THE NATIONAL HISTORICAL GEOGRAPHIC INFORMATION SYSTEM (NHGIS)

Robert B. McMaster, Mark Lindberg, and David Van Riper

Department of Geography
University of Minnesota
Minneapolis, MN  55455
mcmaster@umn.edu

ABSTRACT

The National Historical Geographic Information System (NHGIS) is a five-year NSF-funded project designed to create a comprehensive U.S. census database—at the census tract and county level—for both the geographical and attribute data. Technological change presents an unprecedented opportunity to make these data readily available for social science research; thus bringing the complete census within reach of social scientists unlocking the potential of two centuries of data collection, and stimulating research in economics, history, sociology, geography and other fields.

The project consists of three major components: data and documentation, mapping, and data access.

- The data and documentation component gathers all extant machine-readable census summary data; fills holes in the surviving machine-readable data through data entry of paper census tabulations; harmonizes the formats and documentation of all files; and produces standardized electronic documentation according to the recently developed Data Documentation Initiative (DDI) specification.
- The mapping component creates consistent historical electronic boundary files for tracts, counties and larger geographic units.
- The data access component creates a powerful but user-friendly web-based browser and extraction system, based on the new DDI metadata standard.

INTRODUCTION

The United States Census, which dates back to 1790, collects a massive amount of data on the characteristics of the population every ten years. Commonly called the Population of Census and Housing, it is one of the core data sets for social science researchers, and enables geodemographic analyses at multiple scales and data resolutions. Social science researchers and others have utilized the census to study myriad social problems over the past two centuries, including changing patterns in poverty, income, occupation, and housing. A small percentage of these studies, however, have focused on the spatial dimension of the problem as the census tract and other boundaries were not readily available. The two major goals for the NHGIS project include the creation of a comprehensive spatio-temporal database at tract and county levels for the entire United States and enabling robust spatio-temporal analyses of census data, that is, allowing comparisons of census data with different enumeration boundaries through interpolation.

The three major stages of the NHGIS project include:

- Gathering Census Data
  - Collect historical and contemporary Census summary data on population and housing, agriculture, and others
- Creating Boundary Data
  - Geocoding Census tract boundaries back to 1910
  - Geocoding County boundaries for each census year back to 1790
- Developing a Web-based Access System
  - Creating Metadata
  - Integrating Census data
  - Integrating Boundary data
Project Goals

The overall goals for the NHGIS project are to create a comprehensive spatio-temporal database at tract and county levels for the entire United States, and to enable robust spatio-temporal analysis of census data – comparing census data with different enumeration boundaries through interpolation.

Acquiring statistical data

The United States government has been gathering data on its population since the late 1700s. While we mostly think of the Census of population, there are other censuses as well, including those that gather information on the economy and agriculture. The NHGIS project will gather as much data as possible at the county-level, and all tract level data dating back to 1910, when the first cities in the United States were tracted. The following illustration (Figure 1) details the various censuses at the county level, illustrating the temporal availability.

![Figure 1. The statistical data included in the NHGIS](image)

The census tract-level editing process.

The process of reconstructing census tract boundaries begins with the Census’ 2000 TIGER files as the base. From these files, tracts for both 2000 and 1990 can be directly obtained. The 1992 TIGER files (updated after the 1990 census) provide the 1980 tract boundaries. For earlier decades the tract-level boundaries are created backwards in time from scanned paper census maps to the first decade that a city was tracted which, for many cities in the eastern United States, dates to 1910. The enumeration unit of analysis for this tract work is the county, that is, tracts are scanned and processed together at the county-level. When complete, boundary files and a significant amount of census statistical data will be process for all census tracts and county units. As of the 2000 census, the entire United States had been tracted. County boundaries are based on tract boundaries where possible. For rural areas after 1910 and all areas before 1910, county boundaries are created backward in time, reflecting changes recorded in historical maps and other sources. The specific steps in reconstructing the tract boundaries include the following 7 steps, with further details provided in McMaster and Lindberg, 2003.

1. Assemble data
2. Construct tracts
3. Review tracts
4. Aggregate tract counties into metropolitan areas (also states)
5. Generalize aggregated files
6. Final review of files
Figure 2. The discontinuity between the scanned map and the digitized tracts

Figure 2 illustrates the discontinuity between the digitized tracts and the scanned map from the 1960 census paper maps. Using the scanned map in the background, the “digitizer” is responsible for identifying the correct census boundaries on the 2000 detailed TIGER database. In this figure the displacement between the scanned and TIGER representation can be noted.

**Editing Water Features.**

A significant problem arises with the representation of water boundaries. The water boundaries were inconsistently depicted among the decades. In certain instances, tract borders were drawn out into water while in some years they were drawn along shoreline (Figure 3). The tract boundaries along rivers were either drawn down the center or along one of the edges. Additionally, among the decades the river morphology often changed, resulting in shifting tract boundaries. Another issue is the creation of reservoirs by damming rivers/streams, thus completely altering the geometry of tract boundaries.

