Building an Archive of U.S. Census Related Data Products

by John Blodgett[*]
Urban Information Center
University of Missouri St. Louis

The consortium for International Earth Sciences Information Network (CIESIN), in conjunction with the Urban Information Center at the University of Missouri St. Louis has established a public archive of United States census data. The data are available via the Internet using FTP and/or a WWW browser. Currently the archive contains map boundary files for all common geographic units used in the census (including census blocks, block groups, tracts, counties, etc.) in a standard ascii (BNA) format. Data extracted from the 1990 decennial census Summary Tape File 3 (STF3) are provided in a format that makes it easy to link with the boundary files to create thematic maps with widely available GIS software such as Atlas*GIS, ArcView and MapInfo. These data files are organized by state and geographic level. Using the archive researchers should be able to readily retrieve files that will allow them to analyze and/or map data for areas as small as blocks or block groups. for anywhere in the U.S.. This paper discusses the content of the archive as well as describing some of the details of how and why it was constructed.

Archive Overview

The archive of census-related data products (CRDP) is a subset of the data archive maintained by SEDAC (the Socio-Economic Data Application Center) at CIESIN. It is a joint venture with the Urban Information Center at the University of Missouri St. Louis. The UIC develops programs (in SAS) for accessing the raw census files and creating data products. CIESIN then adapts these programs and creates the "assembly line" applications (using SAS and/or Unix shell scripts) that actually generate all the specific data files that are the data archive. The CRDP archive currently consists of over 11,000 files (some of which are mini-archives created with the zip software, which when unzipped may turn into several data files) and over 4 gigabytes of data (compressed — multiply by 3.3 to get uncompressed size.) It contains the following basic kinds of files:

-Mapping boundary files in ascii (BNA) format. These files can be used in desktop mapping packages to create maps of various types of geography such as counties, census tracts or blocks.

-Extracts of 1990 census files with over 200 variables such as total population, total households, median household income, age and race distributions, etc. Intended to match up with the boundary files to make it easy to create thematic maps showing these data values. (The boundary and census extract files are stored in almost identically structured parallel directories within the archive.)

-Street intersections files. One file per county in the U.S. with one record for each pair of intersecting street features (as found in the TIGER files) with latitude, longitude coordinates.

-ZIP equivalency files. One file per state in the U.S. with one record for every populated 1990 census block with the latitude, longitude coordinates of an internal point and a long list of other geographies associated with the block, including ZIP code. These files can be used as a powerful tool to build geographic correspondence tables.

-A special collection of enhanced county to county migration files showing the characteristics of persons moving within the U.S. between 1985 and 1990. The Census Bureau's product code for this data collection is STP28.

How to Access

The archive can be accessed via anonymous FTP at ftp.ciesin.org by going to the directory /pub/census. You can also use a web browser to go to the CIESIN demographic data home page (http:///www.ciesin.org/datasets/usdemog-home.html) and then selecting the hypertext link to /pub/census. The latter method is preferred for your first visit since it gives a more user-friendly interface and can help you get a better overview of how the archive is organized. Experienced users may find that a direct FTP connection is more efficient, especially for retrieving files in quantity. Also, the web browser FTP function transmits all files in binary mode only; this is a slight problem for text files transmitted to PC environments since you do not get the extra carriage-returns added between records as you do with regular FTP in text mode.

Most of the data in the archive (everything but small text metafiles) is in zip-compressed format. You will need to have software to unzip at your end (pkunzip works fine); the archive has a special directory (/pub/census/src) with a selection of binary versions of zip/unzip that you can retrieve and use. Having access to and familiarity with using FTP and the unzip program is critical for anyone wanting to access the archive for more than just a few files.

The Directory Structure

The archive directory structure is fairly simple. Under the "root" directory (/pub/census/) are two relevant subdirectories:

src, containing the zip/unzip software and

usa, containing all the data for the U.S. (whenever we refer to a directory it will be assumed to be a subdirectory of /pub/census/usa.)

There are four relevant subdirectories of usa:

-stf: containing the census extract files (derived from the Summary Tape File 3 census tabulations, hence the name "stf")

-stp: containing the county to county migration files (derived from stp28, the Census Bureau's code name for the county to county special tabulation product).

-tiger: containing the mapping boundary files and the street intersections files.

-zipeq: the ZIP equivalency files.

(There are actually others, but we are restricting our discussion of the archive to these four primary subdirectories.)

Each of these major subdirectories has its own substructure to help organize the data within them. An important thing to keep in mind is that the stf and tiger directories have parallel structures so that when you navigate your way to retrieve a boundary file from tiger you should be able to repeat the path through the stf structure to obtain the corresponding attribute file (demographic extract).

