Chapter 6

GIS, Historical Research, and Microdata

GIS Data

Since most census data is geographic in nature, it lends itself well to geographic analysis. Geographic Information Systems (GIS) are collections of software and data for conducting geographic analyses and making maps. Special GIS data files called vector files store coordinates to form geometries that represent points, lines, and areas. Each file represents a particular type of feature that covers a specific extent: a file for counties for the US, a file of census tracts for a state, or a file of roads for a particular county. The files are georeferenced, which means they are drawn to scale using a specific spatial reference system that ties them to real locations on the earth. These reference systems allow GIS data files from multiple sources to be overlaid in a GIS project. In addition to the vector files, a different format called a raster represents continuous surfaces as a series of grid cells of equal size, where each cell has a value that represents something about the surface. Rasters are also georeferenced, and satellite imagery, air photos, and scanned paper maps such as topographic maps are stored in the raster format. For census mapping applications, rasters are useful as base maps to provide context for vectors.

Desktop GIS software provides a lot of power for cartography and geographic analysis, particularly in combination with GIS data from multiple sources. For example, with census tract boundaries, population data for those tracts, and a point layer of public libraries created from the Institute of Museum and Library Services data files, you can map census tracts by the population under eighteen and select all tracts that fall within a mile or two of a public library. This allows you to measure youth population near each library, as well as measuring the population that falls outside this zone. You can measure the distance from each tract to the nearest library to generate measures of accessibility for different areas. Vector layers can be overlaid on top of raster maps or web mapping services, such as the OpenStreetMap, to provide context for your layers. We will explore census mapping for LIS research in the final chapter.

The Census Bureau publishes vector GIS files that represent all of the legal and statistical areas that they publish data for, as well as files that represent features that are used for drawing these areas, including roads, railroads, water bodies, and other landmarks. These GIS files are part of the TIGER geographic database, which is the system the Census Bureau uses to update and maintain all of the geographies it uses for its operations. The features in this database follow the geographic summary level and nesting rules to maintain structural integrity of boundaries so that the boundaries of smaller features fit within the appropriate larger ones.

All of the TIGER files are published on the Census Bureau's website and can be freely downloaded for use in desktop GIS software and in web mapping applications. The vector format is published in several different file types, the most common being the ESRI shapefile, which is an open legacy format that's widely supported. Other options include Google KMLs, GeoJSON, and the native TIGER file format. Most vector formats can be used in any GIS package. ArcGIS Pro is the best-known proprietary GIS package, while QGIS is a popular free and open source alternative.

The Census Bureau publishes several different iterations of the TIGER files for different mapping purposes. The official TIGER files represent the precise legal and statistical boundaries delineated by the Bureau. Since the boundaries of features incorporate both land and water, the shapes may appear unusual at first glance and may not be ideal for thematic mapping. The Bureau publishes a derivative of TIGER called the Cartographic Boundary Files, where large coastal water bodies have been removed so that features better represent land areas. The linework of these files has been generalized to smooth out boundaries and remove small features like islands that

