php.

United States Census Bureau. n.d. Tracts and Block Numbering Areas. Accessed July 10, 2018. https://www.census.gov/history/www/programs/geography/tracts_and_block_numbering_areas.html.

Virginia State Board of Elections. 2017. Agenda, November 27, 2017. http://townhall.virginia.gov/L/GetFile.cfm?File=C:\TownHall\docroot\\meeting\151\26912\Agenda_SBE_new_v1.pdf

Ward, M. H., J. R. Nuckols, J. Giglierano, M. R. Bonner, C. Wolter, M. Airola, W. Mix, J. S. Colt, and P. Hartge. 2005. "Positional Accuracy of Two Methods of Geocoding." *Epidemiology* 16(4):542–547.

Weiner, R. 2018. "Appeals Court Won't Block Republican from House of Delegates over Ballot Mixup." *Washington Post*, January 10, 2018. https://www.washingtonpost.com/local/virginia-politics/appeals-court-wont-block-republican-from-house-of-delegates-over-ballot-mixup/2018/01/10/f2ae472e-f55a-11e7-a9e3-ab18ce41436a_story.html.

Welte, J. W., W. F. Wieczorek, G. M. Barnes, M.-C. Tidwell, and J. H. Hoffman. 2004. "The Relationship of Ecological and Geographic Factors to Gambling Behavior and Pathology." *Journal of Gambling Studies* 20(4):405–423.

Whitsel, E. A., K. M. Rose, J. L. Wood, A. C. Henley, D. Liao, and G. Heiss. 2004. "Accuracy and Repeatability of Commercial Geocoding." *American Journal of Epidemiology* 160(10):1023–1029.

https://doi.org/10.1017/pan.2019.44 Published online by Cambridge University Press