Realigning Historical Census Tract and County Boundaries

David Van Riper
Research Fellow
Minnesota Population Center
University of Minnesota – Twin Cities
dvanriper@gmail.com

Stanley Dallal
ESEA
dallal@esea.com

Abstract

The National Historical Geographic Information System (NHGIS), housed at the University of Minnesota's Minnesota Population Center (MPC), is a National Science Foundation-funded project to digitize and disseminate historical United States census data. We have created shapefiles that depict historic census tracts and county boundaries for the entire country for all censuses completed from 1790 to 2000. Due to improved mapping technology, the boundaries for these censuses tend to be out of alignment with each other. We are realigning these boundaries to the most accurate 2010 Census to benefit researchers that analyze historical spatio-temporal population trends and patterns. The MPC started the conflation processing in July 2008, and the project is scheduled to finish in 2011. ESRI's ArcGIS software suite combined with ESEA's MapMerger ArcMap extension have provided the tools for aligning the historic census boundaries to the most recently updated census features.

Introduction

The National Historical Geographic Information System (NHGIS), housed at the University of Minnesota's Minnesota Population Center (MPC), is a National Science Foundation-funded project to digitize and disseminate historical United States census data. There are two main parts of this project. First, the MPC is collecting as much of the historical census data as possible and is creating a website to disseminate the information. Second, the MPC is creating shapefiles that depict historic census tracts and county boundaries for the entire country. Scholars may combine the attribute data with the boundaries and analyze spatial and temporal trends in migration, ethnicity, health, income and other topics at multiple scales (e.g., national, state, county, neighborhood) over 200 years. More information on NHGIS can be found at http://www.nhgis.org .

The historical census tracts and county boundaries that are a product of NHGIS were created from the Census Bureau's TIGER/Line Census 2000 files. Since those files were released, the Census Bureau started the Master Address File/Topologically Integrated Geographic Encoding and Referencing Accuracy Improvement Project (MAF/TIGER AIP). This project improved the accuracy of spatial features (roads, water, etc.) used to create census tracts and other census units (United States Census Bureau 2001b). Unfortunately, this accuracy improvement has serious implications for the NHGIS boundary data (Figure 1). NHGIS leveraged features in the 2000 TIGER/Line files to construct historic census tract and county boundaries (United States Census

Figure 1. The 2000 census tracts derived from the 2008 TIGER/Line files (in red) differ from the NHGIS 2000 census tracts, derived from the 2000 TIGER/Line files (yellow).

Bureau 2001a). We are reusing boundaries if they remain constant between decades and are using existing TIGER/Line features—such as rivers and roads—when they form historic boundaries. If these features move because of the improved accuracy, the historic boundaries will not overlap with entities from the 2010 census or subsequent censuses. The realigned NHGIS boundaries will help people extend their studies of spatio-temporal trends and patterns to future decennial censuses and future releases of yearly American Community Survey data. A scholar who has analyzed changes in neighborhood composition between 1990 and 2000 can extend that analysis to 2010 because the historic boundaries will be realigned with 2010 boundaries.

In light of these more accurate TIGER/Line features, the MPC developed a plan to align the historic census tract and county boundaries to the new features. The historic boundaries were originally created using ESRI's ArcInfo 7.0 and ArcGIS 8.1 software packages, and the MPC wanted to continue using ESRI products during conflation. ESEA's MapMerger ArcMap extension (www.esea.com) was chosen as the software to assist with the conflation process. MapMerger allows us to perform the alignment accurately and faster than would otherwise be possible, while working within a familiar ArcGIS environment. We worked with geodatabases at the county and state level. The project uses Dell OptiPlex 755 computers with 4 gigabytes of RAM, 500 gigabyte hard drives, and dual monitors. The scope of the project is as follows:

Census Tracts
For 1990 and 2000, the entire country is subdivided into census tracts. There were approximately 64,866 census tracts in 2000 and 60,513 census tracts in 1990.

For the 1910-1980 censuses (tracts were first used in 1910), only counties that participated in Metropolitan Statistical Areas were tracted. As we go back in time, we have fewer and fewer census tracts to align.

Counties
The Census Bureau has tabulated data for counties since the first census in 1790. Thus, we will align county boundaries for all censuses from 1790 to 2000. Luckily, there are many fewer counties than census tracts! There were approximately 3,100 counties in 2000.

Process

MPC staff spent three months designing the conflation process. They identified the data requirements, the data pre-processing requirements, the conflation workflow using MapMerger, and the data post-processing requirements. After designing the process, a pilot state was completed, and some modifications to the process were made. Finally, additional staff was added to execute the conflation process.

Data Requirements
The source data for the conflation process are the original NHGIS census units, created from the TIGER/Line Census 2000 files. These data are stored as SDE feature classes. The target data for the conflation process are the 2008 TIGER/Line shapefiles, which were the first release of the TIGER/Line files after all counties had completed the MAF/TIGER AIP (United States Census Bureau 2008).

