Digitizing, Districting, and Data: Creating an Open-Source Precinct Shapefile for Ohio

By Katie Jolly

Abstract

One of the most recent advances in voting rights research has been the use of GIS and computation to create new metrics for fairness of maps. Voting precinct shapefiles are necessary in order to spatially evaluate election results, bring gerrymandering cases to court, and create viable alternative districting plans. Even as more government data has been made public over the last few years, many counties still do not publish their precinct boundaries in an accessible format. Ohio is particularly challenging for precinct shapefile collection. As part of a larger project, we set out to create an open-source, statewide shapefile from county shapefiles or PDFs. In a handful of counties, there were neither. Instead we developed a methodology that uses geocoded addresses from the public voter file to approximate precinct boundaries. In theory, precinct boundaries do not split census blocks, so we wrote a nearest neighbor classification algorithm to allocate blocks to precincts. This method loses some of the accuracy of digitizing PDFs but saves considerable time and produces comparable results. As the 2020 census and election approach, the implications of districting plans become more important to understand. Drawing precinct boundaries is a first step, and this project aims to make that process easier for future researchers and community members.

I. About me and this project

I am a senior in the geography and applied math & statistics departments at Macalester College. My interests right now are civic data, geographic data science, and public policy. Many of my projects have been related to how all three of those things can come together in a meaningful way to make an impact.

This past summer I worked at the Voting Rights Data Institute at Tufts and MIT in Cambridge, MA. Broadly, we worked on projects related to the mathematical and computational aspects of gerrymandering research. My main project was creating an open source precinct shapefile for Ohio. We started with a list of 88 county names and over the summer built a full

shapefile and joined it to as many statewide election returns as possible since 2010. I gained an appreciation for county government employees, Twitter as an academic resource, and crowdsourcing data collection.

In this talk I'll start with some motivation for why we wanted to do this particular project, then talk about the creation of the shapefile. In that section I'll mention the digitizing process, but I'll spend the bulk of the time on algorithms we wrote to approximate precinct boundaries from voter registration data. After that I'll talk briefly about how we joined the shapefile to the election returns. I'll end with a few questions that came up throughout the project about open-source and reproducible research, and some ideas I have for future work related to this research.

II. Why are precinct shapefiles important for voting rights?

Precincts are the smallest unit at which election data are reported. Counties can also provide a coarse picture of election returns, but precincts give significantly more detail to the story. In order to make nuanced arguments about the fairness of districts, we need to know about the precinct boundaries (*Geo-Enabled Elections*, 2018).

In the United States, precinct boundaries are left up to the county to decide. On the surface that is not a bad thing, but a problem arises when there is no central database for that information. Often the information isn't digitized in an easily shareable way. In my ideal world there would be a centralized repository of shapefiles that is required to be updated whenever there is a boundary change. However, we are pretty far from that in terms of data capacity, and I realize that's a larger problem than just precinct boundaries.

Recently in the news, with the election and decennial approaching in 2020, there has been a resurgent interest in polling place accessibility, voter registration campaigns, and (sometimes hidden) voter disenfranchisement. For example, there's an ACLU lawsuit in Georgia now over

closing polling places in predominantly African-American areas.The Supreme Court also recently upheld Ohio's practice of purging inactive voters from the voter rolls (Liptak, 2018). Many of the most pressing voting rights challenges require precise knowledge of where polling places are and who is using those polling places, as well as which voters are affected by new policies.

This project aims to make it easier for community members to learn more about their own precincts and create one way for people to support concerns about equity and fairness with numbers. We focus on Ohio, but we hope our methodology can be repeated in other states as well and support the work of groups working on similar research.

III. Creating the shapefile

In June we started with a list of the 88 counties in Ohio and a group of about 10 fellows. For each county we looked online for a publicly available shapefile. Around 30 had shapefiles up-front. Counties fell into one of four categories: had a shapefile, had a web map, had a PDF map, had nothing available (at the time).

After the initial internet search, we called the county Board of Elections and/or a secondary county official (a GIS specialist, for example). Our conversations with them depended on what information we had available to us already. For counties with shapefiles, we asked when that particular plan was adopted and if they had access to any previous plans' shapefiles. For counties with a web map as the best option, which was fairly uncommon, they were almost always able to send the underlying shapefile, as well as answer all of the usual shapefile questions. Counties that had only PDF maps available online either said that the PDF was the best and only option or had a shapefile that just had not been released online, often due to a lack of strong GIS infrastructure.

