Making Sense of the Census, Part II: Working with Census Data

Thomas H. Grayson
24 January 2002

Objective

Extract data from the 1990 U.S. Census of Population and Housing and produce a map that display them.

Resources and Tools

STF 3A documentation from the U.S. Census Bureau
U.S. Census data in DBF format
Microsoft Access
ArcView

Review: How the 1990 U.S. Census Data Are Organized

Note: The notes from the introductory census lecture may prove helpful.

The STF 3A (Summary Tape File 3A) contains long-form, sampled Census data organized using Census-defined geographic boundaries, for example:

State (summary level 040)
County (summary level 050)
Tract (summary level 140)
Block Group (summary level 150)

You can access data on **population** (people) and **housing** (dwelling units).

The data are provided by the Census in dBASE (DBF) format on CD-ROMs. These are available at Rotch Library.

Data are provided at many levels of aggregation, represented by summary levels, some of which are shown above.

Relevant Census documentation includes:

- Notes from the "Introduction to the U.S. Census of Population and Housing" lecture
- List of Tables (Matrices)
- STF 3A Subject Locator
- STF 3A Table Definitions
- Data Dictionary (Table Elements)
- Summary Level Sequence Chart
- "How to Use This File" (for locations of tables in specific DBF files)
- Text list of the combined state/county FIPS codes (County codes are available in more convenient form from EPA)

Overview

Identify the census data items needed in the STF 3A documentation

If normalizing the data is required, identify the universe and plan on extracting that too

Find the census "tables" where the data items and universe are stored

Find the items within those "tables" that are needed

Identify and find the dBase files that contain the relevant "tables"

Extract the data from the census dBase files using the query tools in Microsoft Access

Build a unique census record identifier for the extracted data so the records can be joined to a geographic theme's attribute table

Save the results in a dBase file so that they may be read easily by ArcView

Use the results to generate a thematic map in ArcView

Example Task

Produce a thematic map of:

Middlesex County, Massachusetts

that displays

Median household income

for each U.S. Census

Block group.

Process

Note: These instructions were written for use on a Windows 2000 PC where the local hard disk is drive C:, using Microsoft Networking as configured at the CRL. Windows 95, 98, ME, and NT 4.0 should perform similarly, but these platforms have not been tested.

Grasp the Structure of the Problem.

Our task has a defined geographic scope.
We are interested in data for **Middlesex County, MA**. We don't care about other states or other counties in Massachusetts.

The task has a defined level of analysis.
We are interested in data at the **block group level**. We are not concerned with other levels of aggregation such as counties or census tracts.

We are looking for a specific data item.
While the census contains a rich set of data, for this task we are searching only for **median household income**.

Identify the Table of Median Household Incomes.

Examine the STF 3A Subject Locator. Search this page for "Median Household Income" at the left margin. Ignore entries that contain the word "imputation" or where the text is not at the left

margin. You should see "Households" immediately below it. By looking in the column to the right of "Households," we can see that the information we want is in table **P80A**.

Identify the Required Data Items in the Table.

Look at the STF 3A Table Definitions. Find table P80A by searching for the string **"P80A."** (including the period) in your web browser. This table has only one item, "Median Household Income in 1989," in item **P080A001**. Note that other tables have many data items (e.g., the "Household Income in 1989" table, P80).

Identify the Files Where the Tables Are Stored.

Look at the "Using the File" section of the STF 3A documentation. Here you will see listed which tables are in each of the 35 DBF files in STF 3A under the heading "Data Tables In Each Segment." A look at this list indicates that table P80A is in segment **STF314**. The state abbreviation is appended to the segment name in the file name containing the data. Hence, this segment corresponds to the file **stf314ma.dbf** on the CD-ROM (and on the network copy of the CD-ROM you will access shortly).

Locate the DBF Files on the Network or CD-ROM.

