

Federal Census Files

By Frank LoPresti
frank.lopresti@nyu.edu

The first Federal Census was taken in 1790 under the auspices of Secretary of State Thomas Jefferson. The predominant need for this poll of the population was, and still is, to create and manage Congressional districts. Every state gets to send two senators to Congress—no more, no less...no need to count. On the other hand, the number of persons each state sends to the House of Representatives is proportional to its population. To this end, Article I, Section 2 of the US Constitution mandates an “actual enumeration” of the population. This requires a periodic physical headcount of the country’s residents, and prohibits statistical guessing or estimation of the population. For this purpose, Congressional redistricting committees use the Redistricting File and Summary Files 1 and 2 (SF1 and SF2) from the Federal Census report.

COLLECTING THE DATA: THE LONG AND THE SHORT OF IT

Each decade, about one sixth of the “households” in the United States fill out a long Census form. This data is compiled into Summary Files 2, 3, and 4 (SF2, SF3, and SF4). The remaining five-sixths of the households are given the short

form. The basic data common to both forms yields the total counts compiled in file SF1. To reapportion Congressional seats, these counts are quickly released in the Redistricting File. No statistical sampling is used there.

The short form collects personal data (“P” variables, including household size and occupants’ relationships, sex, and age) and household data (“H” variables, including location, occupancy, and tenure of occupancy). The long form asks many valuable but invasive questions about each resident’s race, education, military service, income, health, employment, and the residence’s utilities, insurance, number of rooms, etc.

ANONYMITY: WE INSIST UPON IT EVEN IF WE CAN’T PRONOUNCE IT

When this sensitive data is collected, it goes into files where it becomes a P line if it relates to a person, or an H line if it relates to a residence. Several P lines and one H line containing the person’s address make up a household group of data in the original, “secret” database. Since this database contains peoples’ addresses, it will never be released to the public.

The issue is how to distribute useful data without compromising people’s personal information, since without *some* geographic information the Census files would be useless. To address this problem, Census workers identify households geographically using Blocks, Tracts, Zip Codes, Counties, and PUMAs (described below).

A Census Tract is usually defined to include around 1,500 households. A Block group will usually be about one half or one third of that size, or even smaller, depending on the local geography and other logical dividing lines. In NYC, for instance, a Block will sometimes be a single apartment building. If people’s addresses were removed from a Census file but Block location (within 500 households) remained available, that information would still be insecure. If someone could identify your Block, the type of house you live in, the number of rooms in your house, and whether you rent or own, they could probably figure out exactly which residence in a given Block was yours. This is potentially very bad business. Imagine, for example, the opportunities for a criminal seeking likely victims for a robbery or scam.

To avoid this, the Census aggregates data, showing only tables of information at the small Block level to protect the individual identities of you, your neighbor, and the entire Block population. Since the Block-level Census files provide only certain cross-tabulated tables—age by race, race by income, etc., they do not provide the kind of detailed information that a “bad guy” might need.

In this way, Summary Files (SF1 to SF4) protect privacy by summarizing the information into tables. The richest data, however, is the raw data from the H lines and P lines of the long form, which the Census releases as Public Use Microdata Samples (PUMS) data files. This raw data is stripped of all geographic variables except an identifier called the PUMA (Public Use Microdata Area). The PUMA groups include about 200,000

households. Since that comprises only about 1/6th of the total households in the United States, the PUMA data are weighted, which allows for statistical replication of the entire population. The data can then be used for any statistical purpose.

HOW TO FIND AND USE CENSUS FILES

Many repositories and CDs containing Census data are available. NYU is a member of the data archive at the Inter-university Consortium for Political and Social Research (ICPSR). ICPSR’s Census collection (<http://www.icpsr.umich.edu/topical.html#CENSUS2000>) is available free of charge to researchers at NYU and includes SAS and SPSS code to format these files. We recommend that you use a high-speed connection such as that provided at NYU to download these large files.

Geolytics (<http://www.censuscd.com/>) is also a very impressive resource. Used by many researchers, these value-added CDs are available for a fee. They offer a Windows-based program for downloading subsets of the various files, spanning the decades from 1970 to 2000. An application is included which allows variables (where it makes sense) to be taken across several decades using year 2000 geographical boundaries.

Help and additional information about using Federal Census data is available through the ITS Social Sciences, Statistics & Mapping Group. Contact Frank LoPresti at 1-212-998-3398 or frank.lopresti@nyu.edu for more information.

Frank LoPresti heads the Social Sciences, Statistics & Mapping Group of ITS Academic Computing Services.

The First Annual ITS Staff Art Show

In March 2004, ITS held its first annual staff art show, providing ITS employees the opportunity to share their talents with their colleagues. This informal show featured an impressive variety of work, including photography, illuminated digital art, sculpture, and painting. The exhibit was on display for the entire month, enlivening the halls of our facilities at 10 Astor Place. For a busy department focused on providing for the technical needs of the NYU community, this fun show was a welcome opportunity to learn more about the artistic talents of our colleagues, and is sure to become an annual tradition.



David Ackerman and Eduardo DeLeón appreciate a series of multimedia paintings by Callie Hirsh.



A detail of a back-lit digital print created by Philip Galanter.