Chapter 13

Geography as data

LEE HACHADOORIAN AND RUTH BUCK

CHAPTER SUMMARY

The U.S. Census Bureau was established to enumerate all of the residents of the country every ten years. Its geographic units, and the counts of people attached to them, are the basic stuff that districts are built from. This chapter will tell us about Census and electoral data products and the spatial tools that let us manipulate them.

1 INTRODUCTION

The last two chapters discussed key concepts in geography and geographic thought as they relate to U.S. redistricting. This chapter offers an introduction to the data side of the story. We'll introduce common terminology, survey the data sources, and discuss the use of geographic information systems (GIS) software and other mapping tools to make it all come together.

From achieving population balance between districts, as required by the Supreme Court, to determining whether or not a districting plan complies with the Voting Rights Act, we find Census and electoral data, coupled with the powerful capabilities of GIS, at the heart of redistricting practices.

This discussion begins with a look at the Census. Census data act as the ground truth about what kinds of people live in what locations for most redistricting purposes. It must be brought together with a second and much messier data source: precinct and election results data from state and local authorities. After that, we spend a fair amount of time introducing the GIS tools used for contemporary geospatial data analysis and display. Finally, we review some puzzles and challenges that are specific to the geospatial data of redistricting.

WHO'S WHO

It's important to remember that most of the materials and practices we describe in this chapter go far beyond the application to redistricting. Demographers, human geographers, spatial statisticians, public health scholars, and urban planners all make heavy use of census data and GIS.

In fact, GIS more broadly encompasses the entire system of tools and technologies that can be used to work with spatial data, including satellites, GPS devices, drones, web-mapping servers, spatial databases, and geoprocessing in a variety of programming languages. Our discussion here focuses on a subset of domain-relevant technologies and methods.

2 THE CENSUS AND ITS PRODUCTS

The Census divides up the nation into geographies from coarse (states) to fine (census blocks). In essentially every case, electoral districts are made from these geographic units. The relationship of the Census to redistricting goes back to the nation's founding: Article 1, Section 2 of the U.S. Constitution mandates an enumeration of the population every ten years for apportioning membership in the House of Representatives. Who is counted and how they are classified has changed over the decades, subject to both technical advances and broader social and political changes.

Margo Anderson, in *The American Census: A Social History*, gives us a look at the Census as a record of American social classification practices [1]. For example, the first Census, in 1790, asked for the total number of people in each household, according to the following categories: free White males under sixteen; free White males sixteen or older; free White females; "other free persons"; and slaves. Only the name of the head of the household was collected.

Since then, the Census has changed dramatically. Over the course of the 19th and 20th centuries, a number of economic and social questions were added. Concern over its length and infrequency led to the separation of the Decennial Census into a short form and a long form, with the short form covering only basic questions about age, race, and ethnicity for each member of the household.[1] The long form surveys culminated in the development of the American Community Survey (ACS), a detailed annual survey of about 1.5% of the population, which began in 2005 and has become the nation's leading source of socioeconomic data.[2] Unlike the ACS, which is based on sampling the population, the Decennial Census attempts to create a person-level database of the *entire* population.

From a nuts-and-bolts perspective, the Census Bureau begins with a Master Ad-

[1] The Decennial Census also includes information on housing and *group quarters* like dorms, prisons, and military bases, but for the purposes of this chapter we will focus on the information it collects regarding population.

[2] The ACS asks respondents many pages of detailed questions regarding income, education, access to cars and the internet, and so on. Because it is based on a survey, the ACS data are frequently used in rolling five-year averages.

dress File (MAF), a list of residential addresses where Census forms will be mailed. Households return the forms by mail, or, beginning with the 2020 Census, they can complete the form online. In the event of a non-response, Census canvassers will visit an address to determine who, if anyone, is living there. Everything the Census collects begins here, on the ground, and everything it assembles from these gathered data is considered a *product*, which is then *released* on a schedule. Current privacy law protects the full records from disclosure for 72 years, so the Census releases the data as counts that are aggregated into various geographic units.

2.1 CENSUS GEOGRAPHIES

HIERARCHY AND CONCEPTS

Census data products take a variety of forms, one of which is *geographies*: spatially described subsets of territory. Users may call them geographic areas, geographic units, or geographic entities.[3] This section will sketch out their structure and some of their many uses.

The Census Bureau provides demographic and socioeconomic data for a staggeringly large number of geographic areas. A hierarchical structure, depicted in Figure 1, helps to keep track of all the scales and interrelationships. The center line in the figure is called the *spine*, especially the six-level structure

Nation—State—County—Tract—Block Group—Block.

The smallest units, census blocks, completely cover the territory of the United States. These nest inside block groups, which nest inside counties, and so on up to the nation. Geographies that fall outside of the spine may not nest neatly and may not entirely cover the larger unit to which they belong. For example, Congressional districts fall wholly within states, but do not necessarily honor any other geographic boundaries until the smallest unit, blocks. A particular redistricting plan may strive to keep counties whole, but this is not guaranteed, so counties do not appear in the hierarchy under Congressional Districts.[4]

Off-spine geographies include legal or administrative geographic units such as Federal American Indian Reservations and school districts. In addition to units like these that are decided externally, the Census Bureau and other federal agencies *create* some geographic units for the purpose of data dissemination. For example, the Office of Management and Budget defines *metropolitan / micropolitan statistical areas* (MSAs) based on commuting patterns and uses those to report statistics that are useful to researchers and planners. Another example is Zip Code Tabulation Areas (ZCTAs). The Post Office creates Zip Codes *only* for the purpose of delivery route planning, but because there is now a wealth of industry and marketing data available by Zip Code, the Census Bureau has built corresponding ZCTAs for which demographic data are reported. These are examples of geographies off the central

[3] Or you might see hybrid terms like *areal units* for units pertaining to areas.

[4] The Census Bureau releases data for many "part geographies," so it is possible to download demographic and socioeconomic data for "Counties split by Congressional Districts," for instance. This would show demographics for the part of a county that falls in a specific Congressional district.

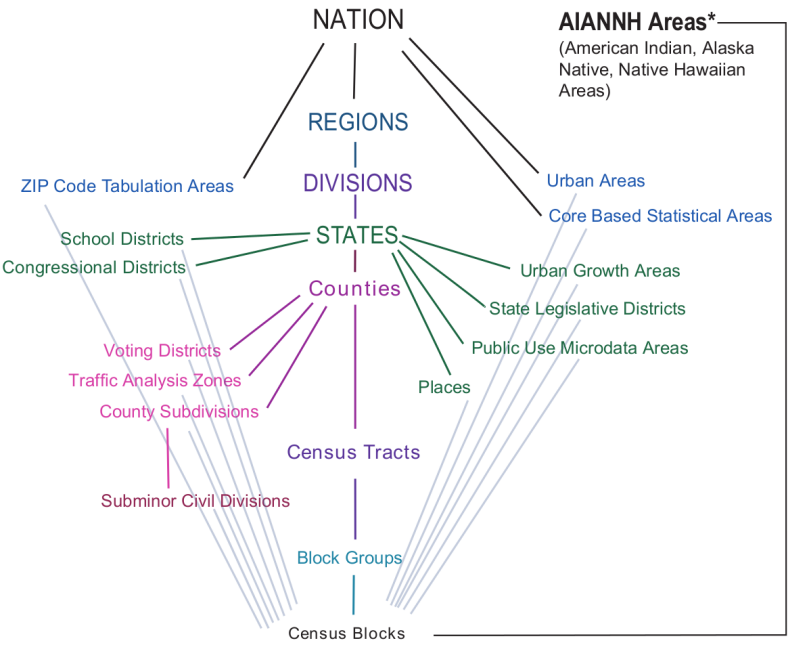

Figure 1: Hierarchy of Census geographies, from census.gov [24]

spine; they are made up of census blocks but do not necessarily nest in any of the other important units.