The last and most difficult category were those counties that had nothing available online. Those calls went one of three ways. First, they would be able to send us a PDF or shapefile, which was the best situation. Second, and more likely, they were able to mail us a hard copy of a map. These were often paper maps or highway maps that had precinct boundaries drawn by hand with markers. Third, and slightly less likely, they really had no maps they were able to give us. These counties were generally small and rural, although some were surprisingly populated. Fourth, the county officials never answered the phone after repeated attempts or hung up on us. That only happened in one or two counties, though.

The counties that had no maps were the intellectually interesting ones. Later in this paper I will discuss more in depth how we handled drawing precincts for those counties. Ultimately we used geocoding and the addresses listed in the voterfile to approximate the boundaries.

In total, we were given shapefiles from 50 counties, PDF or paper maps from 31 counties, and no maps from 7 counties. The next few sections describe our methodology for each of the cases. Note that for counties that provided a shapefile, often the only work on our part was creating a projection file when one was not already included so I do not discuss that methodology in this paper.

IV. Digitizing

In this section I refer to digitizing as the full process including both georeferencing and vectorizing.

Among counties that provided us with PDF or paper maps, there was a wide range of quality and quantity. Some counties had a single PDF map of all the precincts whereas other counties had one precinct per PDF, which could be as many as 100 or so. Other counties mailed us

paper maps, which also had a wide range of quality. There were even counties that split precincts among several PDFs. To digitize we used OpenStreetMap in QGIS.

There were approximately 450 total images to digitize. We hosted “digitizing parties” and invited everyone at VRDI with the promise of a free dinner and new technical skills. At the first one, I led a short lesson on georeferencing and projection basics. We also made a step-by-step guide for georeferencing that was specific to these maps.

During the second party we started vectorizing the georeferenced maps. This was prone to a few more technical issues, including difficult angles and fine-tuning. Due to time constraints we vectorized by selecting census blocks that were within the precincts because in most places, precinct lines followed census blocks.

From a project management standpoint we usually assigned the georeferencing of a county to a group of two to four people and then assigned only one person for the vectorizing. This helped minimize opportunities for human error.

In total we estimated that we spent about 400 person hours on the project. Some of that time comes from the fact that many people had never used GIS before and we had to redo a few counties. In general, it’s a long and tedious process. We are currently working with artificial intelligence researchers to brainstorm faster and more innovative ways to georeference and vectorize maps. As of now, by hand seems to be the only viable option.

V. Voterfile approximations

The more challenging case was when no map existed that we could use. Much of the research I did was related to how to use other publicly available data to approximate the precinct boundaries in these counties. In the end we found that geocoded addresses from the voterfile could draw “accurate-enough” precinct boundaries in the absence of a better option. Groups like

the National States Geographic Information Council are also using voter addresses to geocode precinct boundaries with address range shapefiles.

I wrote an algorithm that takes the voterfile as input and then uses the Census Bureau API to find the block ID for each valid address. From there I assign blocks with valid addresses to a particular precinct. This usually classifies between 40 and 60 percent of the blocks in a county (see Figure 1 in appendix). The harder part is classifying the blocks without valid addresses.

I'll use Noble county as an example to demonstrate how our classification algorithm works. The algorithm is written in R but I'm sure it could be written in Python as well. I'm a much more comfortable R user so it is my language of choice.

The algorithm starts by finding the rook-contiguity neighbors for each census block. Then it identifies which blocks are unclassified but have at least one classified neighbor. For all the blocks in that category it assigns the most common precinct of the surrounding blocks.

The process then works iteratively. After the algorithm assigns precincts to unclassified blocks with classified neighbors, it starts again. Each county usually takes four to eight iterations to classify each unclassified block. (See Figures 2-5 for maps from Noble County).

There are a few things I have in mind to improve this algorithm in the future. I'd like to build a classification model that takes into account municipal boundaries and major roads because those are often dividing lines for precincts. Another improvement would be to require a higher number of minimum classified neighbors (i.e., more than one), in order to classify a block. I'd also like to combine this census block method with address-range geocoding to have the maximum possible information about precinct locations.

VI. Checking for accuracy

One major question we were left with is whether or not our precinct approximations were "accurate." We did not have a great way to measure accuracy. In order to measure anything we needed some sort of baseline. We decided to approximate precincts for counties that had provided us with official shapefiles. I'll use Clark county as an example for the voterfile accuracy checks. I chose this example because it includes Springfield, a moderately sized city, to illustrate the different issues we see with rural and urban precincts.