The TIGER Directory

This is by far the largest and for many the most important portion of the CRDP archive. It includes a series of doc and txt files to provide various documentation and geographic code lists that can help you find things. It has its own "0code" subdirectory where you can find some code (SAS programs) that may be retrieved and used to process the data in SAS. For example, one of the files in tiger/0code is called bna2sas.sas and is a program to convert bna format boundary files back into a format that the SAS/GRAPH GMAP procedure (and, some day, SAS/GIS) can use. But the essence of the tiger directory is its collection of over 50 subdirectories corresponding to the states and territories. To access data for Missouri you use:

```
cd /pub/census/usa/tiger/mo .
```

Each of the tiger/ss subdirectories has an identical structure of four subdirectories. (Note: you really do not have to "learn" all this — as you issue cd commands in FTP to "descend" the directory tree, helpful readme/.message files are automatically echoed to the screen to guide you.) Most of the data is contained within the bna_st subdirectory; it is where boundary files for all the substate geographic levels are found except for the block level files. The bnablk subdirectory is where you can find the set of block boundary files, one file per county in the state. The csvisd (Comma Separated Values, InterSections Data) subdirectory is where you can find the street intersections data. A fourth subdirectory may be present for some states and may or may not be empty. It is called xptsdl and contains a set of SAS transport format files that represent the TIGER line files for each county. It is beyond the scope of this paper to deal with these files; suffice it to say that they would be of interest to a very small audience of persons who wanted to use SAS to access street level data from TIGER.

Descending down one more level to the bna_st subdirectory may be one of the more daunting steps of your journey through the archive. When you type "ls" to get a listing of this directory, what you see will vary from state to state, but will have a common structure. The explanations for the file names here are provided in the .message file from the previous level. You need to look at those notes very carefully so that when you look at the /tiger/mo/bna_st directory listing you will understand that, for example:

ap29.zip is the A-pumas file for the state (29=mo FIPS code).

bg291740.zip is the block group file for metro area (MSA) 1740 in Mo.

You could get the file tiger/msacodes.txt to have a listing of all the MSA codes in the U.S. that you'll need. Another way to know the codes is to first get the stf data for these areas; when you cd to stf/mo you'll see a .message file that tells you the names to put with the metro area codes. You would then know that 3760 is the code for Kansas City and that would let you deduce that t293760.zip (back in the tiger/mo/bna_st/mo directory) is the tract boundary file for the Missouri portion of the Kansas City MSA. What you'll note is that there is one bg (block group) and t (tract) boundary file for each metro area in the state plus one for the pseudo-MSA 9999 (remainder of state). The other files are statewide: bp29=B PUMA's, c29=counties, fm29=FIPS MCD's, fp29=FIPS places (cities), m8_29=1980 MCD's, and t8_29=1980 tracts (where defined).

The TIGER line files from which these mapping files were created were supposed to be "perfect" with respect to what is called "topological closure" — which really just means you can take all the line segments in them and create chains of segments that form the boundary of any geographic area and those line segments should form a complete, closed polygon. But perfection is difficult to achieve and our efforts in trying to achieve a 100% complete mapping data base has not yet been attained; we are at 99.8% and the chaining rate tends to be best at the smallest levels. The file tgr29all.zip (in the Mo directory) is a report file that contains information on the chaining process for this state.

If you select either the bnablk or the csvisd subdirectories from the tiger/mo subdirectory, then an "ls" will yield a similar list of 115 files corresponding to the 115 counties in the state. Note that with it is very easy with FTP to go to one of these subdirectories and enter the command

```
mget *
```

to retrieve all the files in that subdirectory. But be careful before you do this as it may involve transmitting more data than you will have room for on your disk. Also keep in mind that these are all compressed files and that they will typically grow to be 3 to 5 times larger when they are unzipped. You need to consider strategies for how you are going to process this much data and where you are going to store the results. And as long as we are saying be careful, the most frequent error users make when accessing the archive is to forget to issue a "binary" command before getting ".zip" files. You should only have binary turned off when retrieving text files.

The STF Directory

This directory has a structure which very closely parallels the tiger directory. It has 51 state-level subdirectories (states and DC only) and a "us" subdirectory that is rather special. It also contains a number of explanatory text files (such as "xtabs3" and "fafvar") that should be retrieved and examined carefully. There is also a 0code subdirectory with some useful programs for processing the files you get. It includes a SAS program that will read the pair of .csv files as downloaded from the archive and turn them back into the SAS datasets from which they were originally derived. The us subdirectory, a very recent enhancement, allows you to access all data of a certain type (e.g. tract level data or ZIP level data) in a single subdirectory. This directory is done by creating a series of links to the actual data files which physically reside in the 51 state-level subdirectories.