The hierarchical structure is reflected in the system of unique identifier codes attached to the geographic units, denoted GEOIDs. For example, states are identified by a two-digit FIPS (Federal Information Processing Standards) code (e.g., "01" = Alabama, "42" = Pennsylvania). Congressional Districts use an additional two digits; for instance, 4207 labels the 7th District in Pennsylvania.

GEOGRAPHIC PRODUCTS

Ultimately, all of the geographies described here are stored in a massive database called the TIGER system and are made available as a set of *TIGER/Line files* and related products.[5] If you want anything like a canonical set of mapping products for American geography, this is it. When a new geography is supplied by a state, it is standardized and processed into TIGER/Line format, correcting errors (e.g., gaps in coverage) and aligning units.

[5]TIGER stands for Topologically Integrated Geographic Encoding and Referencing. As the Bureau documentation says, "The TIGER/Line Shapefiles are the fully supported, core geographic product from the U.S. Census Bureau."

13.1 WHAT'S IN A BLOCK?

The *census block* is the basic building block of all census geographic units. Block boundaries are keyed to streets, roads, railroads, water bodies, legal boundaries, and "other visible physical and cultural features" [23]. In a city, a block might be coincident with a city block, while in a rural area a block might be a larger plot of farmland.

Given that the Census counts people by residence (i.e., where they sleep), and block construction is determined by the physical and social landscape, some census blocks may have odd shapes or zero population. In some cases, a census block picks out a traffic circle or a winding segment of a multi-lane highway. This helps to illustrate the difficult balancing act between the role of blocks in logically segmenting the territory of the country and their role in finely enumerating the population.

The use of census blocks to cover the entire country was initiated in the 1990 Census. By the time of the 2010 Census, the country was divided into over 11 million census blocks, of which over 40% (more than 4.8 million) have no reported population at all. Over half a million blocks (541,776 out of 11,078,297, to be precise) are wholly composed of water. Figure 2 shows census blocks in Philadelphia.

In populated areas, as blocks often divide from each other along streets, districts built out of blocks will also have the property that along their edges, people are separated from their across-the-street neighbors—a benign consequence of this choice of units, but somewhat counterintuitive. In California public meetings, some voters vocally objected to being in separate blocks from co-residents of a housing development that is separated by a street, while being joined across rear lot lines with another housing development [17]. Urban researchers have for some time questioned the usefulness of these census units for local analysis, noting that dividing lines *behind* homes are more natural for understanding how people think about neighborhoods [5].

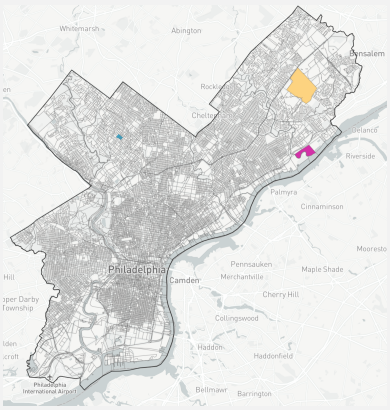

Figure 2: Census blocks in Philadelphia. The Northeast Philadelphia Airport (yellow) is its own census block, population 2. (It's unclear who is considered to live there; perhaps homeless people.) The most populous block in Philadelphia is the Riverside Correctional Facility (pink), population 4535. But most census blocks are city blocks, like the five marked in the Germantown neighborhood (blue), which have between one hundred and three hundred people each.

Depending on the purpose of a map, factors such as scale, appearance, or even storage size, may require a reduction in the level of detail. The process of simplifying geographic information—removing detail, either manually or algorithmically—is called *generalization*. The objective of generalization is to reduce the precision of geographic data in order to suppress information that is irrelevant or inessential to a map's purpose without sacrificing clarity or accuracy of relevant spatial relationships [27].

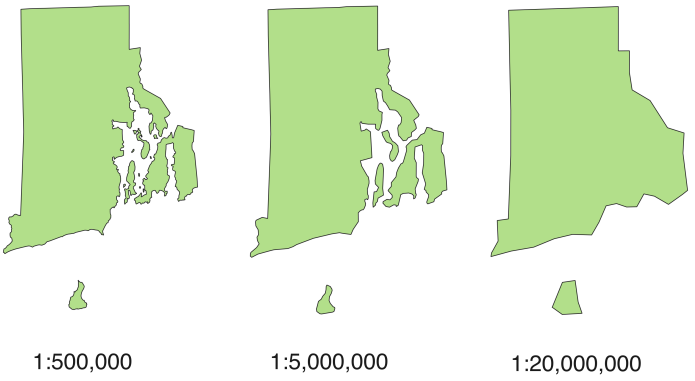

Figure 3: Three different levels of generalization of Rhode Island.

In addition to the full-resolution TIGER/Line files, the Census Bureau also provides generalized cartographic boundary files at 1:500,000, 1:5,000,000, and 1:20,000,000 resolutions. Figure 3 shows the state of Rhode Island at these three levels of resolution. The purpose of these files is to be suitable for printed maps, which are not well served by accuracy down to feet or inches.

2.2 WHO GETS COUNTED, AND HOW?

The aspiration of the Census is to enumerate all the people resident in the United States on April 1 ("Census Day") of each year ending in 0. This is an increasingly difficult task as the population has grown to over 300 million, and as you can imagine, it was especially thorny in 2020, when the coronavirus pandemic shut down most of the country in mid-March.

Population counts will ultimately be reported in *tabular* form—i.e., in tables or spreadsheets. Besides the main tables, they also offer *cross-tabulations* ("cross-tabs") showing intersection counts for pairs of primary variables, and *special tabulations* ("special tabs") for alternative aggregations answering common queries that are not addressed in the main tables. In this section we introduce some of the complexity of how the numbers are produced.

13.2 PRODUCING RACE COUNTS

Among the demographic data that the Census collects is self-identified race and ethnicity, using categories that reveal changing social conceptions of race. For example, in the late 19th century, the Census included a "Chinese" racial category that subsumed all East Asians. South Asians were considered "White," but later were assigned to a new "Hindu" category, which included Muslim Pakistanis. Now, South Asians and East Asians are grouped in the racial category "Asian."

Figure 4 shows the question asking about race in the 2010 Decennial Census.

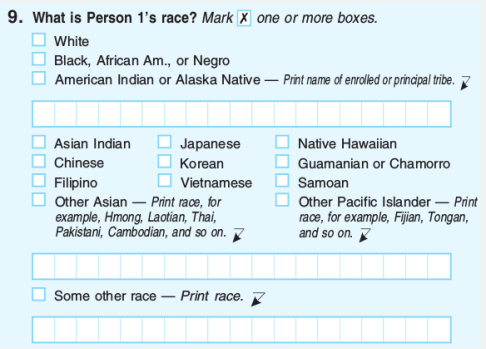

Figure 4: "What is Person 1's race?" from the 2010 Census.

If a race is not reported, it may be inferred or *imputed* from data gathered about other household members, or from the previous census household record. The Census processes this information and then reports race counts by the following six categories.

- White
- Black or African American
- American Indian and Alaskan Native
- Asian
- Native Hawaiian and Other Pacific Islander
- Some Other Race

Since the 2000 Decennial Census, the questionnaire has allowed respondents to select multiple races. Since there are six racial categories, there are $2^6 - 1$ or 63 possible combinations (you can be in or not in each of the six, but you can't be enumerated in no category at all). In 2010, the great majority of respondents were enumerated in some single race category; only 2.9% of respondents were ascribed two or more races.