The first method we thought of for measuring accuracy was finding the percent of the total area in the approximation that was misclassified. To calculate this we took the absolute value of the difference between a precinct in the base and approximation and divided by to to account for double-counting area. This worked well enough. We found that most maps had between 2 and 5% of the area misclassified. We were pleasantly surprised by how low those numbers were. When we calculated the error for Clark County we found that only 2.9% of the area was misclassified.

There were still some unsatisfying parts of that measurement though. It did not tell us a lot of interesting information about the accuracy and lacked nuance. It was a pretty baseline metric. As an alternative, we used the *sabre* R package (Nowosad & Stepinski, 2018; Nowosad, 2018), which is a spatial implementation of the validity measure, v-measure, that comes from computer science to evaluate clustering algorithms. This measures calculates both completeness and homogeneity. In this case, completeness measures how well the precincts in the official shapefile fit into the approximated precincts, and homogeneity measures how well the approximated precincts fit into the official precinct boundaries. The values are between zero and one, with one being the goal. The v-measure is the harmonic mean of homogeneity and completeness. V-measure allows us to quantitatively compare two categorical maps. In other words, it gives us a way to measure the association between two separate regionalizations of the same domain.

When we calculated the v-measure for Clark county we got 0.95. The value for homogeneity was 0.96 and completeness was 0.95, all indicating high spatial agreement between the two precinct shapefiles.

We can also find the regional inhomogeneity for each of the precincts. This tells us a little bit more about where the inaccuracies are within the county, rather than just a global measure. In Clark county we found that the worst precincts were more urban than rural. The shapes of urban precincts are more irregular, making them more difficult to approximate using our census block method.

As a secondary measure we also calculated goodness-of-fit using mapcurves. For these calculations the official shapefile was used as the reference layer. We found that Clark county has a goodness-of-fit value of 0.88. In general we prefer the v-measure to the mapcurves method for assessing accuracy.

VII. Joining election data

We obtained the data for 2016 election returns from the Ohio secretary of state website. We then created a lookup table for precincts by manually matching precinct names and codes from the election data to the precinct information from the shapefiles we merged. After filtering out invalid precincts from the merged shapefiles, we performed a full join on both precinct and county names to obtain a complete shapefile with the election data attached.

At this point we have not been able to match each precinct from the shapefile and the election returns. Of the 9298 precincts total, 2.25% of the precincts (209) in election results are missing from shapefile and 3.915% of the precincts (364) in shapefile missing from the election results.

VIII. Research challenges and theories along the way

I. Benefit of an open source project

From the start of this project, and in keeping with the mission of the Metric Geometry and Gerrymandering Group, we knew it was important for this work to be open-source and transparent. The intention was always to create a shapefile that anyone could download and use, as well as see the work that went into it. To this end we were also intentional about using open-source software and code instead of graphical user interfaces (GUIs) so that the work would be as reproducible as possible. All of the R scripts we used are available and I would be happy to talk more about them or how we used QGIS.

There are now copies of the shapefile in its current state on GitHub, along with the data in geopackage and geojson format to make it as easy as possible to use. I hope that people start to move away from shapefiles to other file formats, but I included a variety of file formats for usability.

ii. Crowdsourcing the work

When we started this project, we gave an estimate of two weeks for a draft of the shapefile. We had no idea that it would take about 3 months total. After we started to gather images and shapefiles we really saw the scope and importance of the project. We knew we needed a new plan for finishing all of the work.

I had participated in crowdsourcing for other projects before, including some with digitizing elements to them. We knew it wouldn't be possible to finish the project with just our small group working on it, but it was also scary to hand off aspects of the project to other people. We compromised on a model where we could crowdsource certain aspects, like digitizing maps, in a controlled environment.

iii. Digitizing skills

One unexpected outcome of this work was that so many people would get the chance to learn how to digitize maps when they otherwise would not have. A good friend of mine who is currently a mathematics Ph.D. student came to our digitizing parties to help out with the work and for the food. He caught on quickly and ended up being a great collaborator on the project. Later, when he was working on one of his own projects, he was given a paper map of a city but he really needed a digital version in order to analyze it. He pulled out the guide we had written for the digitizing parties and created his own shapefile of this city map with all the features included. He told us he honestly never would have thought it would have been possible to digitize it before learning about GIS. I was happy that aspects of this project that I considered to be somewhat unremarkable, like teaching everyone how to digitize the maps, ended up being pivotal for other projects that I cared about as well.