As mentioned in the discussion of the tiger subdirectory, one of the nice things about working in the stf directory is that when you cd to a state subdirectory, a .message file is displayed to the screen which is an annotated table of contents. For example when I type "cd stf/mo" I see:

```
CONTENTS of mo
File        # recs   Description
county<a/b>    115    MO counties
mcd<a/b>     1,367    MO mcd/ccd's
tr1740<a/b>     28    tract/BNA's COLUMBIA, MO
tr3710<a/b>     31    tract/BNA's JOPLIN, MO
                         .....
placec<a/b>  1,014    MO places (cities)
place<a/b>     960    MO places (cities) <sic>
metro<a/b>       5    MO (c)msa's
zip1740<a/b>     9    5-dig ZIPs (1991) COLUMBIA,
MO
...
zip9999<a/b>   714    5-dig ZIPS (1991) Non-metro
rmndr of MO
```

What you are reading here is a file that was actually written by the same program that wrote all the files that it is summarizing. The program did not know they were going to be zipped so it does not show the ".zip" extension that is actually present on each of the files. The "<a/b>" notation indicates that all these lines represent a matched pair of files; what we really have are two files, countya.zip and countyb.zip, that contain the data for the 115 Missouri counties. (When you retrieve these files and unzip them what you have inside are countya.csv and countyb.csv.)

You may note similarities with the tiger subdirectories but important differences. The names are not identical; c29.zip is used for the BNA file in the tiger subdirectory, county is used here. The breaking up of the tract and block group data by metro area is the same, although "tr" is used here in naming the tract files, just "t" for the mapping files. The metro files here have no match in the tiger mapping files. Neither do the ZIP data files. This is an important point: the archive does not have files for mapping by ZIP, but it does have census data by ZIP.

Processing Retrieved Data

We have talked about the structure and content of the data archive. Now we want to discuss how a user might process the information in the archive once it has been transferred back to their computing site. Our primary emphasis will be on mapping applications within a GIS (geographic information system).

Obviously, the first step in processing almost any of the files retrieved from the archive is to de-compress it using unzip or pkunzip. If you do not have one of these programs you can download one for your operating system from /pub/census/src. If there is no module there for your system you may have a problem. MVS is one platform for which I know of no public software that will "unzip" data sets.

BNA files

Mapping boundary files are in "bna" format. This is a relatively simple ascii format that is used by Strategic Mapping, Inc. (the producers of Atlas*GIS/PRO software) as the standard format for importing/exporting their mapping files. In order to use one of these .bna files you are first going to have to convert it to a format that is directly usable by whatever software you plan to use to produce your maps. If you are using one of the SMI products (Atlas*GIS or Atlas*PRO for DOS or windows) then you need to also obtain their IE (import/export) utility program. The BNA files in the archive have been created with the capabilities of the ie utility in mind. If you go in to edit/browse a .bna file you will see something that looks like this:

```
 "211739802.00106A","106A","12117390802.00106A","blk",4
-83.940732, 38.067433 -83.942302, 38.066794 -83.940591,
38.067890 -83.940732, 38.067433
```

These 3 lines represent one polygon. The first line is the header record which identifies the feature with a series of 4 identifier fields and then gives the count of points to follow. The next two lines contain the latitude longitude coordinates of the points which describe the shape of the polygon. This particular example (borrowed from the example in Henk Meij's APDU paper) is for a census block in county 21117 (somwhere in Kentucky.) It is block 106A in tract 9802.00 (BNA 9802.00 to be technically correct, but we hate to use the Census Bureau's correct terminology for these tract equivalents since "BNA" already has one meaning in these discussions). The first quoted field on the header record is called "polid" and is a unique string consisting of the FIPS state (21) and county (173) codes, the tract/bna code (9802.00) and then the block code. (When and if we ever put block level data in the archive you can be sure that the first field on the corresponding data file for this block will have a value identical to this polid field for ease of linking the data to the polygon.) The second field on the header, "106A" is called "name2" and gets stored as the secondary name field when ie imports the file. This field is typically used as a name field that is often used as a polygon label for identifying the polygon on plots. The 3rd field on the header, "name3", has a value identical to the first but with a digit "1" pre-appended. This gets into some technical stuff about Atlas*GIS, including the differences between the DOS and Windows versions. In the Windows version, this field is used as the unique (across all layers in a file, not just for this layer) internal identifier. In DOS that identifier was generated for us and we did not have to worry about it, but in Windows we have to create our own. Finally, the 4th field on the header line has the value "blks". This is used as a layer identifier value. It says to store this polygon in a layer to be named blks. If you wanted to create a mapping file with counties, tract and blocks, all together in one mapping file, you could concatenate your 3 bna files for the different levels and import them as a single file. The result would be a single geographic file with 3 layers; the names of those layers would be based on the value that appear as the 4th field of the header records in these files. You can always easily rename them after the geographic file is created. Atlas users should note that concatenating files in this way will make importing coverages much faster because Atlas will only have to index the files once. When you add layers one at a time, the program has to keep going back and regenerating the indices which can be very time consuming for large files.