The form produces some difficulties in self-reporting because it does not map perfectly onto American racial discourse. For example, many Americans think of Hispanic/Latino as a race, but because the Census treats it in a separate ethnicity question, the form forces Hispanic-identified people to make a race choice from this list. Nationally, in 2010, about 65% of Hispanic people identified as White and 27% as Some Other Race (an option selected by only 0.3% of non-Hispanic people). This phenomenon had significant regional variation. For another example, the lack of a Middle Eastern/North African category leads to MENA self-identification spread over multiple categories (Asian, African American, White, and Some Other Race), making it ultimately harder to conduct social science research on this group, even though it has strong social and racial salience.

UNDERCOUNT AND OVERCOUNT

While the Decennial Census is intended to be a complete enumeration of the population, it obviously can't achieve perfect accuracy. The *undercount* is the number of persons resident in the United States on April 1 of a census year who are not enumerated. *Overcount* occurs when a reported number is too high, such as if a person is included twice.

The stakes for a correct count have increased considerably over the course of the last half-century. The growth of federal aid programs, which often distribute resources in proportion to population, has meant that state and local governments have a considerable financial stake in the accuracy of the Census. Civil rights legislation has included employment provisions that rely upon a count of racial minorities. And enforcement of the Voting Rights Act hinges on counts of population by race and sometimes by language group.

Since 1950, the Bureau has published an assessment of overcount and undercount for each Census, using both demographic analysis and a post-enumeration survey.[6] They estimated approximately 16 million omissions in the 2010 Census (5.3% of the population). While it may seem obvious that the Census Bureau will miss counting some individuals, the source of the overcount is, at first glance, puzzling. 85% of erroneous enumerations in the 2010 Census—persons counted who should not have been counted—were duplications, usually due to individuals whose residence is ambiguous, such as college students living away from home, incarcerated or military population, children with two custodial parents, and households with multiple properties. 15% were erroneously enumerated for other reasons, including persons who were born after or died before April 1 and "fictitious persons."[7]

The deeper issue is not merely obtaining more accurate overall numbers. The problem is also one of a *differential undercount*, deeply correlated with demography and geography. The Bureau's own estimates of undercount disaggregate as follows:

- Black: 2.07%
- Hispanic: 1.54%
- Native Hawaiian/Pacific Islander: 1.34%
- Asian 0.08%
- Native American: 4.88%/−1.95%
- White: −0.84% (a net overcount)

Notably, the Native American net undercount breaks down as an undercount of 4.88% on reservations, but −1.95% (a net overcount) off reservations.

The Bureau's report finds that many other characteristics also correlate with count inaccuracies. Residents of owner-occupied housing are consistently overcounted, while renters are undercounted. Very young children are undercounted, while teenagers are overcounted. Both males and females who are 50 years and older were overcounted in the past three Censuses. At younger ages, the pattern diverges: 30- to 49-year-old women are overcounted, while 18- to 49-year-old men

[6] Read about the Census Bureau's Coverage Measurement techniques: www.census.gov/programs-surveys/decennial-census/about/coverage-measurement.html.

[7] In 2010, there were 10 million erroneous enumerations and 6 million whole-person imputations—people from whom the Bureau did not collect sufficient information, but inferred characteristics to include in the count. The overcount thus almost exactly offset the undercount, yielding a net overcount of 0.01% [19].

are consistently undercounted.[8]

Because we have a great deal of external evidence that points to patterns in the undercount, statisticians at the Census Bureau could adjust the data to compensate for known disparities. But in 1999, the Supreme Court ruled (in *Department of Commerce v. United States House*) that statistical adjustment could not be used when deciding how many members of Congress are apportioned to each state. Intriguingly, they left the door open to the possibility of statistical adjustment used at the state level, such as for redistricting, and several states are now considering their options for count correction in an unprecedentedly challenging and contested census year.

2.3 SPECIFIC PRODUCTS FOR REDISTRICTING

The malapportionment court decisions of the 1960s put significant pressure on the Census to tabulate population counts for small-area geographies. To achieve the principle of "One Person, One Vote" in practice, and to judge compliance with the Voting Rights Act, public officials wanted more and more detailed census data, at the finest spatial resolutions possible, as quickly as possible.

In 1975, Congress passed a bill called Public Law 94-171 in order to meet these needs. In response, the Bureau took up the mission of "provid[ing] states the opportunity to specify the small geographic areas for which they wish to receive decennial population totals for the purpose of reapportionment and redistricting" [25].

THE PL 94-171

Since this legal shift in the 1970s, the Census Bureau has been required to provide redistricting data to the designated authority in each state by April 1 of the year following the Decennial Census. These tables are now eponymously referred to as the *PL 94-171 data*; they contain counts by race and ethnicity in every census block in the nation, with a second table reporting the same counts among voting age population.

States generally complete their redistricting process by the end of the year ending in 1, drawing new Congressional and legislative districts for use in the primary and general elections of the year ending in 2. In addition to these products, the Census Bureau issues a special data release in the year ending in 3 containing information about the newly enacted districts. In 2021, everything is expected to happen late because of pandemic-related data collection problems and an unprecedented level of interference from the White House.[9]

[8] Note that geographic imprecision when locating addresses can lead to under- or overcounts at the block level that disappear at larger levels of geography. An estimated 2,039,000 people, or 0.7% of the United States population, are enumerated in the correct county of residence, but the wrong "block cluster" (group of nearby census blocks). So the practice of fine-tuning population to ±1-person population deviation between districts may very well be shifting blocks whose measurement errors are larger than the population differences they are trying to correct.

[9] The most recent PL 94-171 was delayed until August 2021, causing many states to scramble in order to meet their timelines to create new maps.

13.3 THE CITIZENSHIP QUESTION

The Census Bureau often releases revised or specialized data separately, in so-called special tabulations. An important one for redistricting is the Citizen Voting Age Population (CVAP) by Race and Ethnicity. This special tabulation is based on the American Community Survey, in which respondents are asked about their citizenship status. Because only citizens are eligible to vote in state and federal elections, CVAP provides a much better estimate of the demographic balance in the electorate than population or voting age population alone. When Voting Rights Act challenges are brought on behalf of Asian or Latino plaintiffs, CVAP is always used in the supporting data.

However, citizenship status has been carefully kept out of the Decennial Census "short form." This is because, even though census form responses are supposed to be firewalled from other government agencies, people who are in the U.S. with legal concerns about their residency are likely to be intimidated by an official form asking for their citizenship status, making them potentially far less likely to be enumerated—and for apportionment purposes, this can't be corrected statistically. Without the question on the short form, citizenship numbers have not been included in the PL 94-171 data.

This too has become highly politicized. The conventional wisdom holds that excluding non-citizens from redistricting would benefit Republicans, and so several "red states" have made moves to do their redistricting on the basis of citizen-only population. In 2016, the Supreme Court heard a case called *Evenwel v. Abbott* in which Texas sought to use citizenship data in this way; its finding re-affirmed that total population is the traditional basis for redistricting but left the door open for future options.

The Census Bureau is technically part of the Department of Commerce, whose leader is a presidential appointee. As part of the Trump Administration's far-reaching efforts to control the levers of government, the Secretary of Commerce demanded that the Census Bureau add a citizenship question to the short form. When enactment was halted by the Supreme Court, the Bureau was instructed to use other means (like administrative records) to estimate citizenship numbers and include them in the PL94-171—a clear attempt to pave the way for citizen-only population balancing. Citizenship numbers were once more blocked from the redistricting data in this cycle, but future litigation is certain.

THE REDISTRICTING DATA PROGRAM

The Redistricting Data Program is a small division within the Census Bureau that supports the redistricting-specific needs of state and local officials. In the years leading up to each official Census Day, they conduct two main tasks: collect suggestions for changes to census block boundaries, and collect a snapshot of precinct boundaries from the states to create a data product called VTDs (voting tabulation districts, sometimes confusingly called "voting districts"). Both tasks rely on liaisons in participating states to coordinate with the Census Bureau. The opportunity for the states to specify the geographic areas for which they wish to receive redistricting data is granted by law [26].

