UC Santa Barbara
GIS Core Curriculum for Technical Programs (1997-1999)

Title
Unit 2: Demographic Data

Permalink
https://escholarship.org/uc/item/6fq2q9pq

Authors
2, CCTP
Matthews, Kevin

Publication Date
1998

Peer reviewed

UNIT 2: DEMOGRAPHIC DATA

Written by Kevin Matthews, George Mason University

Context

Demography is the study of human populations with an emphasis on statistical analysis. Data describing a human population is referred to as demographic data. This unit is concerned with the task of locating demographic data, including data describing the socioeconomic characteristics of the population. In all, information about the age, ethnic, gender, income, housing condition, and other socioeconomic variables will be considered. Besides academic researchers such as Geographers, Sociologists, and Political Scientists, professionals that use demographic data include, but are not limited to policy analysts and makers, urban planners, market analysts, and transportation planners. Virtually any organization that utilizes information about people need demographic data.

The scope of this unit is to provide guidelines and suggestions to identify demographic datasets and evaluate whether the dataset can be used within a GIS effectively. Assuming that the cartographic database representing the geographical observatioins (can be geometrically symbolized by points, lines, or polygons) is in place in the GIS, besides demographic and socioeconomic variables, the demographic database must include either explicitly the locations (e.g. x-y coordinates) of the geographical objects, or geographic identifiers through which tabular demographic data can be linked to the geographical objects in order to use this database in a GIS.

Example Application

An ongoing research effort led by an advocacy group for the lower income's access to transit sought to determine how bus and rail transit fares were distributed in the Washington DC metropolitan area. Necessary data included cost of fares to and from all locations, bus stops and rail stops. For this example, it is assumed that this data was already compiled. In order to evaluate the level of access to transit for the lower income group, however, demographic data were needed.

Traditionally, census tract level is used in many socioeconomic analysis. Therefore, census tract level data of population and housing for the Washington DC Metropolitan Statistical Area (MSA) were needed. In addition to the transit data described above, the

relevant census variables required to evaluate the accessibility of lower income people to transit include the number of automobiles owned per household, median household income, and percent of population below poverty.

A research librarian was consulted in order to expedite the search for this data. The first option discovered was the socioeconomic and demographic data collected by the United States Census Bureau. Tract level data for all variables listed above were found in the summary tape files (STFs). The data was available in digital form at the level of geography needed (i.e., tract level data). When linked to the TIGER files, the cartographic database for census mapping, the selected census variables can be mapped. There was one problem with using this dataset. That is, it was collected in 1990 and was therefore deemed inappropriate for use in this study due to its being out of date.

The next option led to research for more current data. The local bus and rail transit authority for Washington, DC was then consulted. This organization, however, only had data for administration and operations. They had no demographic data and limited amounts of spatial data. The researchers were directed to the Federal Transit Administration's website which contained bus and rail transit data including level of service. The transit authority also suggested contacting the local transportation planning organization for the Washington Metropolitan Area.

The Metropolitan Washington Council of Governments (MWCOG) indicated that updated data did exist and were available. The researchers, however, would have to purchase the data. They were also suggested contacting the Bureau of Transportation Statistics who distributes the Census Transportation Planning Package (CTPP). The CTPP contains various levels of geography including transportation analysis zones and associated demographic datasets for 1990. The problem with using the CTPP was that the updated demographic data to be purchased from MWCOG were aggregated into transportation analysis zones (TAZ) rather than census tracts.

The purpose of this example is to show the process one might go through in order to locate secondary demographic data sources that can be used in a GIS. Specifying the data needs in the beginning of the project will save time and money. Data needs specification involves choosing appropriate variables, defining the geographic extent of the study area and determining the level of geography to be used.

After the data needs are specified, the effort to locate demographic data should begin by identifying secondary data sources. This identification process can be expedited by consulting an academic librarian or contacting a government agency that specializes in the required data.

