Pulling It All Together: Developing the Spatiotemporal Layers to Support Location-based Integration

Wendy Thomas and Tracy Kugler
Minnesota Population Center, Minneapolis, MN, USA
Corresponding author: Wendy Thomas, email: wlt@umn.edu

Abstract

The promise of open data and statistics for sharing and integrating data from multiple sources is great. It is especially hopeful for combining data from different disciplines to explore the interaction of human activity and the environment. However, without adequate infrastructure to support a clear linkage along spatial and temporal dimensions, the researcher is left on their own to develop these all important relationships. The difficulties are great, for example, data without clear spatial boundaries, statistics representing dissimilar points in time or averages of multiple points in time, inconsistent data availability over time, or mismatched spatial boundaries between data sources. As the research focuses in on smaller areas, urban, sub-urban, and regional statistics, the problems increase. The Minnesota Population Center (MPC) at the University of Minnesota is dedicated to addressing these issues. Known primarily for its large integrated collection of world-wide census microdata, IPUMS, the MPC has also made available historical collections of U.S. Census and other aggregate statistics along with the spatial boundary files associated with those geographies in the National Historical Geographic Information System (NHGIS). The NHGIS is currently increasing its usability by creating time series for common statistical tables. A new project funded by the U.S. National Science Foundation, Terra Populus, brings together population, environmental, land use, climate, and areal data through location-based integration. The project will provide an organization and technical frame work to preserve, integrate, disseminate, and analyze global-scale spatiotemporal data describing population and the environment.

Key Words: data integration, linked data, spatial boundaries, urban and regional statistics

1. Introduction

The promise of open data is great. More and more data (micro, aggregate, and spatial) are accessible to the public for discovery and analysis. However, without adequate infrastructure to support a clear linkage along spatial and temporal dimensions, the researcher is left on their own to develop these all important relationships. The difficulties are great, for example, data without clear spatial boundaries, statistics representing dissimilar points in time or averages of multiple points in time, inconsistent data availability over time, or mismatched spatial boundaries between data sources. As research focuses in on smaller areas, urban, sub-urban, and regional statistics, the problems increase.

Funding agencies are supporting a number of programs intended to enhance the ability of researchers to find, link, and efficiently analyze data at a variety of geographic levels. Two of these projects, the National Historical Geographic Information System (NHGIS) and Terra Populus (TerraPop), are located at the Minnesota Population Center and funded by the United States National Science Foundation. These projects provide users with a consistent spatial basis for comparison over time. NHGIS provides historical population data and boundary (shape) files for all U.S. census data geographies back to 1970, Census Tracts back to 1910, and State and County data back to 1790. This work has been supplemented by the Integrated Spatio-Temporal Aggregate Data Series (ISTADS) project, funded by the Eunice Kennedy Shriver National Institute of Child Health & Human Development at the National Institutes of Health. ISTADS defines time series of

topical population data through integration of the metadata, identifying where the same data has been produced (or can be generated through aggregation) over time and creating time series tables for the required geography on demand. The TerraPop project provides broader spatial and topic coverage, including population and environmental data for the world.

These projects focus on addressing the spatiotemporal issues which arise when integrating data sources. In addition, they employ the lessons learned from the Integrated Public Use Microdata Series (IPUMS) on topical data harmonization to integrate and harmonize data on a topical level. TerraPop will use both the semantic integration approach of ISTADS as well and the data fusion approach of IPUMS.

2. Problem Statement

Due to the lower costs of capturing and storing data, increasing use of metadata and storage standards, and the open-data movement, data analysts are enjoying a level of access never before achieved. At the same time they are facing issues of harmonization and integration that are difficult and expensive to do.

Clear valid data linkage requires commonality on two of three dimensions: spatial, temporal, and topical. To compare or contrast two spatial areas requires data for the same time periods on comparable topics expressed in similar ways. To look at change in an area over time, there needs to be consistency in the spatial area and topical comparability. To link and analyze data from different sources and topical coverage, the temporal and spatial coverage need to be comparable.