Both block boundary suggestion and VTD collection should be thought of as a back-and-forth between the state liaisons and the census staff. On the census side, the submitted geography will be cleaned and aligned to make it conform to the TIGER protocol so that it can be used seamlessly with other data products. Because the VTD process is part of the Decennial Census effort, the Census Bureau publishes VTDs only once every 10 years. The Census Bureau neither keeps the geographies updated nor publishes ACS data for the VTDs during the intercensal period.

The Redistricting Data Program makes a valiant effort to standardize the wildly varying election administration units in the states, at least at one timestamp in the ten-year census cycle. We turn to that broader question now.

3 ELECTION DATA AND THE PRECINCT PROBLEM

Census geographies are an elaborate and crucial resource for understanding the human geography of the U.S. in spatial terms, but there is another fundamental piece of the redistricting puzzle: election results. These are not only needed for a wide range of analytic tasks for redistricting—from competitiveness to partisan skew—but are also an ineliminable part of Voting Rights Act enforcement, which relies on a showing of racially polarized voting, linking electoral history to demographics.

3.1 PRECINCT BOUNDARIES

The smallest unit at which election results are recorded and released is called a *precinct*. Usually, each one has a single polling place where people physically go to vote, but this is not always a one-to-one match. Here, we use the term "precinct" to refer to the electoral geography: the area whose residents are all handled together in voting administration terms. Perhaps surprisingly, and unlike census geographies, precincts are not drawn or maintained in a standardized way across or even within states!

Several states do have clear laws regarding the management and data transparency of precinct boundaries. In Minnesota, for example, Election Law requires municipal clerks to notify the secretary of state within 30 days of any precinct boundary change, who must then update the statewide precinct boundary database [20].

However, in many other states, the state does not track precinct boundary changes between redistricting cycles. Ohio, for example, requires that the list of registered voters be updated and all affected voters be notified of any precinct changes, but it does not explicitly require that those boundary changes be reported to any central election authority [7]. And in other states, counties or county subdivisions have complete control regarding the election precincts within their jurisdiction. Precincts are split, merged, or completely redrawn with such regularity that there

is no guarantee that Precinct A in County X covers the same territory in a special election in 2017 as it did in the general election in 2016.

This reflects a broader difference between what state elections officials often refer to as a "top-down" versus a "bottom-up" approach to election administration. The secretary of state's office in a top-down state (e.g., North Carolina, Wisconsin, and Minnesota) exercises significant control in how elections are administered and in how data from elections are stored and disseminated. In these states, precinct boundary data and corresponding election results can often be downloaded from the state geospatial data portal or from the Secretary of State's website.

In a bottom-up state like Ohio or Missouri, however, county and municipal governments are given almost complete autonomy in conducting elections. As these local governments have differing ordinances and widely varying mapping technology capabilities, precinct boundary data also vary in both quality and format. For example, during work on a project intended to collect the precinct boundaries used in Ohio's 2016 general election, members of the Voting Rights Data Institute (co-led by one of us, Ruth Buck) contacted officials from each of the eighty-eight counties. Forty-six provided data in GIS format, twenty-seven had PDF maps, eighteen mailed paper maps (sometimes held together with scotch tape, with precincts drawn in magic marker), and seven counties either refused to or could not provide precinct boundary data in any format. In those instances where the county could or would not share its data, we had to estimate precinct boundaries using addresses in the state voter file.[10]

To summarize: Reconstructing precinct geographies can be a complicated and labor-intensive post hoc process.

3.2 MAPPING ELECTION RESULTS

Another complexity in this work is that precinct-level election returns are generally provided in a tabular format such as a spreadsheet, showing total votes by precinct for each candidate in each election. This means that the names appearing in these tables of results must be matched up with names of units in a mapping format if we are to have any hope of visualizing the elections. This is unfortunately not always straightforward, even in states where precinct boundary data and election returns are both publicly available. There are almost always disparities in the names and even the total number of precincts between the precinct boundary data and the tabular election returns.

Once these disparities are resolved to the best of one's ability, there is rarely a *natural key* (i.e., a code that is meaningful, such as the two-letter state postal code, as a opposed to an arbitrary serial identifier) for connecting the datasets. Also, while states do have an obligation to provide certified election results to the Federal Election Commission (FEC), the manner in which states report and publish their precinct-level results varies wildly. In Mississippi, 2016 precinct-level

[10] A voter file, or voter registration file, is a database that usually contains the names, addresses, party affiliation, and precinct assignments of registered voters. Information included in the voter file varies from state to state, as do the cost and the bureaucratic obstructions to obtaining the file. Political campaigns will often buy processed voter files from commercial vendors to aid in targeting likely voters.

election results were reported by the Secretary of State's office as PDF scans of paper printouts. For the redistricting analyst, time-intensive data cleaning would be necessary before these results could be used.

Additionally, not all reported votes are tied to a precinct: ballots cast through absentee, provisional, or early voting are often reported only at the level of the election authority, such as a county board of elections. Federal overseas absentee votes, including ballots cast by U.S. military, are also often reported state- or county-wide rather than at the precinct level.[11] At the present time, it's obvious that these challenges will only expand. Elections in pandemic conditions drove people to alternative voting modalities in record numbers this year, with many estimates indicating that as many as 50% of votes cast nationally were early, absentee, or by mail. There are a variety of methods in spatial statistics that can be used for assigning votes reported at a coarser geography (like a county or a house district) to a finer geography (like a precinct). Unfortunately, there is no solution that is reliable across contexts.[12]

The picture is made more complicated by the difficulty of prescribing election procedures to the states, as the Constitution makes it explicitly a state affair. This could be addressed, for instance, with federal regulation requiring the reporting of election data in a modern, spatial format. Until then, the work will continue to rely on difficult and time-intensive data preparation.[13]

4 GIS: SHAPES AND ATTRIBUTES TOGETHER

4.1 WHAT IS GIS?

Contemporary mapping is heavily reliant on spatial software. The *International Encyclopedia of Geography* defines it this way:

> A geographic information system (GIS) is a computer system (desktop or web mapping software) for capturing, storing, querying, analyzing, and displaying geospatial data. Geospatial data describe both the location and attributes of spatial features. [4]

[11] In Alaska, absentee, provisional, and early votes are only reported at the house district level. Those votes, however, were about *one-third* of all votes cast in 2018 in both the gubernatorial and U.S. congressional elections in that state.

[12] This problem of data disaggregation is discussed further below in §4.2.

[13] Voting rights litigators must conduct this work each time they seek to press a Voting Rights Act case. Some academic and civic groups have attempted to collect and curate election geodata and make it publicly available. Several groups are currently working to collect and make publicly available high-quality election geodata. During the 2010 redistricting cycle, political scientists from Harvard, MIT, and Stanford created the Harvard Election Data Archive, which contains election geodata from most states (projects.iq.harvard.edu/eda [14]). In this cycle, efforts include MGGG States from the MGGG Redistricting Lab (github.com/mggg-states) and OpenPrecincts from the Princeton Gerrymandering Project (openprecincts.org). For tabular election results alone (with no geography), OpenElections has results in various stages of cleaning up to mid-cycle (openelections.net) but the most comprehensive and up-to-date data can be found from the MIT Election Data + Science Lab (electionlab.mit.edu/data).

GIS developed and spread in the late twentieth century alongside the rise of personal computing. Before this, redistricting (like many other planning operations) was conducted largely on paper. Now, GIS technology suffuses the redistricting process, from the gathering and dissemination of census data, to the creation of the electoral districts, to the analysis of the impacts of specific redistricting plans.