Learning Outcomes

The following list describes the expected skills which students should master for each level of training, i.e. Awareness/Competency/Mastery.

Awareness:

Identification of Data Needs. The expected learning goals of this section asks the student to specify their data needs before starting any project. Evaluating whether demographic data can be mapped.

Competency:

Search strategies for locating demographic data. The learning goal for this section deals with the general guidelines for locating demographic data for use in a GIS.

Mastery:

Secondary data sources. The section deals with the types of demographic data available.

Complementary Units:

- Unit 1 - Data Acquistion
- Unit 20 - Using Text Editors
- Unit 21 - Using Spreadsheets
- Unit 22 - Merging tabular data with spatial data

Awareness

Identifying data needs:

Before one starts searching for demographic data, the data required for a study have to be specified as clearly as possible. A possible approach to specify data needs is to develop a list of data that will best answer a specific question or a broad set of questions being asked. Answers to three broad set of questions should be kept in mid when trying to locate demographic data.

1. What variables (i.e., population characteristics) are needed?
2. What are the required characteristics of these variables?
3. Can the data items be mapped in a meaningful manner?
4. What is the study area or geographic extent of the data?
5. What is the level of geography within the study area?

What variables are needed?

Based upon the question(s) asked in a study, the analyst should identify a variable or a set of variables likely to be found in demographic databases that can provide information relevant to the study or address specific questions. By analyzing data describing these variables, answers may be derived to answer some research questions. However, one should be prepared to identify a set of "proxy" variables to approximate the ideal variables when the ideal variables are not available. Ideal variables do not always exist in real world databases.

What are the characteristics of variables?

Besides identifying variables, the analyst should also specific the characteristics of those variables. These characteristics include, but not limited to the measurement scale (nominal, ordinal, interval, ratio) of the variables, the temporal aspect of the variables, and the reliability or sampling framework of the data. All these characteristics should be specified as criteria to be used for locating databases.

This unit is primarily concerned with locating secondary data.Secondary data are data which have been previously collected by some other researcher or organization. For a good reference on collecting your own geographic data for use in a GIS, see Haring, et.al. It is in an organization's best interests to know what data they need, what data is available and to be able to thoroughly research to determine whether their data needs can be met with existing data.

Can the data be mapped?

When demographic data are reported at the disaggregated (individual level or microdata) level, locations of individuals have to be included in the database in order to link the demographic data to locations on the map. The location information most likely be in longititude-latitutde coordinate system, or street addresses. Using these location information, the observations will be put in the map for analysis and display.

Demographic data at an aggregate level can be mapped if it satisfies the two following criteria: if the data is reported within an areal unit and if a geographic identifier exists for that areal unit. The database must contain information that ties the attributes of an areal unit to a location in space. This piece of information is known as a geographic identifier, or a geocode. It is this piece of information that will be used to join the demographic information (i.e., attribute/non-spatial data) to the spatial database (i.e., the basemap). For example, a geographic identifier can simply be a state name or state abbreviation that has been associated with the demographic information for that state. If that same name or abbreviation is found in the basemap's database, then the demographic data can linked to the map and be mapped at the state level.

The federal standard for geographic indentifiers is a FIPS (Federal Information Processing Standards) code. A geographic FIPS code is recorded as a variable and is associated with each record in the demographic attribute database and spatial database. Each record in the demographic database will have a unique FIPS code corresponding to a unique areal unit in the spatial database.

The FIPS code system for geographic information follows the Census hierarchy. For example, the FIPS code for the tract where George Mason University is located is 510594405.00. This particular geographic identifier gives information about the state, county, and tract number where George Mason University is located. Table 1 shows how the information is broken down.

Table 1:
FIPS Code 510594405.00

	Name	FIPS Code
State	Virginia	51
County	Fairfax	059
Tract		4405.00

note: the associated spatial data for all level of Census geography is found in the TIGER line files maintained and distributed by the United States Census Bureau.