**Sliver Problems.**

Another difficult editing problem arises when very small sliver polygons exist in the dataset (Figure 4). Interestingly, these slivers are often “real” in that they have Tract IDs. These slivers commonly occur at county boundaries and are small enough that they might not be visible when the files are generalized. However, they create serious edge-matching problems when the tract files are edge-matched.
Complete Retracting.

Yet another type of tract-editing concern is when cities complete “re-tract” between decades (Figure 5). As an example, the City of Minneapolis completely retracted boundaries between 1990 and 2000. This situation is analogous to a complete redigitization for a given year since none of the existing boundaries are usable. Retracting is very costly in terms of the time needed to edit and review, but fortunately it is not that common.
NHGIS plans to create county boundaries for every county that was enumerated in a U.S. census. More than one-fifth of present day states will have county data available from the first census, 1790, to the most recent census in 2000. County boundaries will be produced as an extension of our work with tracts. Initial boundaries are created by aggregating census tracts—creating county boundaries by retaining tract boundaries that coincide with county boundaries and eliminating all others. This procedure has the benefit of maximizing the correspondence between the tract and county data bases. Like the tract data, therefore, the county boundaries originate from TIGER files except where boundaries have to be reconstructed from other sources. Like the tract editing, we work backwards in time starting with the most recent boundaries for the 2000 census which are essentially identical to county boundaries in the TIGER files and constructing new county footprints for each preceding census. Editors work on county boundaries on a state-by-state basis.

We are relying heavily on secondary data sources for obtaining early (pre-tract) county boundary configurations. Three sources are paramount because they cover most or all of the country. They are John Long’s series of atlases of historical county boundaries done at the Newberry Library, Thorndale and Dollarhide’s *Map guide to the U.S. federal censuses*, and Richard Forstall’s *Population of states and counties of the U.S.: 1790-1990*. In addition to these, other resources are being used, but generally are pertinent to individual states or territories. The data sources occasionally have significant differences regarding county configurations. For example, Figure X shows county configurations from two data sources for the same census year. Determining an appropriate configuration can be an involved and complex process. Some of the differences in our source materials stem from differences in their purpose. For example, Long’s series presents statutorily defined county boundary configurations but these were not necessarily used by census enumerators at the time. Some older boundaries are indefinite and we therefore expect substantial differences in how the county configurations are shown.

For Minnesota we have examples of data reported for counties that no longer existed at the time of the census and for some that were not legally organized until after the census. For example, the census for 1870 reports 140 persons for Lac Qui Parle County. It had been eliminated by the state legislature in 1868. (Another Lac Qui Parle County was formed in 1871— in a different location.) The census for 1860 reports 350 persons for Monongalia County. It was not created until 1861 however. Census enumerations exist for Rock and Pipestone Counties for all censuses between 1860 and 2000. Someone relying on the tabular data alone might miss the fact that they exchanged names in 1862. Hence in order to analyze population growth between 1860 and 1870 one would have to compare the 1870 Pipestone County with the 1860 Rock County and vice versa.
Providing unique identifiers for county configurations is a necessary step in boundary construction and ultimately for matching enumeration units with the appropriate tabular data. We developed a coding scheme that is essentially an extension of the FIPS county codes. Issues that required solutions include historical counties that no longer exist, some with the same names as contemporary counties, and counties that exchanged names. Clearly, the 1870 Rock County Minnesota should have a different code than the 1860 Rock County. The five-digit FIPS code (two-digit state plus three-digit county) was extended by two digits (one for the state and one for the county) to allow numerical space for coding historical counties and territories.

When work on the county boundaries is finished, the boundary polygons will be aggregated to form additional census enumeration entities including metropolitan areas and states. Therefore at the conclusion of the project we will be able to assemble a variety of census enumeration units as long as they are composed of aggregates of counties.

Figure 6. An example of where source materials do not match for the same census year. The map on the left is from the Atlas of historical county boundaries. Minnesota, and the map on the right is from Map guide to the U.S. federal censuses. Both illustrate county boundary configurations in 1860.

Figure 7. The 1790 and 2000 county boundaries for Minnesota
The Changing Nature of Census Tracts