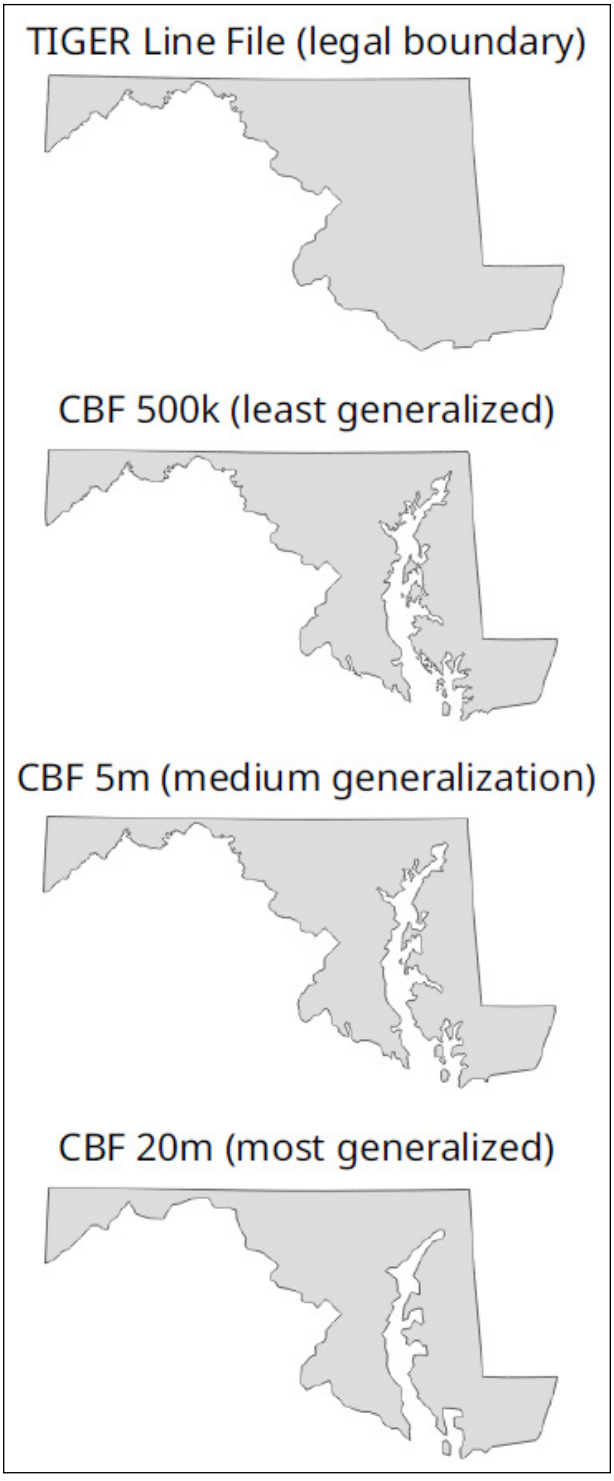

Figure 6.1
Boundary differences between the TIGER files and the Cartographic Boundary Files for Maryland

would not be visible at certain scales. There are different iterations of the boundary files that are appropriate for different scales; the most generalized would be appropriate for a map occupying a postcard, as opposed to the least generalized version that would be more appropriate for a poster. Figure 6.1 illustrates the differences.

Census GIS Resources

TIGER/Line Shapefiles
https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

Cartographic Boundary Files
https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html

Census Geocoder
https://geocoding.geo.census.gov/geocoder

In shapefiles, attributes about the features are stored in a table where each row represents a geographic feature and each column contains attributes that describe the feature. The features that are visually depicted in the software's map view and the records in the table are tied together so that you can use the attributes to label, filter, query, and thematically map the features. In GIS packages, attribute tables operate according to similar principles as tables in relational databases. Columns are designated to store text, integers, or decimals, and one column serves as a unique identifier or primary key. In most instances, vector files do not come with lots of attribute data preattached. Users can add nonspatial data tables, where each row represents a specific geography and columns contain population or socioeconomic data, to GIS software and then join that data table to a corresponding vector file that represents the same features, using a unique ID column that they share in common. This allows you to create thematic maps of the population data, such as shaded-area maps where ranges of population values are classified into categories assigned a specific color.

Mapping census data in GIS is a multistep process that involves downloading the TIGER file that represents the features, downloading population data that contains the attributes you wish to map for those features from data.census.gov, processing the features and data table to prepare them for analysis, adding both files to the GIS package, using the package's tools to join the GIS features to the table using the Census Bureau's GEOID column (which has the unique summary level and ANSI/FIPS codes), and using the GIS package tools to symbolize the features based on the attribute you wish to map. Most of the data processing steps are associated with taking large areas or large numbers of variables and creating subsets out of them or conversely stitching together multiple shapefiles or

data tables into a larger set. All the TIGER files and data tables from data.census.gov come with census identifiers in multiple forms, including the full GEOID field, so features and data tables can be readily joined. If you retrieve data from the API or another source, you will want to be sure that you also retrieve identifiers that will match the vector file.