Pre-Processing
The original NHGIS census tract polygon feature classes for 1910 – 2000 are split apart into county-level feature classes. For a given county, all the tract feature classes are unioned, and the unioned output is converted to a line feature class. A series of attributes (*t2000*, *t1990*…*t1910*) are added to this line feature class. Then, the 2000 census tract polygons are used to select features in the line feature class that served as tract boundaries in the 2000 census. For the selected features, the value of the *t2000* attribute is set to 1. This process is then repeated for the 1910-1990 census tracts. Next, the line feature class is broken into a series of feature classes based on the values of the *t2000*…*t1910* attributes. If a county have census tracts for 2000-1980, the line feature class will be split into three separate feature class. One will contain the lines with *t2000* = 1, one will contain the lines with *t1990 = 1 AND t2000* = 0 (these lines represent boundary changes between 1990 and 2000), and one will contain the lines with *t1980 = 1 AND t1990 = 0 and t2000 = 0* (these lines represent boundary changes between 1980 and 1990) (Figure 2). These county-decade feature classes are then merged together to create a state-decade feature classes.

Figure 2. Pre-processing the NHGIS data yields a feature class for each decade, with each feature class representing the boundary changes between decades.

As part of the 2008 TIGER/Line release, the Census Bureau provides the 2000 census tract shapefile and an "All Lines" [1] shapefile for each county in the United States. MPC staff downloads these shapefiles for all counties in a given state. The shapefiles are imported into a file geodatabase and re-projected to the standard NHGIS projection.[2] A series of attributes (*t2000*, *t1990*…*t1910*) are added to the All Lines feature class. The 2000 census tract feature class is used to select the features from the All Lines feature class that serve as 2000 census tract boundaries. The selected lines are saved out to a new feature class, and after processing all counties in a given state, the new feature classes are merged together to create a statewide feature class.

A different process is used to prepare the 2008 TIGER/Line data for the 1910 – 1990 census tracts. The county-decade feature classes described above are buffered by 1 km and then the county All Lines feature classes are clipped using the buffer (Figure 3). These clipped county-decade All Lines feature classes are merged together to create state-decade All Lines feature classes.

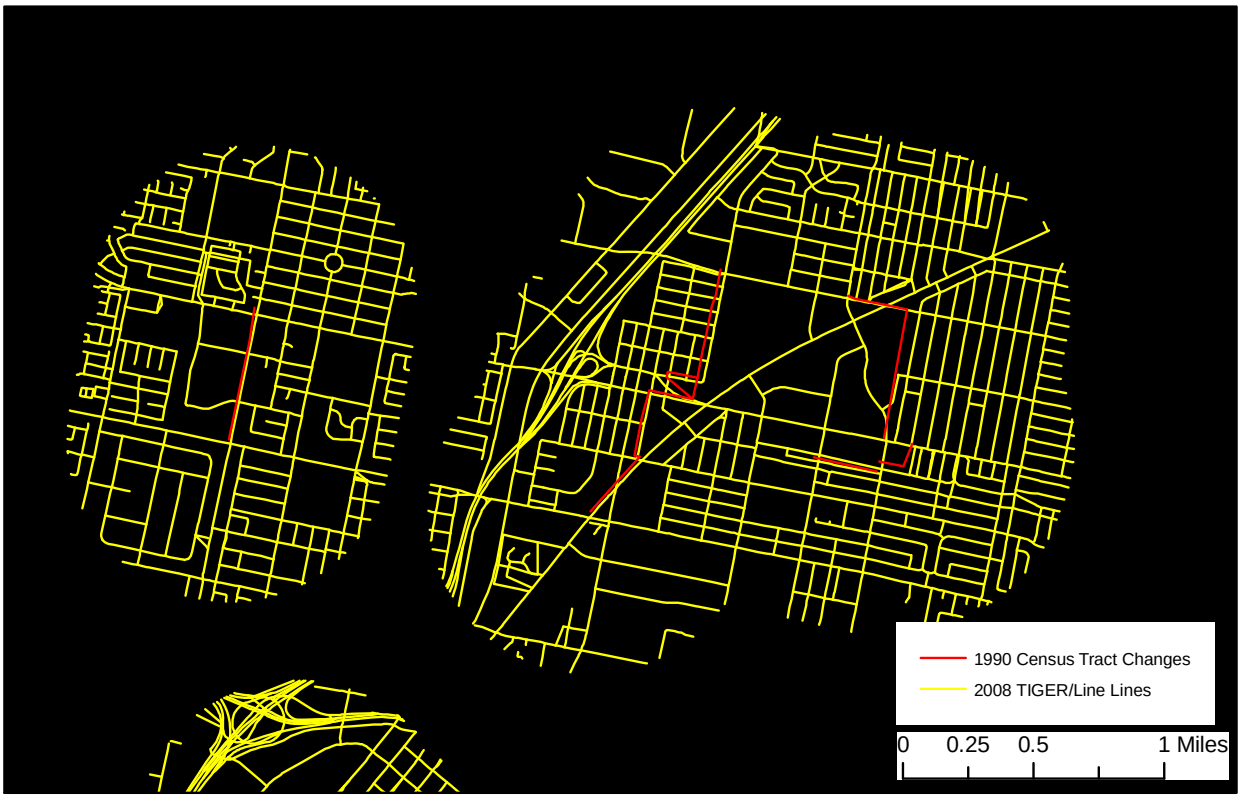

Figure 3. The feature class containing the 1990 tract changes was buffered, and the buffered features were used to clip features from the 2008 TIGER/Line All Lines feature class. The clipped feature class served as the target layer during the realignment process.