Citations

Geo-Enabled Elections. (2018). Retrieved from https://www.nsgic.org/geo-enabled-elections

Liptak, A. (2018, June 11). Supreme Court Upholds Ohio's Purge of Voting Rolls. *New York Times*. Retrieved from https://www.nytimes.com/2018/06/11/us/politics/supreme-court-upholds-ohios- purge-of-voting-rolls.html

Nowosad, J. (2018, September 10). *Sabre: Or How to Compare Two Maps?* Retrieved from https://nowosad.github.io/post/sabre-bp/

Nowosad, J., & Stepinski, T. (2018, April 19). *Spatial association between regionalizations using the information-theoretical V-measure*. Retrieved from https://doi.org/10.1080/13658816.2018.1511794

Pebesma, E. (2018). sf: Simple Features for R. R package version 0.6-3. https://CRAN.R-project.org/package=sf

QGIS Development Team (2018). QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org

Walker, K. (2018). tigris: Load Census TIGER/Line Shapefiles. R package version 0.7. https://CRAN.R-project.org/package=tigris

Wickham, H. (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Images

Figure 1: Choropleth map of Noble County census blocks showing valid addresses (successfully geocoded)

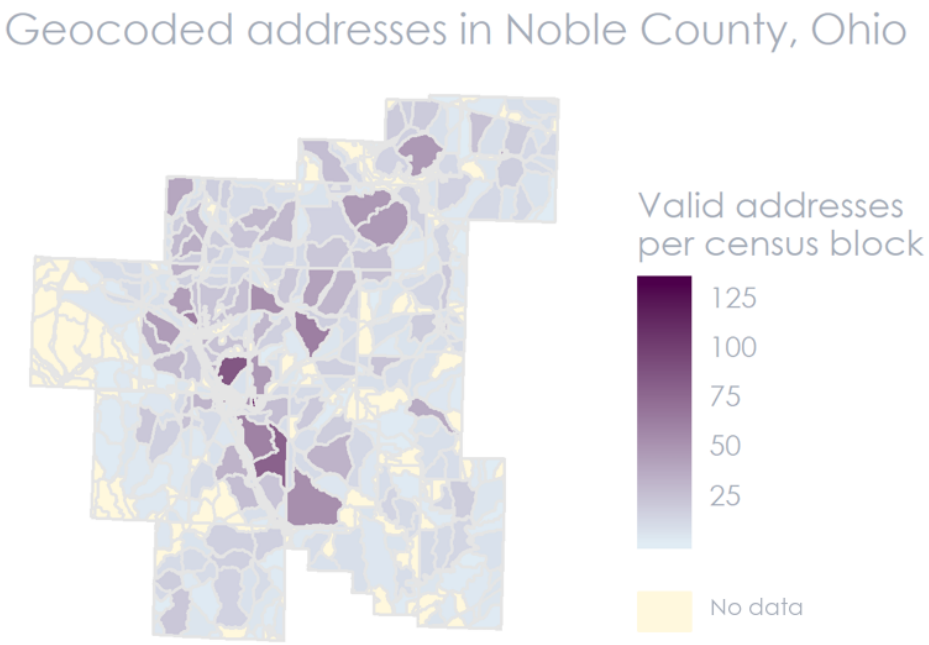

Figure 2: Noble County precincts after one iteration of the classification algorithm