At the request of Dr. Walter Laidlaw, the Census Bureau first delineated census tracts for the 1910 Census of Population and Housing. Laidlaw, the director of the Population Research Bureau of the New York Federation of Churches, analyzed neighborhood characteristics in New York City using census data tabulated for assembly districts. Unfortunately, the State of New York changed the assembly district boundaries, making the longitudinal study of neighborhoods difficult to carry out. Thus, Laidlaw proposed establishing small-area geographic units that were independent of assembly districts. The units, originally called sanitary districts and later census tracts, would have permanent boundaries and would allow for consistent comparisons of areas through time. Laidlaw’s plan called for small units with “permanent” boundaries, but population changes and decisions by local census tract committees have resulted in relatively fluid boundaries. When the Bureau or a local area census committee first established tracts in a county, protocol required an average of 1,500 housing units, or approximately 4,000 people, in each tract. The number of housing units could range from 600 to 3,000, or 1,500 to 8,000 people. Protocol also required that the demographic and housing composition of the original tracts be relatively homogeneous (U.S. Census Bureau 1994). Through time, however, homogeneity decreased as population shifted geographically and changed demographically. Changes in the geographic distribution of the population resulted in tract boundary changes (Figure 8). The Bureau wanted tracts to satisfy the protocol’s requirements. Therefore, tracts that experienced population decline were merged together (Figure 9c), and tracts that experienced population growth were divided into two or more new tracts (Figure 9d). Population change, however, was not the only reason for transfers (Figure 9b). Changes in infrastructure, especially building the interstate highways, yielded transfers. Tract boundaries shifted to interstate highways, taking away part of the adjoining tract. Also, transfers occurred because of local decision-makers, where local census tract committees were responsible for delineating tract boundaries. The committee could change boundaries as they saw fit, often producing transfers. For example, the city of Minneapolis, Minnesota, redrew its tracts for Census 2000 to more closely match the cities’ neighborhoods (City Planning Department 2002). The Census 2000 tract boundaries differed greatly from the Census 1990 boundaries as a result of the decision by local officials.

Figure 8. Two sets of tract boundaries. The lighter boundaries are from the 2000 census while the darker set depicts 1990.

The Interpolation Process
Areal interpolation has been defined as “the process of taking statistical information for one zonation [source units] of an area and converting it to give an estimate of that statistic for another incompatible zonation [target units] of the same area” (Fisher and Langford 1995, 211). Why is this needed? A massive amount of data, especially social science data, is aggregated to areal units. Often, the entities collecting the data define the boundaries of the areal units independently from one another, which results in incompatible units (Goodchild and Lam 1980; Flowerdew and Green 1989; Goodchild et al. 1993; Xie 1995). Comparing data collected over different sets of areal units is difficult because we cannot separate actual differences from the differences introduced by the incongruent sets of boundaries (Gregory 2002). For the case of historical census tract data, the type of areal unit (census tract) is constant through time, but the boundaries of the tracts change through time. Consider the following hypothetical example (Figure 9). We want to know how much the total population changed from 1990 to 2000 in the tracts defined for 2000. Because of the boundary change, a direct comparison based on 2000 tracts is impossible. The 2000 tracts could be aggregated together to equal the 1990 tract, but such aggregation removes the geographical detail offered by the 2000 tracts (Openshaw 1984; Howenstein 1993; Openshaw and Rao 1995; Gregory 2002). Instead, areal interpolation can be used to determine total population change--areal interpolation provides a solution to the problem of incompatible areal units.

![Diagram](image)

(a) **Name change**: changes the number of a geographical unit without a boundary change

(b) **Transfer**: moves part of one geographical unit to another unit

(c) **Merger**: combines two or more geographical units into one unit

(d) **Division**: splits one geographical unit into two or more units

**Figure 9. Types of boundary changes (adapted from Gregory 2002).**

There are several areal interpolation methods for dealing with such nonidentical boundary files. The simplest method is area-weighted interpolation, where estimates may be made by computing area-weighted estimates from one area for another area (Figure 10). Using the example above, 1/3 of the 1990 tract 101 area is assigned to 102 in 2000 and 2/3 of the area was assigned to 103. To interpolate the 1990 population to the 2000 census tracts, multiply the population by the fractions of 1990 tract 101 assigned to 2000 tracts 102 and 103. Thirty-three people in tract 101 will be assigned to tract 102 with the remainder assigned to tract 103. This method is simple, but it makes a questionable assumption – it assumes that the population within a census tract is homogeneously distributed across the entire tract. Unfortunately, this assumption is usually wrong. The NHGIS project will experiment with other interpolation methods to determine if they provide more reasonable estimates.

**CONCLUSION**

The National Historic Geographic Information System (NHGIS) is a five-year NSF-funded project to build a comprehensive census database, for both boundary files and attribute data, for the entire United States at both the county and census tract level. Using both existing digital data, and scanned census maps, a temporal database is being built. At the same time, all available statistical data are being digitized so that users will be able to pursue true spatio-temporal analyses. This database will enable researchers to pursue many types of geodemographic analysis not
possible before. Two research activities for the project include areal interpolation and cartographic generalization. We are now developing and testing a multitude of different interpolation algorithms that will allow researchers to make direct statistical comparisons across decades. For instance, a user might wish to calculate the percent change in poverty from 1960 to 1990 for Los Angeles. Due to the lack of census boundary correspondence between decades, as illustrated previously, such analysis would not be possible without robust interpolation capability. Research on the project is also exploring the creation of multiple scale databases at 1:150,000, 1:400,000, and 1:1,000,000 through the application of generalization algorithms. All boundary and statistical data will be delivered through a Web interface, and available in the summer of 2006.

![Figure 10. Hypothetical boundary change requiring areal interpolation.](image)

<table>
<thead>
<tr>
<th>1990 Tract</th>
<th>Area</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>30</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2000 Tract</th>
<th>Area</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>102</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>103</td>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>

**REFERENCES**