Competency

 The process of locating appropriate demographic datasets is especially efficient when the data needs of the project have been thoroughly specified. Research for a demographic dataset can be accomplished in an university library or via the internet. Many university libraries are government depository libraries - they receive certain standard datasets and publications from the federal government and local governments. Most databases generated by federal government and local governments recently are already in digital formats, though paper copies may also be available. The digital form is always preferable for subsequent data manipulation and analysis in GIS. If data are in hardcopy, one should exhaust all search strategies for digital versions before manually entering the data. Should no digital versions of the data exist, use of a scanner and text recognition software is an excellent alternative to manual data entry. For an overview of formatting a raw text file into a structure a GIS can use see Unit 20 - Using Text Editors and Unit 21 - Using Spreadsheets.

 Today, many federal government agencies, local governments, and private companies disseminate data through internet. Therefore, the internet is also an efficient and effective tool to identify demographic data. See Unit 1 - Data Acquisition. Strategies for internet research include:

1. using keywords in an internet search engine such as Lycos, Alta Vista, Excite, Yahoo, or Infoseek,
2. asking questions on GIS related discussion groups and listservs concerning the existence of demographic data,
3. looking for demographic data within state and federal agency's website. The US Census Bureau and the FedStats sites are excellent starting points, and
4. searching the web for sites that catalog demographic data sources for use in GIS, (example Geographic Information System (GIS) WWW Resource List).

In addition, one may:

- - pose questions to universities which might maintain local/regional demographic and spatial data,
- -consult local and regional planning agencies that use GIS. They are excellent contacts who may lead to further data sources or people who might be able to help with data

needs,

- - conduct literature reviews in academic and professional journals for related subjects in order to ascertain where the authors obtained their data, and
- - realize that data might need to be purchased from private companies who specialize in collecting and/or selling demographic data.

Mastery

Mastery of locating demographic data comes with knowing what demographic datasets compatible with use in a GIS are available. The United States Census Bureau provides an excellent resource for locating state, federal and international demographic data. Most of these databases can support mapping in GIS, though the data may not be available on-line, but have to be purchased.

Among databases disseminated by the U.S. Census Bureau, the most popular databases are probably derived from the census conducted every ten years, such as the Census of Population and Housing. Theoretically, census data are derived from enumerating everyone in the nation. However, not everyone in the nation provides equally detailed information for census. Only approximately 5% of the entire U.S. population were asked to provide detailed socioeconomic and demographic data.

Most census data are aggregated into various levels of enumeration units(see Census Geography). These areal units are reported as individual observations that describe the composition of that a unit's population.

The datasets describing population and housing characteristics can be found in the Summary Tape Files 1 through 4. The denotation is STF-1 through STF-4. STF-1 and STF-3 are generally the most appropriate source for demographic data for use in a GIS in the United States. STF-1 contains detailed geography as well as national and state totals. STF-3 is broken into file 3A, 3B, 3C, and 3D. The only difference between the files 3A, 3B, 3C, and 3D is the scale of the reporting areal unit. The supporting technical documentation for STF-3, which includes detailed descriptions of all the variables within the dataset and census heirachy definitions, can be found here.

Census data were also reported at individual level - the Public Use Microdata Samples (PUMS). The PUMS data are organized by households and household members. Data are gathered for sampled households (1% or 5% samples) and from each members of the sampled households. However, the smallest geographically identifiable area for the households is a Public Use Microdata Area (PUMA), which is usually a county or a county-subdivision.

Bibliography

Yaukey, D. 1985. Demography: The Study of Human Populations. St. Martin's Press, New York.

Haring, L.L., J.F. Lounsbury, and J.W. Frazier. 1992. Introduction to Scientific Geographic Research. Wm. C. Brown Publishers.

Resources

[outdated links have been removed]

Ciesin US Demography Home Page

UK Demography

US Census Bureau

Geographic Information System (GIS) WWW Resource List

FedStats

1997 Statistical Abstract of the United States

Created: May 14, 1997. Last updated: October 5, 1998.