For small areas, linking, comparison, and analysis becomes more difficult due to the nature of the data. Spatial data is primarily expressed as polygons representing administrative areas or grids. The temporal dimension is often more widely spaced or is the aggregation of data collection points over a long period of time (i.e. U.S. American Communities Survey 5 year file is an aggregation of 60 data collection events). Topical data is represented within confidentiality constraints which include the use of cohort groups, top-coding, collapse of classification schemes, and data suppression. In addition, aggregate data is often published with simply an area name, providing little or no information as to the actual footprint of the geographic area or the geographic time stamp. (A geographic time stamp is the valid date of the footprint as opposed to the data. It is common to republish data from an earlier date for a new geography, i.e. 2000 Census Data published for 2003 Congressional Districts.)

3. National Historical Geographic Information System (NHGIS)

As is clear from its name, NHGIS focuses primarily on providing a solid geographic (spatial) layer for analysis of data over time. NHGIS contains aggregate data, data from statistical tables, retaining the relationships between table cells within its metadata using the Data Documentation Initiative DDI standard for describing aggregate data. The data is primarily from the United States Census of Population and Housing (1790 to date) and the American Communities Survey, supplemented by other major state and county level series such as County Business Patterns (Bureau of Labor Statistics), and other censuses on economic and agricultural activities. The project created boundary files for U.S. States, Territories and Counties from 1790 as well as Census Tracts from 1910 on. NHGIS continues to integrate changes to all census supported geographies as they are published. Extensive work was done to link published data, historically described by a

name, to an integrated geographic structure and location coding system which is used to address individual polygons within the shape file.

This work addressed one of the major difficulties in linking historical data over time. Geographic comparability is multi-faceted. Geographic locations as we know them consist of a Name which may be expressed as a different code in multiple systems, and a geographic footprint. Change may result from the change in a geographic footprint (i.e. the separation of the State of West Virginia from the State of Virginia), a change in a code system, or the change of a Name. Unfortunately the change in one of these does not indicate a change in the others. Working with a system based on Name the following common situations could easily be missed: the change of the geographic footprint of the State of Virginia when West Virginia was created; the switch in Names between Rock County, MN and Pipestone County, MN between two censuses; the dissolution of the original Lac qui Parle County, MN shortly after one census and the adoption of the name Lac qui Parle County by another Minnesota County prior to the following census; and the renaming of a County with no area changes.

Working back from the 2000 TIGER/Line Files the NHGIS project created boundary files going back to 1790. The work also made use of the 1992 TIGER/Line data for 1980 census tracts, printed census maps from the 1910-1980 census tracts and small areas, and the *Map Guide to the U.S. Federal Censuses, 1790-1920,* by William Thorndale and William Dollarhide (Genealogical Publishing Co., Baltimore, MD, 1987). Geographic structure and coding systems were extended to allow for the inclusion of historical areas and this was used to link Names found in the data to the appropriate geographic footprint. Care was taken to make sure that the footprint matched the geographic date (rather than the collection date) of the data.

Time was not a major issue as the primary data sets were decennial. However, the topical dimension presented problems as both coverage and detailed changed over time. NHGIS retained the dimensional relationships within the aggregate tables though the use of the DDI metadata structure called an NCube. This structure describes each dimension of a table so that, if there is no suppression, they can be collapsed to calculate single dimensional totals. For example a table of Sex (male and female) by Age (cohorts) could be collapsed to provide totals for each age cohort. The DDI structure also allows easy identification of tables that share a common dimension (i.e. all tables using the same set of age cohorts).

ISTADS takes the metadata within the NHGIS system identifies consistency in topical dimensions over time and between sources and then creates harmonized descriptions where possible to extend the topical comparability of data over time and space. Unlike the IPUMS systems where the results of harmonization are captured through “data fusion”, the creation of new data items containing the harmonized value, ISTADS captures the harmonization in the metadata noting where comparable data is located and any calculations that are needed to create a time-series table on demand. (NHGIS, /documentation/time-series)

NHGIS has also added the School Attendance Boundary Information System (SABINS) which includes both boundary files and aggregations of data from Census Blocks for school attendance areas, or school catchment areas, in selected areas of the United States. Maintaining consistent boundary files which reflect change over time as well as the small area data associated with them, supports the aggregation and/or interpolation of data into new geographies as they arise.