[1] The "All Lines" shapefile contains all features (*e.g.*, roads, hydrography, railroads, municipal boundaries) in a county.

[2] The NHGIS uses three different Albers Equal Area projections for its datasets. One projection is optimized for the contiguous United States. Another is optimized for Alaska, and the last is optimized for Hawaii.

Conflation Workflow

MPC staff defined a standard workflow for the conflation project that all employees involved in conflation followed. Staff uses the MapMerger ArcMap extension to align the NHGIS census tract boundaries for the 2000 Census to the Census 2000 tract boundaries derived from the 2008 TIGER/Line shapefiles. After finishing the alignment process in ArcMap, staff uses MapMerger to export a line feature class to a personal geodatabase. A custom Python script is executed in ArcCatalog to convert the line feature class to a polygon feature class, and the polygon count in this feature class is compared to the expected polygon count from the original NHGIS database. The script reports the results of the comparison to the staff member, and if the counts are mismatched, the staff member returns to ArcMap and fixes the errors using the MapMerger tools. This process continues until the counts match for the 2000 census.

Using MapMerger, staff members then align the lines that represented boundary changes between the 1990 and 2000 census tracts. This process is different from Census 2000 boundary conflation process. The 1990 boundary lines from the NHGIS and the buffered lines from the 2008 TIGER/Line are loaded into MapMerger. The final, conflated Census 2000 tract lines are also added to the ArcMap document. A Select by Attributes… query is applied to these conflated Census 2000 lines to select those that also serve as 1990 tract boundaries (Figure 4). The NHGIS 1990 boundary lines are manually matched and/or spatially adjusted to the 2008 TIGER/Line features. The aligned features are exported to a feature class, and the same custom Python script used for the aligned Census 2000 tract boundaries is executed. The polygons counts for the 1990 NHGIS tracts are compared to the polygon counts for the aligned 1990 tracts, and if there is a mismatch, the staff member fix the errors in ArcMap using the MapMerger tools. This process continues until the counts match for the 1990 census. After finishing the 1990 alignment, the same process is repeated for the 1980 tracts and continues back in time until all decades for a given state are realigned.

Figure 4. The red features represent the tract boundary changes between 2000 and 1990. The yellow lines represent the 2008 TIGER/Lines that will be used to realign the 1990 boundary changes. The blue lines are the 2000 tract boundaries that also serve as 1990 tract boundaries. Staff used the MapMerger tools to realign the red lines to the correct yellow lines.

Post-Processing

After the alignment process is finished for a state, a final post-processing procedure is executed. For each decade a state has census tracts, tabular census data downloaded from the NHGIS website is joined to the aligned census tracts using its unique ID. If an aligned tract does not have a corresponding entry in the tabular data, it is selected and saved out to a feature class. If an entry in the tabular data does not have a corresponding census tract polygon, it is selected and saved out to a table. MPC staff checks the entries in the feature class and table and make any corrections to the polygon attributes or to the feature classes derived from MapMerger.

Status

This project started its production phase in February 2009 and has finished 47 states as of May 2010. These 47 states were finished with the 2 full-time equivalent employees.[3]

Conclusion

Numerous benefits will accrue from realigning the NHGIS county and tract boundaries with the new TIGER/Line features. Accurate spatial datasets are becoming increasingly important as research begins to exploit new technologies. For example, without the proposed realignment,

[3] One graduate assistant and three undergraduate assistants were employed on this project. They each worked approximately 20 hours per week.

highly precise spatial data collected from handheld GPS units or satellite imagery may be incompatible with existing NHGIS boundaries. The realigned boundaries will help users integrate new data with the historic census tracts and counties.

References

United States Census Bureau. 2001a. Census 2000 TIGER/Line Files. Washington, DC: Internet. http://www.census.gov/geo/www/tiger/tiger2k/tgr2000.html [Accessed 3 June 2010].

United States Census Bureau. 2001b. MAF/TIGER Modernization. Washington, DC: Internet. http://www.census.gov/geo/mod/rel_docs.html [Accessed 3 June 2010].

United States Census Bureau. 2008. 2008 TIGER/Line Shapefiles. Washington, DC: Internet. http://www.census.gov/geo/www/tiger/tgrshp2008/tgrshp2008.html [Accessed 3 June 2010].