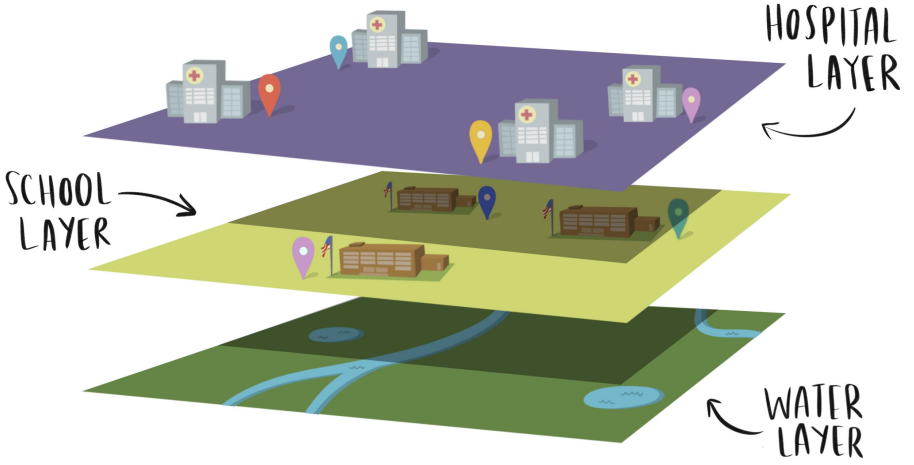

Figure 5: A schematic of data layers in GIS.

All map construction involves selection of what facts of the world to represent. Within GIS, maps are constructed by adding *layers* representing facts from different knowledge domains (concept illustrated in Figure 5). A map of a suburban street may be constructed of layers representing tax parcels, roadbeds, sidewalks, trees, building footprints, etc. A gas station layer might be important for a road map, but wouldn't be included on a map of religious affiliation.

MODELS, STRUCTURES, AND FORMATS

Our information will need to be organized. This is done on the one hand with a relational data model and on the other hand with a graphical interface.

GEOID	Name	Location	Biden votes	Trump votes
36059	Nassau County, NY	(list of vertices)	396,504	326,716
36081	Queens County, NY	(list of vertices)	569,038	212,665
36103	Suffolk County, NY	(list of vertices)	381,021	381,253

Table 13.1: A very small attribute table showing the relation between attributes of three entities. If we wanted to join further data, we would need it to be labeled in a common way, such as by the same GEOIDs, which can then be used as a key.

Here, the *entities* are counties, the *attributes* are the entries in the table, and the *relation* is the whole table's worth of information.[14] (See Harrington [13] for more

[14] Geographers tend to refer to nonspatial data as "attribute data," even though a geometry (such

on relational database design.)

Moving the discussion toward shapes, let's start with a distinction between *vector* and *raster* data (Figure 6). Vector formats store data as sets of coordinates that describe points, lines (sequences of endpoints with a line segment between each two successive ones), or polygons (a sequence of points that forms a closed loop). Raster formats store data as a grid of values. Geospatial phenomena can be represented in either model. For example, a forest could be represented as a polygon (vector) for the purposes of a road atlas, but could be represented by a specific value for grid cells in a land cover raster layer (where other values might represent grass, bare earth, etc.). The vector and raster models have different strengths; the vector model is more often used for hard-edged data, such as human-created political territories or tax parcels, while raster formats are better for the continuous or fuzzy-edged phenomena common in physical geography.

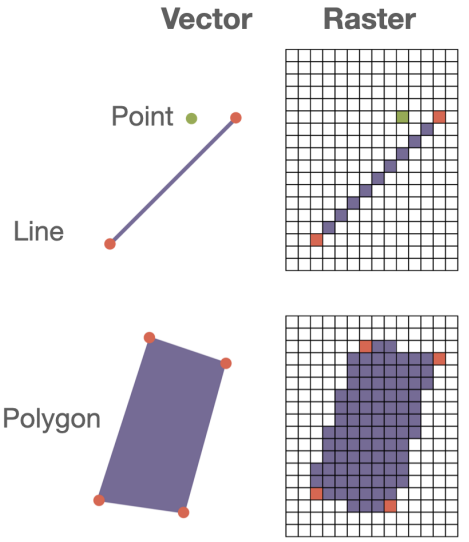

Figure 6: Vector vs. raster representations of points, lines, and polygons.

Census geographies are represented in GIS as vector polygons. Each element of a vector dataset is called a *feature*. For example, GIS might have a spatial layer for cities in which each individual city is a feature. Layers may be organized by type, scale, or theme. Roads are conceptually different from counties, and each would be stored as a separate layer file. Counties and states would also be stored as separate layers. Finally, the *attributes* ("facts") that one might want to know about an entity are virtually limitless, so to keep file sizes manageable, they may be separated thematically (e.g., into economic data and environmental data). As

as the shape of a Congressional District) is technically an attribute as well. Analysts from many fields might also refer to attributes as *variables*, and be interested in constructing models that demonstrate relationships among these variables.

the geometries (feature shapes) are usually the largest part of the file, geometries and attributes can be and frequently are stored separately.

LOCATION AND PROJECTION

The coordinates of a vector object are just numbers, and cannot be pinned to an earth location without specifying a *coordinate reference system (CRS)*. Latitude and longitude measurements are spherical coordinates that must be transformed for display on a flat computer screen or printed map using a *projection* method. Going from a sphere to a screen or map inevitably leads to distortion, but with knowledge of the specific projection method, it is possible to minimize and account for distortion in ways relevant to the project at hand.[15]

It is not uncommon for geographic data to be distributed without the CRS specified. This would be analogous to telling someone the temperature without indicating Celsius or Fahrenheit. The correct scale may be obvious from context, but using the wrong scale would lead to wildly incorrect conclusions. You cannot work with geographic data without interpreting the coordinates in *some* CRS, and assuming an incorrect CRS will lead to nonsensical or misleading results.[16]

All projections also have specific parameters; for instance, many projections are centered somewhere, so they have greatest accuracy close to the specified center point. Some CRS, such as the *state plane coordinate system*, are locally parameterized to have good accuracy on particular states. This makes them appropriate for use in mapping small areas, such as a county or group of counties within one state. Others, such as the USA Contiguous Albers Equal Area Conic are parameterized for mapping larger areas like the continental U.S.[17]

In going from the three-dimensional, uneven Earth to a flat map, all projections necessarily distort. Projection methods can be *conformal*, meaning that they preserve angles as measured at any point; *equivalent*, meaning that they preserve the relative areas of features; or *equidistant*, meaning that with respect to one or two special points, distances are faithfully represented.[18] Conformal projections

[15] To be more complete, here are some of the elements that are built into the process of projecting geographic data. Latitude and longitude coordinates must be interpreted with respect to a *geographic coordinate system* (GCS) and a *datum*. A GCS is a set of parameters that translate between the Earth's shape—an oblate spheroid—and a perfect sphere. Part of the GCS called the datum can be thought of as an anchor point for the GCS (so the precision of the projection will be greatest near that anchor point) [6]. GCS and datum considerations are unlikely to matter to most redistricting practitioners and researchers, but can be of life-or-death importance in engineering and military applications.

[16] For new GIS users, one of the most frequent mistakes is to *assign* a CRS to a spatial layer when one actually wants to *transform the coordinates* of the spatial layer to a new CRS. This is like changing "30 degrees Celsius" to "30 degrees Fahrenheit," rather than applying the arithmetic transformation to yield "86 degrees Fahrenheit." The telltale sign of this kind of mistake is having spatial layers that don't align with each other when viewed in GIS.

[17] Confusingly, detailed mapping is called "large scale" while zoomed-out mapping is called "small scale" in some geography language. To see why, look at Figure 3 and note that 1/500,000 is a larger fraction than 1/20,000,000, even though that figure is more detailed. Other CRSs like the Albers Equal Area Conic can work across the globe but parameters must be set based on the area and scale to be mapped.