The 1990 STF 3A files for Massachusetts and New Hampshire can be accessed directly over the network using the directory **J:\stf3a1990\ma_nh**. The contents of this folder are identical to what you would find on the Census CD-ROM for these states.
Note that this network copy of the 1990 STF 3A files for Massachusetts and New Hampshire is available only here at MIT. If you wanted to access the files elsewhere or you wanted data for other states, you would have to find an alternative source. Other U.S. Census data CDs are available at Rotch Library and other U.S. Government document depository libraries. In addition, you can find the 1990 STF 3A files online from the U.S. Census Bureau via HTTP and FTP.
Optional: Because the Census DBF files are quite large (often 10-20 MB), accessing them from CD-ROMs may be slow, and using them over the network can be both slow and unreliable. In addition, you may want to store them locally so that you have the files permanently available. On a PC, use the Windows Explorer to copy the files from the Census disc to your personal folder in the public space on a PC (e.g., C:\PUBLIC*username*).

Open the Census dBase File in Microsoft Access.

Each segment (i.e., DBF file) contains many census "tables." We want to extract the minimum items needed for our analysis: the key columns and the data items we identified earlier. We can use the query capabilities of Access to perform this task. First, we need to open the dBase file so that we can access it. Add the dBase file as follows:

- Launch Access. When the initial dialog box appears, choose the option to "Create a New Database Using Blank Database." Name the file **census.mdb** and save it in your private CRLspace folder, **H:\private**. (If you don't have a CRLspace folder, use a Zip disk or the local hard disk instead.)

- In the Database window, make sure the "Tables" tab is active; this is the default after creating a new blank database. This is the window that is called 'census : Database'.

- Click on the "New" button to open the "New Table" dialog box. Select "Link Table" here and click OK. ***Be careful not to select "Import Table" here by mistake.*** Because the STF3A files for Massachusetts are fairly large, you don't want to make your Access database unnecessarily large by importing them, especially since we are interested only in a subset of the data.

- In "Link" dialog box, set the "Files of type:" drop down list item to **dBASE IV (*.dbf)**. Navigate to the directory where the files are stored; for Massachusetts

and New Hampshire, this is **J:\stf3a1990\ma_nh**. Then select **stf314ma.dbf** and click on the "Link" button.

- Now the "Select Index Files" dialog box will appear. Click the "Cancel" button to make this box go away.
- If all goes well, you should now see a message box that states "Successfully linked 'stf314ma'." Click "OK" to acknowledge the message.
- You'll now be back at the "Link" dialog box. You could link more tables into your database if you wanted to at this point. Since we need no more for this task, click "Close" to dismiss the dialog box.
- You should now see the linked table in your Database window with a "->dB" icon to the left of the name.

Find the County Code.

Recall that we are only interested in extracting data for Middlesex County. This means we need to find its FIPS (Federal Information Processing Standard) code. The census disk contains a file called **cnamesma.dbf** that lists all the county codes for Massachusetts. However, it's much easier to look this up using one of the lists on the web:

A text list of the combined state/county FIPS codes for the entire U.S. from the U.S. Census Bureau. Authoritative, but hard to read
A more readable list at California State University, Northridge
Another list, including 1990 population counts, from CIESIN SEDAC
Yet another list, this time from the U.S. Environmental Protection Agency. Organized by state; click on the state abbreviation to get a table for a particular state. This one is probably the easiest to use.

By reviewing one of these sources, we can quickly determine that the county code for Middlesex County is **017**.

Find the Appropriate Summary Level.

Census data is tabulated at many levels of aggregation: states, counties, census tracts, and so on. These levels nest within each other in a hierarchy. All levels of the geographic hierarchy -- what the Census calls a "summary level" -- included in the STF 3A tables. All the levels will be visible if we perform a query without restricting the summary level. For this exercise, we are interested only in the block group summary level. To find the numeric code for this summary level, examine the Summary Level Sequence Charts. In the row marked "State--County--Census Tract/Block Numbering Area--Block Group" the table indicates that the correct code is **150**. Note that this should not be confused with summary level 090, which uses the "place" geography above block group in the hierarchy. In many states (but not Massachusetts), using the summary level that uses the "place" geography will eliminate all block groups that are not within either incorporated communities or other densely populated but unincorporated "census designated places." Also, block groups and hence tracts may cross place boundaries, complicating the analysis.

Select the Desired Rows.