While GIS gives you a lot of power and flexibility, the learning curve can be steep. The Census Bureau does package pre-joined vector files with a selection of basic population data to create special geodatabases. These can lower the curve by eliminating data processing work. For non-GIS users, the Bureau publishes a number of static maps depicting census boundaries, as well as selections of thematic maps. For users who want to do basic data exploration and mapping, data.census.gov and other free tools such as the Census Reporter provide basic web mapping capabilities, where you can select a type of geography and variable and make a map. There are also a number of proprietary library database products that public and academic libraries subscribe to. These tools add value by pulling a lot of disparate data together in one place and providing easy-to-use tools for creating thematic maps, doing basic querying and analysis, and downloading data.

Census Database and Mapping Products

PolicyMap
https://www.policymap.com/

SimplyAnalytics
https://simplyanalytics.com/

Social Explorer
https://www.socialexplorer.com/

GeoLytics
https://www.geolytics.com/

Historical Data

Historical analysis, whether it's studying recent or long-term trends, presents a number of challenges. In terms of access, most public data portals such as data.census.gov and the APIs do not archive data that's more than a few decades old. There are also few census products that compare change between two periods in a single table or file. The ACS comparative profiles compare two nonoverlapping five-year periods. There are no official DEC tables (i.e., part of the summary files with official table identifiers) that compare two successive censuses, but there are some special products published on the DEC program websites.

The primary public archive for all historical summary data is the National Historical Geographic Information System (NHGIS). This is one of many projects that's part of the IPUMS series of repositories created at the Minnesota Population Center at the University of Minnesota (Kugler and Fitch 2018). Users must register and create an account, but it's an academic, nonprofit project and registration is free. Users are asked to cite the NHGIS as the source and use the data for noncommercial purposes. Summary data from every DEC, all versions of the ACS, pre-twenty-first-century data from the Business Patterns, and several other special datasets are available. GIS vector boundary files for each decade are also available for depicting areas as they appeared for each decennial census. As in the advanced search in data.census.gov, users apply filters to choose a dataset, year, geography, and topic and can then browse through the returned tables and download them. Most of the data is presented nominally, the way it was originally published. The NHGIS has assembled a limited series of time series tables that allow you to compare the same or similar variables across several decades, from the 1970 census to the present.

One of the challenges in studying the census over time is that census geography changes; statistical areas are updated for each DEC, and legal areas change on an ongoing basis. One approach for grappling with this challenge is to use normalized instead of nominal data. Normalized data has been modified so that data from several points in the past has been altered to fit present-day boundaries, so comparisons within the same area can be made. Census tracts are used as the basis for creating normalized data, as they have been drawn and numbered in a way to insure some degree of consistency over time. Converting the past data to modern boundaries is achieved by aggregating tracts (if two old tracts were eventually combined into one), splitting tracts (if one old tract was subsequently split into two), or apportioning data in instances where a split was more complex.

The NHGIS publishes a series of normalized data tables (referred to as geographically standardized tables) from the 1990 to 2010 census using 2010 geography, for states down to block groups. The Longitudinal Tract Data Base produced at Brown University publishes a crosswalk for users to normalize their own data for census tracts from 1970 to 2010 using 2010 geography, and it publishes a limited number of normalized data tables for users to download (Logan, Xu, and Stults 2014). The GeoLytics company has been a long-time producer of normalized census data, and its Neighborhood Change product provides normalized tract data from 1970 to 2010. Once the data for the 2020 census is fully released, it is likely that each

organization will create new series with data normalized to 2020 boundaries.

Other resources allow users to adjust nontract geography over time. For each DEC, the Census Bureau publishes block relationship files that crosswalk blocks from the previous DEC to the current one, as blocks are completely renumbered and their boundaries can change. As blocks are the smallest geographies, they can be aggregated to form any geography in the hierarchy. The Bureau also publishes an ongoing list of changes to county boundaries from 1970 to the present that can be used to adjust data. While county-level changes do occur, they are less frequent in modern times and most adjustments are straightforward. The composition of metropolitan areas changes quite frequently, at least two or three times a decade. Since metro areas are composed of counties, county-level data can be aggregated to create consistent metro definitions for comparison over time.