[18] Interestingly, to achieve this, one further cheat is needed: a special point on the sphere will typically have to be represented by an arc on the flat representation.

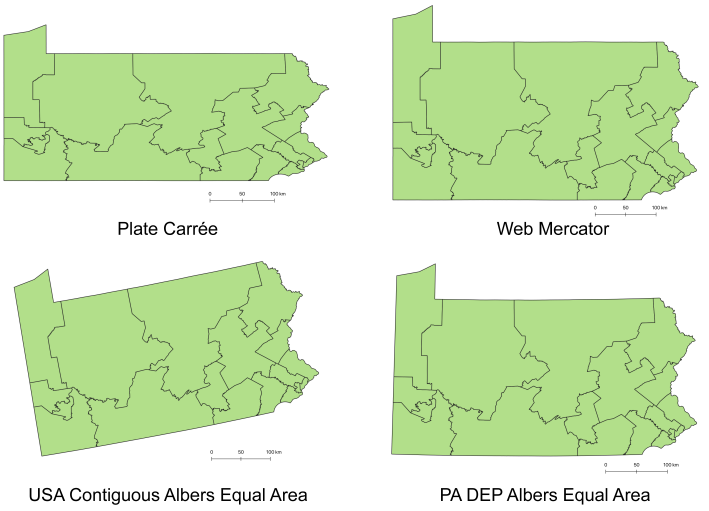

Figure 7: Pennsylvania in four map projections.

may be the best choice for navigation purposes because the right angle of a street intersection will appear as a right angle on the map; equivalent projections may be most appropriate for demographic mapping. There are compromise projections that try to balance different kinds of distortion, especially in zoomed-out mapping (e.g., maps of the world), but no projection can be conformal, equivalent, and equidistant at the same time. Of course, the curvature of the Earth is imperceptible over small distances, so for locally parameterized CRS, such as state plane zones, there is very little loss in this tradeoff. The state plane zones are all conformal, and they have a maximum scale distortion of 1 part in 10,000, for a maximum measurement error of 52.8 feet in 100 miles, which is generally considered to be acceptable for most applications.

FILES AND SOFTWARE

Broadly speaking, the industry giant in GIS is the multi-billion dollar ESRI corporation, which makes a package called ArcGIS that is so ubiquitous that for many casual users it is referenced interchangeably with the whole idea of GIS. The most common vector file format in redistricting analysis is the shapefile, a format created by ESRI in the 1990s and eventually published as a standard for data interoperability [10].[19] The format is widely supported by government agencies, including the Census Bureau.[20] A widely used free and open source alternative is QGIS.

[19]There is some ambiguity as to whether the shapefile format can be considered an "open standard." It is not included among the standards published by Open Geospatial Consortium, the major international geospatial standards body [21]. The published technical documentation is incomplete, and in 2012, geospatial programmers reverse-engineered the unpublished spatial indexing component of the format [15]. There are many other geospatial data formats, but the shapefile is so ubiquitous that there is a temptation (which should be avoided!) to think refer to *any* geospatial data set as "a shapefile."

[20]When gathering information for spatial entities such as political and administrative geographies, the possible attributes are limitless. However, many software applications are limited in the number

ESRI also makes packages specific to redistricting, and some states have built state-specific ArcGIS plugins, such as the RedAppl application in Texas, which is notoriously associated with ruthless redistricting in Texas by some authors [11]. Within redistricting, the dominant commercial software package is Maptitude for Redistricting (often shortened to Maptitude), made by a smaller company called Caliper.[21] And there are a number of others, notably autoBound (by CityGate GIS), that are selected by various localities for their line-drawing purposes.

It's important to realize that redistricting in these software programs is essentially all based on human selection (using keyboards and mouses), rather than relying on algorithms that draw the lines for you.[22] They display building blocks with data visible on a dashboard and let the user assign each building block to a district interactively.

Today there is growing momentum toward web-based, free, and open-source alternatives. Three of the most popular available tools are Dave's Redistricting App (davesredistricting.org), made by a volunteer team of Microsoft-affiliated engineers; DistrictBuilder (districtbuilder.org), made by a company called Azavea; Districtr (districtr.org, Figure 8), made by the MGGG Redistricting Lab. The impetus behind this open software push is to demystify and democratize redistricting for the public.

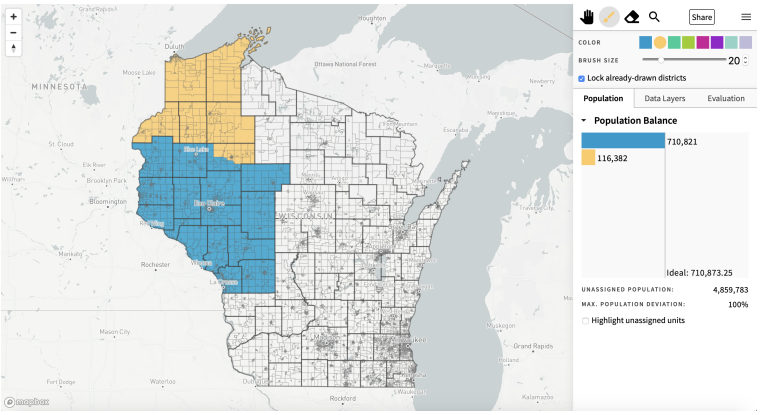

Figure 8: A screenshot from Districtr.

of columns in a single table. For the American Community Survey, for example, although the Census Bureau does indeed manage the entire (very large) dataset, the attribute data are split among many tables merely to keep file sizes manageable.

[21] As a further indication of ESRI's dominance of the GIS market, Caliper publishes Maptitude for Redistricting as both standalone software and as an extension for ArcGIS.

[22] In their 2020 release, Maptitude included some algorithmic districting options for the first time, but they are quite clearly not the primary mechanism for line-drawing. See caliper.com/mtredist.htm.

4.2 JOINING DATA TO GEOGRAPHIES

ONE-TO-ONE JOINS

We have already encountered a case where two kinds of data must be joined: the spatial data of precincts and the attribute data of vote totals (see Table 13.1). You can perform joins within desktop GIS, but for complicated operations many users prefer to work with spatial database systems or geospatial data packages within programming languages like Python and R.[23] All that is needed is a matching key, such as GEOIDs, that specifies the correspondence between the entities in the two datasets. This is another extraordinarily useful aspect of working with census data: they are painstakingly constructed to facilitate joining.

By contrast, electoral data are frequently a mess. Different agencies may use different and poorly documented conventions and abbreviations (even within the same state). Joins between these data sources may therefore be difficult without manual inspection, or by applying methods that are not usually available in desktop GIS such as natural language processing or fuzzy matching. Matching on names rather than unique identifiers also creates the potential hazard of multiple matches.

ONE-TO-MANY AND SPATIAL JOINS

A more complex but very useful operation is a one-to-many join with aggregation. This is when one row in the spatial data or tabular attribute data is joined to many rows in the other table, and then an additional operation is performed to combine some of the rows. For example, if the analyst has polling place wait times, and wants to create a map showing the average wait time by county, this requires both aggregation (grouping the polling places by county and averaging the wait times) and joining the resulting average to the county layer. These steps can be accomplished in desktop GIS, but can often be executed more easily and reliably in a programming language where you can write code to perform the operation systematically, saving the script for later inspection.

Finally, an analyst may have two sources of data that need to be joined spatially by common locations. For example, consider the polling place data mentioned above. Assigning each polling place to a Congressional District requires a *spatial join* that queries the districts to figure out where each polling place belongs. Spatial joins are more "expensive" (computationally intensive) operations than attribute joins; with large datasets such as census blocks, spatial joins can take minutes or hours to run on desktop-grade processors. Spatial joins can be sped up with proper spatial indexing (a technique that organizes spatial data) and by writing results to disk so that certain expensive operations don't have to be repeated.[24]

[23] PostGIS is a major spatial database program. The Python package called geopandas and the R package called sf are also popular choices.