Create a new query in Microsoft Access. Add the **stf314ma** table to the query. As we noted before, we want to restrict the summary level (**SUMLEV**) to "150" and the county (**CNTY**) to "025. Add the following columns to the query grid and set the criteria as noted below:

Column	*Description*	*Criteria*
SUMLEV	**Summary Level**	**"150"**
STATEFP	**State FIPS code**	

CNTY	**County FIPS code**	**"025"**
TRACTBNA	**Census tract/Block Numbering Area ID**	
BLCKGR	**Census block group ID**	
LOGRECNU	**Logical record number**	
P080A001	**Median household income in 1989**	

When you are ready, run the query. When accessing the files over the network, this process may take several seconds, but under a minute.

Examine the Block Group Geographic Layer in ArcView.

Launch ArcView. In a new View window, add the theme of Massachusetts block groups in the shapefile **K:\11.208\arcviewfiles\stateplane\mablkgrp.shp** to your project. Display the **Mablkgrp** theme's attribute table using the **Theme > Table** menu item. Examine the format of the block group identifier, **Bkg_key**. The **Bkg_key** column contains numbers such as "250173001001", "250173182009", "250173872026". Compare this column with the data returned by the Access query. The **stf314ma** table has no column that matches **Bkg_key**. We do, however, have all its component parts in the **STATEFP, CNTY, TRACTBNA,** and **BLCKGR.** We need to use these components to assemble a block group identifier that is compatible with **Bkg_key**. Observe that **Bkg_key** contains numbers of the form

SSCCCTTTTXXG

where

SS is the state FIPS code (e.g., **25**),

CCC is the county FIPS code (e.g., **017**),

TTTTXX is the census tract number (e.g., **300100**), and

G is the census block group number (e.g., **1**).

Let's take a closer look:

SSCCCTTTTXXG

250173001001

250173182009

250173872026

Note that all these codes include leading zeroes; the **Bkg_key** is always 12 characters long. We need to construct the equivalent **Bkg_key** in the **medhhinc.dbf** table. What makes this process difficult is the representation of the tract number. The **Tractbna** field in the STF tables uses the format TTTTXX,but the XX part is *omitted* if it would be zero (e.g, "3001", "3182", "387202"). In other words, the field is sometimes 4 characters long (when the two trailing zeroes are omitted), and sometimes it is 6 characters long.

Formulating a Solution.

So, at this point we have the block group geography available, plus some interesting data to link up to it, but no easy way to join the data we extracted from the census tables to the block group theme. To fix this, we need to create a **new field** in the Access query that contains the needed key.

Add a Column to the Access Query.

The tract number poses some difficulties. The **TRACTBNA** field in the STF tables uses the format TTTTXX , and the XX part is omitted if it would be zero (e.g, "3001", "314398"). In other words, the field is sometimes 4 characters long (the two trailing zeroes are omitted), and sometimes it is 6 characters long.
To cope with this, we can use two of the functions built into Access:

Function	***Purpose***	***Examples***
Len(*string*)	Returns the length of *string*	**Len("ABCD")** returns 4 **Len("ABCDEF")** returns 6
IIf(*test_expr, true_expr, false_expr)	Evaluates *test_expr.* If *text_expr* is true, returns *true_expr.* Otherwise, returns *false_expr.*	**IIf(numval > 1000, "High", "Low")** For numval=2000, returns "High" For numval=10, returns "Low"

We can use these functions to test whether a particular tract identifier (**TRACTBNA**) has more than 4 characters. If it does, nothing needs to be done. Otherwise, we need to "pad" the **TRACTBNA** field with two extra zeroes to fill out the length. The following expression will do the job:
[STATEFP] + [CNTY] + [TRACTBNA] + IIf(Len([TRACTBNA]) > 4, "", "00") + [BLCKGR]
Since we want to give this new column a meaningful name, rather than the name "Expr1" that Access will assign by default, we can add the name **BKG_KEY** as shown below:
BKG_KEY: [STATEFP] + [CNTY] + [TRACTBNA] + IIf(Len([TRACTBNA]) > 4, "", "00") + [BLCKGR]
We want this column to be the first in the resulting table. So, click anywhere in the first column in the query grid, then select **Insert > Columns**. A new blank column will appear. In this column, copy-and-paste (or type) the expression above into the "Field:" cell.
Run the query again. Notice how the **BKG_KEY** column has 12 characters regardless of whether the **TRACTBNA** column has 4 or 6 characters.
Note that we would have to modify this procedure if there were any tract numbers with 5 digits as weel as 4 or 6. *(What would you do?)*

Turn of Unneeded Columns.