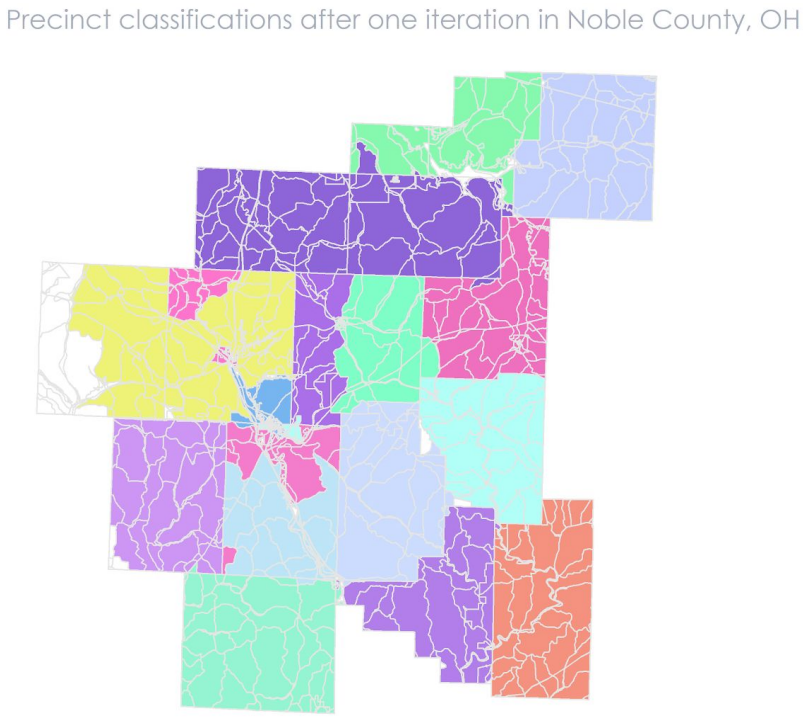

Figure 3: Noble County precincts after two iterations of the classification algorithm

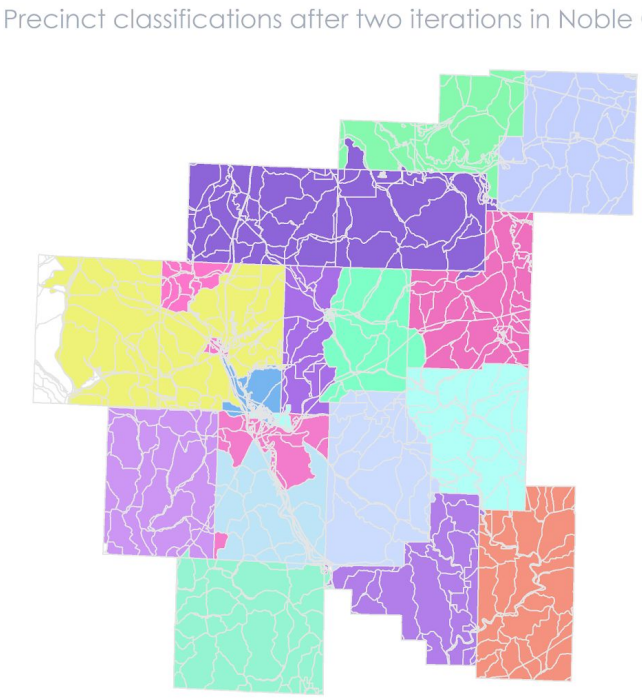

Figure 4: Noble County precincts after three iterations of the classification algorithm

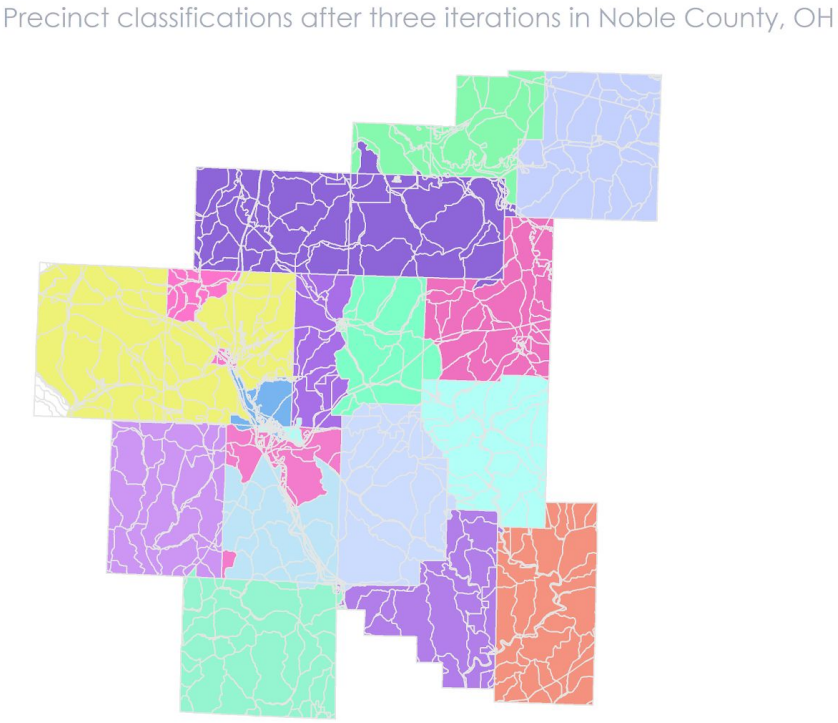

Figure 5: Noble County precincts after six iterations of the classification algorithm (fully classified)