[24] For instance, we were able to join Pennsylvania's 400,000+ census blocks to a set of hypothetical Congressional Districts in PostGIS in about 20 minutes without spatial indexes, or 2 minutes with properly defined spatial indexes.

INTERPOLATION

If data are provided for smaller units that nest in larger ones, then it can be joined to the larger units by aggregation, as we've seen. But if data exist on one set of geographic units (like CVAP on block groups) but needs to be joined to a different, unrelated set of geographic units (like precincts), then a more sophisticated operation is needed. The problem of estimating attribute data for a new set of *target* units based on data provided for a set of *source* units is known in geography as the areal interpolation problem.

Areal interpolation methods can be classified as *intelligent*, meaning they make use of ancillary information such as land use or census data, or as simple, which incorporate no additional data. Simple areal weighting, which allocates data to a target unit proportionate to the area of overlap between the target and source units, is widely used due to its simplicity [8]. Areal weighting requires no external data and can be performed using most desktop GIS software without special code or plugins. For many voting-related or demographic variables, however, areal weighting is a poor choice. When trying to interpolate CVAP from block groups to census blocks, for example, areal weighting will award the largest census block the most population despite the fact that the largest blocks by area often have the smallest populations (as in Sidebar 13.1).

Many official state election data products, such as in North Carolina, Wisconsin, and Texas, use an intelligent areal interpolation method that Schroeder terms "target count weighting" [22].[25] This method relies on a third, smaller *control* unit that nests in both the source and target units and acts like a common denominator to do the operation. The Wisconsin State Legislature's Legislative Technology Services Bureau, for example, releases a decade's worth of election data interpolated onto a single election year's precincts by essentially disaggregating votes from an older set of precincts (the source units) to census blocks (the control units) based on total population, and then aggregating the votes from the blocks up to the new set of precincts (the target units).

5 SOME SPECIFIC CHALLENGES

DATA VINTAGES

As geographies and associated data change over time and show no exact agreement between data sources, it is preferable to use geographies and demographic data from the same source (or at least from compatible sources) and also from the same time stamp. For example, American cities and municipalities often change borders due to annexation or secession. Official populations may grow or decline suddenly, and this growth or decline may be due to a gain or loss of legal territory.

[25] For examples in state data, see `https://www.ncleg.gov/Files/GIS/Base_Data/2016/Numeric/Data_Processing_Notes_2016.pdf` (NC), `https://redistricting.capitol.texas.gov/docs/pubs/Data_2011_Redistricting.pdf` (TX), and `https://data-ltsb.opendata.arcgis.com/datasets/2012-2020-election-data-with-2020-wards/explore` (WI).

Census geographies are released in *vintages* named for the release year, like wines! For the Decennial Census and ACS 1-year Estimates, the matching year is obvious. For ACS 5-year Estimates, there is a convention that the matching year is the final year of the period: for instance, the 2013–2017 ACS 5-year Estimates should be used with the 2017 vintage TIGER/Line files.

The Census Bureau continually updates the TIGER/Line files, and they publish information on changed geographies every year. For example, in 2016, Florida, Minnesota, North Carolina, and Virginia reported Congressional Districts that changed due to a mid-decade redistricting. These might be considered "real" changes to the geography of Congressional Districts. But on the other hand, if you compare the Congressional District geometries from the 2015 and 2016 TIGER/Line files, you would find that 386 of 444 features show changes between these years.[26] *These are not reflecting changes to the legal boundaries of the districts!* The updated geographies are entirely due to tiny adjustments to the census blocks that are regarded by the Bureau as accuracy improvements, which might add vertices to the polygons or move vertices by a matter of feet or inches. Thus, you cannot simply compare geographies through shapefiles directly to find out if they have changed from year to year. And you also have to be careful if your layers come from different vintages, because operations like intersection, subset, and shared boundary can go awry if the entities are recorded in a subtly mismatched way.

MEASURING COMPACTNESS

Earlier, in Chapter 1, you read about measures of *compactness*, which are aimed at describing the shapes of districts as either eccentric and convoluted or simple and regular. Certain compactness metrics, particularly those based on area versus perimeter, have been shown to be highly sensitive to changes in resolution and projection [2, 9, 16]. This is a practical impact of the choice of coordinate systems beyond "squashed" or distorted appearances (see Figure 7).

Each state might have a different best choice for its coordinate system and projection for redistricting. The Pennsylvania Department of Environmental Protection uses an Albers Equal Area projection that is locally parameterized to be suitable for displaying the entire state, making it reasonable to use for compactness measures (although the user must set some parameters to do so). The lack of a canonical choice across or even within states is a great reminder that even objective-sounding metrics actually require quite a substantial amount of user discretion.

WATER AND ADJACENCY

Speaking of choices that are under-specified, consider the question of whether to include bodies of water in electoral districts, and whether to consider areas separated by water as being adjacent. There is no universal standard about the

[26] This number includes 435 districts with voting members, 6 non-voting delegate districts, and 3 water areas that are legally parts of the territory of some state but not assigned to any Congressional district.

inclusion of water. To see the strange effects this can produce, look at the map of Michigan's congressional districts shown in Figure 9.

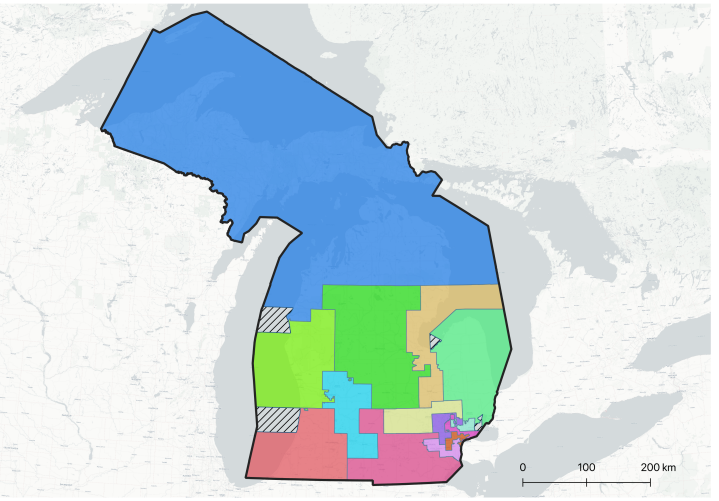

Figure 9: Michigan Congressional Districts drawn after the 2010 Census. Note that the districts zigzag in and out of the lake—the diagonal hatched regions are lake areas not assigned to any district in the TIGER/Line files.

The sawtooth pattern turns out to be explainable as a GIS artifact. Michigan's 6th District (bottom left in the figure) includes the entirety of Van Buren County, but only part of the area of Allegan County. The district geography is therefore constructed from all of Van Buren County, which legally includes the water area in Lake Michigan, and Allegan county subdivisions, tracts, or blocks, for which the water area is omitted.

Water is also important when it comes to thinking about what land area is next to what, which is important since districts are typically expected to be contiguous (i.e., made up of one connected piece). In some cases, redistricting seems to have been done without regard for physical geography, like when land areas separated by water are combined in a single district. But this can be necessary, such as for islands, and it can also be done to further other legitimate redistricting goals (such as complying with the Voting Rights Act or respecting communities of interest).[27]

For example, New York City covers part of the continental mainland and three islands (Staten Island, Manhattan Island, and Long Island), divided into five major pieces called boroughs. When drawing congressional districts after the 2010 Census, Staten Island was kept whole but had to be connected to *some* territory across water in order to get up to the needed population. On the other hand, Manhattan had more than two districts' worth of population. Rather than placing any

[27] When researching state House districts in Alaska, Caldera et al. found that the level of restrictiveness applied to water adjacency had profound implications for minority representation. When districts were randomly drawn on the more permissive adjacency network, there were significantly more House districts with a majority of Alaska Native population than when working with the restrictive network with fewer edges [3].

district wholly within Manhattan, the island was split among four congressional districts, all of which cross water to include significant territory and population from Brooklyn, Queens, or the Bronx.