For our purposes here, we really only need to bring the columns **BKG_KEY, LOGRECNU,** and **P080A001** into ArcView. Therefore, turn off the "Show:" checkbox for the columns **SUMLEV, STATEFP, CNTY, TRACTBNA,** and **BLCKGR.**

Save the Query.

Close the query. When prompted for a name, call it **medhhinc**.

Save the Results as a dBASE DBF file.

ArcView preferred database format is a dBASE DBF file. Hence, we will save a copy of our query's output as a dBASE DBF file. In the Database window, make sure that Query tab is active and that the query **medhhinc** is selected. Now, choose, **File > Save As/Export**. Follow the dialog boxes to save an **external** file in your **H:\private** folder called **medhhinc.dbf**. *Make sure to specify "dBASE IV (*.dbf)" in the "Save As Type:" field of the final dialog.*
We are now finished with Access and will be doing the rest of the work in ArcView.

Open the 'medhhinc.dbf' Table in ArcView.

In ArcView, add the table **medhhinc.dbf** to the project. Note that the table contains only the rows and columns we selected earlier.

Join the 'medhhinc.dbf' Table to the Attributes of the Block Group Coverage.

Use the common field in both tables to join them together. Select the **Bkg_key** column heading in **medhhinc.dbf**, then in the theme attribute table, "Attributes of Mablkgrp.shp." After verifying that "Attributes of Mablkgrp.shp" is the active window, use **Table > Join** to join the tables. The **medhhinc.dbf** table should disappear if you have selected the right table when initiating the join. *You must have the correct table active when selecting **Table > Join** or else you will not be able to map the new attributes!*

Create a Thematic Map.

We're now *finally* ready to make a thematic map!
Use the column **P080a001** to create a graduated color map. Set the null value to be zero and display the "No Data" range. You should see only Middlesex County block groups shaded in. Note that only the block groups in Middlesex County are shaded; that's because we only extracted census data for this one county. If we had pulled off data for the other counties shown (e.g., Essex, Norfolk, Plymouth, and Suffolk), they would have been shaded as well.

Consider a Different Example.

Suppose we had wanted to visualize something a bit different from median household income--the fraction of incomes less than $15,000 in each block group. To compute this, we would need to add together several columns from the "Household Income in 1989" table (P80). In addition, we would need to normalize the data by dividing it by the appropriate universe, here "Households". We can compute the universe, total households, by adding up all 25 columns in table P80. Alternatively, we can obtain the total households from the table "Households" (P5) in item P0050001. The value obtained by summing the 25 columns in table P80 should be the same as the single column P0050001; you can check up on the Census Bureau by comparing these values! The sum of the household counts in various income categories would need to be divided by this value. By taking another look at "Using the File", we can see that tables P5 and P80 are stored in different DBF files. We can use the field "Logrecnu," the logical record number, to link the extracts from the two files together.
Why normalize the data? Comparing the raw numbers of housing units per block group may be deceiving, as the total number of households will vary from one block group to the next. By dividing the number households with income $15,000 by the total number of households, we obtain a *fraction* of the total occupied households with incomes under $15,000. This fraction *may* be compared fairly among block groups.

THE END!

The initial source for these notes is a lecture presented by Laura Lebow on January 24, 1995, as recorded by Qing Shen. Those notes were adapted for online data files and MapInfo by Thomas H. Grayson, Fall 1996. Modified by Thomas H. Grayson for MapInfo 5.0®, Microsoft Excel 97®, and Microsoft Query® in January 1999. A version of these notes prepared for the class 11.521, Fall 1998, by Thomas H. Grayson. These notes, in turn, were modified for 11.520, Fall 1999, by Anne Kinsella Thompson with minor edits by Thomas H. Grayson. This document is represents a merging of the IAP 1999 11.208 notes (Excel®/MSQuery®/MapInfo®) with the 11.520 Fall 1999 notes (ArcView® only), and presents an Excel®/MSQuery®/ArcView® procedure.