4. Terra Populus

The goal of TerraPop is to "provide an organizational and technical framework to preserve, integrate, disseminate, and analyze global-scale spatiotemporal data describing population and the environment." (Overview, 2012) TerraPop is a collaboration of the Minnesota Population Center (MPC), University of Minnesota Institute on the Environment, University of Minnesota Libraries, CIESIN at Columbia University, and the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan, funded by the United States National Science Foundation (NSF). TerraPop expands both the geographic and topical coverage of NHGIS. Its primary task is to "collect, preserve, integrate and describe datasets that measure changes in the world' population and environment over the past two centuries." (TerraPop, 2012) TerraPop will include economic data, population data on individuals, families, and households, as well as environmental data on land use, land cover, climate, and other topics. Data will be expressed as areal, raster, or microdata depending upon the source.

TerraPop will also contain vector-based geographic information (GIS) data delineating administrative and census unit boundaries. The data contained in this system will be geographically represented as grids or polygons. The TerraPop system will provide researchers the means to reconcile different spatial scales and multiple time scales. Semantic integration of topical and classificatory information will be based on in-depth analysis of the metadata and the development of harmonized coding and classifications which allow for various levels of comparison detail dependent upon the research question and limitations of the geographic and temporal selection.

The data within TerraPop represents microdata, aggregate data and raster data. Analysts will be able to obtain results in each of these formats. For example, census microdata may have land use, land cover or climate characteristics attached. Microdata from population and environmental sources can be aggregated to administrative areas or attached to gridded locations. Tools and procedures will be created to integrate, disseminate and analyze data within the system.

For small area and regional researchers, TerraPop fulfills the role of a "Rosetta Stone" for linking data from multiple sources accurately through a spatiotemporal crosswalk. By creating a consistently coded set of boundary files for municipalities, counties, territories, and states and providing clear links from data sources to these polygons over time, TerraPop clarifies two often difficult to determine dimensions for comparison and analysis. An initial public version covering Brazil (1960-2000) and Malawi (1998 and 2008) population, land cover, agricultural land use and climate data, will be available by the end of 2013. While still early in its development, TerraPop holds the promise of becoming an invaluable source of data and metadata for future urban, regional and small area research.

5. Conclusions

In order to do linking and comparison the ideal situation is to be able to hold two of the three dimensions of spatial, temporal, and topical constant so that analysis can be done on the third. This means that the metadata must make the spatial and temporal dimensions of the data set very clear, linking to standard descriptions of spatial areas wherever possible. This is increasingly important when one works with small areas as most data at this level is aggregate rather than microdata. The limits on the comparability of aggregate data increase the importance of being able to exactly specify the spatiotemporal dimensions.

Making data "open" deals with only part of the problem. Funding support for systems which address the major spatiotemporal and topical harmonization issues expands the ability of researchers to look at change over time and comparison between locations. It opens the possibility of new research areas that look at the interaction of humans with each other and their environment.

The work at the Minnesota Population Center also highlights the importance of standards for metadata and for the use of public available structures that can be referenced to supply consistent and common definitions for spatiotemporal dimensions of data. Publication of shared resources provides a common set of geographic structures, as well as geographic name and boundary information allow researchers to reference a common source, helping to ensure a shared understanding of the spatial dimension of published data over time.

References

Minnesota Population Center (2010), National *Historical Geographic Information System (NHGIS)*, www.nhgis.org

Minnesota Population Center (2012), *Terra Populus, Integrated Data on Population and Environment (TerraPop)*. www.terrapop.org

Minnesota Population Center (2012), *Terra Populus Overview*, www.terrapop.org/presentations