So even when districts are required to be contiguous, the rules for crossing water are unclear for legislatures or for courts, and therefore for researchers trying to understand the space of possibilities. Islands off the coast may make up their own census units, but may also be part of a larger geography that includes the mainland and the water in between. When geographies are not physically adjacent, should allowable connections be based on distance measurements, or the transportation network, or demographic similarity? This is another case of interpretive ambiguity that becomes very visible at the level of geospatial data representation, but can't be resolved without deliberation on the larger goals of good redistricting.

6 CONCLUSION: TECHNOLOGY CUTS BOTH WAYS

There are widespread misconceptions about both the availability of reliable geospatial election data and the ease of manipulating it. It doesn't help that popular media sources publish many maps that look like precinct-level election results—for instance, on election night. Although these maps may look authoritative and reliable, the data sources for them are unclear at best: in many cases, no public sources *could have* provided the data that are displayed. This creates an impression of availability that combines with other factors—including confusion about the control of precincts and commercial incentives against widespread public access—to impede reform and modernization of data maintenance and transparency practices.

In this chapter, we have tried to show some of the ways that data you may have taken for granted are actually built and handled. Some are meticulously curated by the Census Bureau and some are scraped and digitized and re-shaped and matched by many hands—and until there is regulatory reform, election data will stay scattered and messy. Within best practices, there are still a substantial number of user choices to make.

We are often asked for our opinions on whether, on balance, the trend toward more powerful geospatial technology has made gerrymandering worse.[28] The blatant partisan gerrymanders following the 2000 and 2010 Censuses do seem to coincide with the increasingly widespread adoption of GIS. However, geospatial technology is just another tool being used to redistrict, and the sophistication attributed to gerrymanderers seems to us to be overblown. While many of the professional redistricting consultants use proprietary software and commercialized data, researchers and engaged members of the public can benefit from the current boom in the open data and civic tech movements, and from rising attention to redistricting from diverse academic researchers. We have considerable hope for the future.

[28] Many authors have considered this question, such as McGann et al., who point to the availability of computer software for redistricting prior to 2000, among other factors, in dismissing the idea [18].

REFERENCES

[1] Anderson, Margo J. 2015. *The American Census: A Social History.* 2nd ed. New Haven: Yale University Press.

[2] Barnes, Richard, and Justin Solomon. 2018 "Gerrymandering and Compactness: Implementation Flexibility and Abuse." *arXiv:1803.02857[Cs],* March. http://arxiv.org/abs/1803.02857

[3] S. Caldera, D. DeFord, M. Duchin, S. Gutenkust, and C. Nix. "Mathematics of Nested Districts: The Case of Alaska." Statistics and Public Policy (2020).

[4] Chang, Kang-Tsung. 2017. "Geographic Information System." Edited by D. Richardson. *International Encyclopedia of Geography: People, the Earth, Environment and Technology,* March. Wiley. https://doi.org/10.1002/9781118786352.wbieg0152.

[5] Clapp, John M., and Yazhen Wang. 2006. "Defining Neighborhood Boundaries: Are Census Tracts Obsolete?" *Journal of Urban Economics* 59 (2): 259–84.

[6] Clarke, Keith C. "Geodesy." In International Encyclopedia of Geography: People, the Earth, Environment and Technology, edited by Douglas Richardson, Noel Castree, Michael F. Goodchild, Audrey Kobayashi, Weidong Liu, and Richard A. Marston, 1–10. Oxford, UK: John Wiley & Sons, Ltd, 2017. https://doi.org/10.1002/9781118786352.wbieg0456.

[7] Code, Ohio Revised. 1977. "3503.17 Precinct Boundary Changes."

[8] Comber, A., W. Zeng, Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. Geography Compass. 2019; 13:e12465. https://doi.org/10.1111/gec3.12465

[9] Duchin, Moon and Bridget Tenner. *Discrete geometry for electoral geography*, preprint. Available at https://arxiv.org/abs/1808.05860.

[10] ESRI. 1998. "ESRI Shapefile Technical Description." White Paper J-7855. Redlands, CA: Environmental Systems Research Institute, Inc.

[11] Forest, Benjamin. 2004. "Information Sovereignty and GIS: The Evolution of 'Communities of Interest' in Political Redistricting." *Political Geography* 23 (4): 425–51. https://doi.org/10.1016/j.polgeo.2003.12.010.

[12] General Assembly of Pennsylvania. 2011. "Congressional Redistricting Act of 2011."

[13] Harrington, Jan L. 2016. *Relational Database Design and Implementation.* 4th ed. Amsterdam; Boston: Morgan Kaufmann/Elsevier.

[14] *Harvard Election Data Archive.* https://projects.iq.harvard.edu/eda.

[15] Lawhead, Joel. 2011. "Your Chance to Make GIS History." *GeospatialPython.com.*

[16] Li, Wenwen, Michael F. Goodchild, and Richard Church. 2013. "An Efficient Measure of Compactness for Two-Dimensional Shapes and Its Application in

Regionalization Problems." *International Journal of Geographical Information Science* 27 (6): 1227–50. https://doi.org/10.1080/13658816.2012.752093.

[17] MacDonald, Karin. 2019. "California's Statewide Database." Presentation at the Voting Rights Data Institute, Tufts University, July 10, 2019.

[18] McGann, Anthony J., Charles Anthony Smith, Michael Latner, and Alex Keena. 2016. *Gerrymandering in America: The House of Representatives, the Supreme Court, and the Future of Popular Sovereignty.* New York, NY: Cambridge University Press.

[19] Mule, Thomas. 2012. "Census Coverage Measurement Estimation Report: Summary of Estimates of Coverage for Persons in the United States." 2010-G01. Washington, DC: U.S. Census Bureau.

[20] Office of the Minnesota Secretary of State—Elections Division. 2019. "2019 Minnesota Election Laws."

[21] Open Geospatial Consortium. n.d. "Open Geospatial Consortium Standards." http://www.opengeospatial.org/docs/is.

[22] Schroeder, J.P. (2007), Target-Density Weighting Interpolation and Uncertainty Evaluation for Temporal Analysis of Census Data. Geographical Analysis, 39: 311–335. https://doi.org/10.1111/j.1538-4632.2007.00706.x

[23] U.S. Bureau of the Census. 1994. *Geographic Areas Reference Manual.* Washington, D.C.: U.S. Dept. of Commerce, Economics, and Statistics Administration, Bureau of the Census.

[24] United States Census Bureau. "Standard Hierarchy of Census Geographic Entities," October 27, 2010. https://www2.census.gov/geo/pdfs/reference/geodiagram.pdf?#.

[25] U.S. Bureau of the Census. 2018. "Redistricting Data Program Management." https://www.census.gov/programs-surveys/decennial-census/about/rdo/program-management.html

[26] U.S. Bureau of the Census. 2014. "Establishment of the 2020 Census Redistricting Data Program." 79 FR 41258.

[27] Weibel, Robert. 1997. "Generalization of Spatial Data: Principles and Selected Algorithms." In *Algorithmic Foundations of Geographic Information Systems,* edited by G. Goos, J. Hartmanis, J. Leeuwen, Marc Kreveld, Jürg Nievergelt, Thomas Roos, and Peter Widmayer, 1340:99–152. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-63818-0_5.