Note that the last field on the header is a numeric node count with a value of 4. But if you look at the data on the next two lines you'll notice that the last coordinate pair matches the first so what we really have is a 3-point polygon; this block is a triangle.

CSV Files

".CSV", for "comma separated values", is the standard extension for these files in the Windows environment. All stf3 extracts are stored in this format, with the first record of each file containing the field names for the variables instead of data values. This format was specifically designed for ease of import into the ATLAS*GIS/PRO software packages. Importing these data files and linking them with the corresponding regions file is very simple with the DOS version of the A*G products. There are some problems with the Windows version which are associated with the new "standard" for processing CSV files which are incompatible with the DOS standard. The problem has to do with the software changing character fields which do not contain non-numeric characters to numerics, even though their values are enclosed in quotes. This same problem occurs when importing the values into Excel.

CSV files can be easily read by SAS using the appropriate infile statement options. There are a number of sample programs in the archive which illustrate this. The x3csvsas.sas program in the stf/0code directory should be retrieved and used by anyone wanting to process these data with SAS. (In some cases these data are also available in the form of SAS export-format files in stf/xpthix subdirectories — but not for all states.)

It is beyond the scope of this paper to go into detailed descriptions of the kinds of data analysis applications that can be performed using these data. There is a good chance, however, that examples of such applications will be added to the archive in the future.

Possible Future Directions

There are a number of areas in which the archive is looking to expand in the near future. Among the kinds of data that have been discussed for possible inclusion are:

-1980 STF3 data comparable to the 1990 data.
-1990 STF4 data.
-1990 STF3 data allocated to PUMA geography.
-1990 place-of-work data from the CTPP package.

In addition to more data, there is a strong possibility of new processing options for persons accessing CIESIN's WWW site. Custom on-the-fly data extracts and using interactive maps to allow users to choose geography (by pointing to the map) and to display results are among the items that are being considered. An experimental mapping application (called "YAVEMS", for "yet another very experimental map server") is already available as a prototype of the sort of things to come.

And finally, there are plans to prepare a detailed User's Guide to help people find and utilize the considerable but occasionally confusing contents of the archive.

Credits

Most of the credit for the CIESIN archive goes to Hendrik Meij of CIESIN. While we in Missouri wrote some programs Henk is the one who had the vision of putting those programs to work and creating this public archive where people could benefit from the resulting data products. We began building the archive prototype in 1992 as our "hobby project". It received no formal institutional support until it was adopted by the SEDAC people in 1994. Clearly, the people at SEDAC/CIESIN deserve considerable credit for the support they have provided the project since then.

The Missouri State Census Data Center in particular and the Census Bureau's nationwide State Data Center program in general provided the environment in which we were able to develop our software tools for processing the census files. Since we were an SDC agency the Bureau provided much of the data used in the archive at no charge. And, of course, none of this would have been possible without the excellent raw materials we had to work with, virtually all of which was provided by the U.S. Census Bureau. They have also been very supportive whenever we have had to go to them with questions about their data products.

Much credit goes to the data archive at Lawrence Berkeley Labs (LBL) where we were able to go and obtain (via FTP over the internet) many of the raw STF files and some of the TIGER line files. We also had files provided by the New Jersey State Data Center (at Princeton) and the California SDC. We have also received much valuable feedback, constructive criticism and inspiration from many people all over the country — most of them subscribers to the SASPAC-L listserver. I would name names but the list would be too long and I'd surely leave some out.

And, last but not least, much of the real credit for the success of this project goes to the Internet itself. Not only is the archive a concept that would not make sense without the internet, it is one that was created on and because of the Internet. It was the net that brought together two people with a shared interest in the application, one from Missouri, the other from Pennsylvania - who otherwise would probably have never heard of each other, much less been able to collaborate on such a technical, data-resource intensive undertaking as the building of this archive. The Internet was the critical tool that allowed quick sharing of ideas, of source code, of sample files, of user comments, of sample outputs, and finally of the resulting products. And when we thought we could never afford to obtain all the STF3 data on tape and convert it, we were able to overcome that barrier by using the Internet: LBL supplied the raw data as .dbf files on from their cd-rom based file server and a person named Richard Hockey in Australia was able to make us aware if his SAS code to convert .dbf files to the SAS format we needed. That tool came to us via the SAS-L listserv and then via FTP to fetch the advertised code from its home library in Australia.

* Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.