Historical Census Data Resources

NHGIS
https://www.nhgis.org/

Longitudinal Tract Database
https://s4.ad.brown.edu/Projects/Diversity/Researcher/LTDB.htm

Census Block Relationship Files
https://www.census.gov/geographies/reference-files/time-series/geo/relationship-files.html

Changes to Counties, 1970 to Present
https://www.census.gov/programs-surveys/geography/technical-documentation/county-changes.html

Historical Census Reports
https://www.census.gov/prod/www/decennial.html

National Archives Census of 1940
https://1940census.archives.gov/

National Archives Census of 1950
https://1950census.archives.gov/

IPUMS Complete Count Data
https://usa.ipums.org/usa/complete_count.shtml

For contemporary historical analysis, the 1970 census is often the initial entry point of comparison for reasons described in chapter 3. It was the first census conducted under the basis of self-identification and the first conducted entirely through the mail. The race and ethnicity categories that we use today had their origins in 1970, although the 1980 census was the first where the standardized Directive 15 categories were used on the 100 percent count form. The digital predecessor to TIGER was developed for the 1970 census, so there were readily available map files to work from, and by this period a significant portion of the country was covered by census tracts (or some corollary), allowing for the creation of normalized files. Fewer geographies are available as you go back further in time, and fewer that cover the entire nation. Census tracts were introduced as a census geography in the 1940 census, and only for urban areas.

Questions on the forms and categories for summarizing data also change, and accommodating these changes is more difficult. As mentioned previously, changes for NAICS can be accommodated through using concordances from one version to the next, but while there is a bridge between the 1997 NAICS and the 1987 SIC system, there are quite a few differences between them. Adjusting dollar values for inflation is straightforward and an absolute must when comparing financial statistics. The Pew Research Center (2020) has created a chart that records how the nation's racial categories have changed over time. In some cases statistics for changing categories can be quantified or estimated, such as the shift to including a multiracial option for the first time in the 2000 census, but in many cases researchers can only note the differences and advise caution in making comparisons. Studies on how specific questions or categories have changed over time, such as native language or language spoken at home (Stevens 1999), can provide guidance for making comparisons.

After seventy-two years, the original responses to the census questionnaires are released by the National Archives, opening up new possibilities for historical research. Prior to 1960, census data was captured on ledgers as opposed to a form for a single household, and these ledgers were microfilmed and sent to libraries throughout the nation. For the 1940 census, scanned PDFs of the ledgers were published on the Archives website, which allowed for remote access. This data is not machine-readable (the responses were written in longhand) and specific persons can be located only if you know their address. With the address, you could identify the enumeration district where the street is located so you can find the file/ledger with data for that street. Private companies such as Ancestry.com digitize and make the records searchable through their products. For social science researchers, the IPUMS publishes a variety of machine-readable, full count data files as part of the IPUMS USA project. The public files have been de-identified (names and addresses removed), but the full individual records are available. Data from 1900 to 1940 is linked, so individuals can be followed over time.

Microdata and the Current Population Survey

Census *microdata* refers to person- or household-level records that represent a sample of responses to census questionnaires, with personal identifying information removed. Some researchers work with microdata because they are interested in studying population or socioeconomic trends from perspectives other than geographical ones, such as broad trends across an entire population and differences in trends between different age, sex, racial, and economic groups. Other researchers use microdata to create special cross-tabulations that are not present in the public summary data.

Microdata records include a number of special attributes called weighting variables, which are used to generate population-level estimates. The weight indicates the number of people or households in the general population that a particular respondent represents. For example, to generate an estimate of the American Indian or Alaska Native population sixty-five and over, you would sum the person weights for all records where the race and age meet these criteria. Public use microdata samples (PUMS) are generated for both the DEC and ACS and can be accessed from the DEC and ACS program web pages on the Bureau's website. IPUMS USA is an alternate choice for creating DEC and ACS microdata extracts, where users generate samples by choosing variables of interest.

To protect confidentiality, estimates for small areas such as census tracts cannot be generated. The Census Bureau has created a special geography called a Public Use Microdata Area (PUMA) for presenting reliable estimates for microdata. Built from blocks and nesting within states, PUMAs are designed to have a population size of approximately 100,000 people. They are identified by a five-digit number and a place name that indicates an area of coverage. Researchers need to be on guard when generating estimates for small population groups, as the sample size may not be large enough to generate reliable estimates. There is documentation for advising users how to estimate reliability.

Beyond the DEC and ACS, the Census Bureau conducts a number of smaller surveys for which it publishes both state and national statistics as well as the microdata for the survey. Of all these surveys, the Current Population Survey (CPS) is the largest and most widely used. Launched in the 1930s, it was one of the government's earliest efforts to employ sample surveying to produce population-level estimates. It is conducted jointly by the Census Bureau and the Bureau of Labor Statistics, and its original purpose was to generate monthly unemployment numbers, a primary role that it continues to serve. The CPS samples 60,000 households a month from a stratified sample that includes every state and the District of Columbia. Census field representatives conduct the surveys in person or over the phone, which makes the responses highly reliable and guarantees a response rate of over 90 percent. The CPS is unique among the census survey programs in that it is longitudinal. The same 60,000 households are sampled monthly for four months, then are rotated out of the survey for four months, and are rotated back in for a final four months. There are a core set of questions that are asked every month, plus a supplemental set of questions that are asked on a set but limited basis. For example, the Annual Social and Economic Survey (ASEC) is a set of detailed socioeconomic questions asked every March, while questions on voter registration and participation are asked every two years in November to coincide with federal elections. The CPS captures many of the variables that are included in the ACS, plus a substantive number that aren't captured elsewhere.

Given the sample size of the CPS, the Census Bureau and the Bureau of Labor Statistics generally publish national and state-level estimates only. Researchers can access all the microdata records from the CPS program page or can create specific extracts from the IPUMS CPS website. Since the survey is longitudinal, there are identifiers that allow researchers to track person and household responses across several months of the survey. There are a number of web-based tools for users who want to generate estimates without having to download and weight the records themselves. For example, the Census Bureau's Micro Data Access Tool (MDAT) and the IPUMS Analyzer allow you to generate estimates from samples in the ASEC.

Microdata Resources

IPUMS USA
https://usa.ipums.org/usa/

IPUMS CPS and Online Analyzer
https://cps.ipums.org/cps/

Current Population Survey
https://www.census.gov/programs-surveys/cps.html

Micro Data Access Tool (MDAT)
https://data.census.gov/cps/mdat/

Census Research Data Centers
https://www.census.gov/about/adrm/fsrdc/locations.html

In addition to the PUMS, there are also restricted microdata samples that researchers can access by submitting research proposals to the Census Bureau.

Strict procedures must be followed to ensure that the confidentiality of responses is not jeopardized. If accepted, the researcher can visit one of several Census Research Data Centers to work with the data onsite. The centers are located throughout the United States, usually at large research universities.

References

Kugler, Tracy A., and Catherine A. Fitch. 2018. "Interoperable and Accessible Census and Survey Data from IPUMS." *Scientific Data* 5: article 180007. https://www.nature.com/articles/sdata20187.

Logan, John R., Zengwang Xu, and Brian J. Stults. 2014. "Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database." *Professional Geographer* 66, no. 3: 412–20.

Pew Research Center. 2020. *What Census Calls Us: A Historical Timeline*. Washington, DC: Pew Research Center, February 6. https://www.pewresearch.org/interactives/what-census-calls-us/.

Stevens, Gillian. 1999. "A Century of U.S. Censuses and the Language Characteristics of Immigrants." *Demography* 36, no. 3: 387–97.