Making Sense of Census Data

Robert Matthews, University of Alabama at Birmingham, Birmingham, Alabama

ABSTRACT

The United States Census Bureau publishes a vast amount of data on many different facets of the U.S. population. One of the most utilized resources is the data available from the decennial census. This data is summarized and stratified in many different ways and can be used for a myriad of purposes. The task of actually extracting particular pieces of information from the census data can be daunting since there are literally hundreds of files containing thousands of individual variables. Not only is the data itself voluminous, but the documentation is also very extensive. An example of extracting three variables from the 2000 census data is presented in this paper to illustrate the tasks involved in reading, understanding, and using the data.

INTRODUCTION

The official U.S. Census is described in Article I, Section 2 of the Constitution of the United States. It calls for an actual enumeration of the people every ten years. The results of the census are used for many different purposes, but one of the most important is the apportionment of seats in the House of Representatives among the states.

Most Census data are available for many levels of geography, including states, counties, cities and towns, ZIP codes, census tracts and blocks, and much more (see Figure 1). The questions that are asked can be grouped into two categories, the “Short Form” and the “Long Form”. The questions on the Short Form are asked of every household in the U.S. The Long Form questions are only asked for a sample of the population (1 in 6 households).

The data from the “Short Form” questions are available in 3 major categories:

- Redistricting Data – used for congressional and state redistricting
- Summary File 1 (SF 1) – data from the Short Form questions
- Summary File 2 (SF 2) – data from the Short Form questions, repeated for 249 population groups

The data from the “Long Form” questions are available in 2 major categories:

- Summary File 3 (SF 3) – comprehensive results from the Long Form
- Summary File 4 (SF 4) – comprehensive results from the Long Form, repeated for 335 population groups

WORKING WITH CENSUS DATA

Our task was to obtain some of the variables available in the Census data for a cohort comprised of a 5% sample of the entire U.S. Medicare population. The steps involved with actually merging the Census data with our Medicare population were fairly straightforward, but somewhat challenging because of, 1) the size of the data, and 2) the complexity of the Census data.

The first task was to obtain the Census block group information for each beneficiary in our cohort. This involved finding a data source that we could use to link the 9-digit zip codes in our Medicare population and obtain corresponding Census block group information. We evaluated several products and chose to use a database from Melissa Data Corporation to perform this linkage. This database contained

approximately 66 million records at the 9-digit zip code level. We also had zip codes at the 9-digit level in our Medicare population and were able to obtain a 98% linkage between our data and the database from Melissa Data. Most of the remaining 2% that did not match were due to either invalid or unknown zip codes or for zip codes that had been added or deleted before the Melissa Data database was created. Our data spanned the years 1999-2006 and the data we obtained from Melissa Data was from 2006. We attempted to locate historical information for prior years, but were unable to identify a suitable vendor.

The next step in the process was to decide which Census files to use. Summary files 1 & 2 were not very useful for our purposes because they contain only a small subset of the variables included on the Long Form. Summary File 3 contains data on such topics as income, home values, ancestry, commute time to work, occupation, education, veteran status, language ability, migration, place of birth, and many others. Since we were interested in creating a measure of Socioeconomic Status (SES) we wanted variables that were not available in SF 1 or SF 2, therefore we chose to use SF 3.

The SF 3 data set consists of one geographic identifier file (Geo ID) and seventy six data files. The Geo ID file is not a "header file" as it is linked horizontally with the data files, not placed on top of them vertically. The Geo ID file is fixed width with no field delimiters while all 76 data files are variable length with comma field delimiters. It is linked to each of the 76 data files using a unique key field named LOGRECNO. There are 77 files (Geo ID + 76 data files) per state for a combined total of 3,927 files (50 states + D.C.). There are also two additional sets of 77 files each for all the states combined and for Puerto Rico. Since each data file contains only certain variables, it is imperative to identify the variables needed for an analysis and only reference the specific files.

The Census data is stratified in many different ways. The variable, SUMLEV, which is available in the Geo ID file, is used to select the stratification level(s). This field identifies the *summary level* (area type) of each record. A combination of the geographic identifier codes for each element in the complete summary level description is used to identify the *specific* area being tabulated. See Table 1 for an example of the Summary Level Sequence Chart and Figure 1 for a diagram that shows how the different stratification levels relate to each other.

The Subject Locator is an index designed to quickly identify the tables in the summary file for particular subjects or topics of interest. This index is arranged in alphabetical order by the name of a subject. Under each subject heading appears the type of entry being tabulated (shown in italics) and the relevant table number (*see Table 2 for an example*).

Summary of steps for identifying variables and merging data with an existing cohort:

1. Use the Subject Locator to identify variables of interest and their corresponding Table Numbers (*see Table 2*)
2. Use the File Segmentation Table to identify the specific data file(s) for each Table Number (*see Table 3*)
3. Use the Summary Level Sequence Chart (*see Table 1*) to locate the desired stratification level
4. Identify SAS® input statements to read each specific data file
5. Read the specific variable(s) of interest into SAS (*code is available on my website*)
6. Merge census variables with existing cohort data for data management and analysis

For our specific purpose, we were interested in obtaining Educational Attainment, Median Household Income, and Total Population from the Census data. Using the Subject Locator we found that Median Household Income was located in Table number P53 (*see Table 2*), Educational Attainment was located in Table number P37, and Total Population was located in Table number P1. We then looked each of these table numbers up in the File Segmentation Table and found that P53 was in file 06, P37 was in file 04, and P1 was in file 01.

The Census Bureau has SAS input statements available for each of the 77 different file types. We downloaded the programs for each of the file types that we needed and created a program that read in the Geographic Identifier file for the specific summary level that we were interested in (SUMLEV=150). We then read each of the three data files containing the specific variables that we were interested in. We

subsequently linked the Geo ID file to the other data files using the LOGRECNO variable and created a permanent file that contained all the census data we wanted. The final step involved merging the permanent census data file with our cohort file to add the new variables to our existing data.

CONCLUSION

Working with the census data is a somewhat daunting task due to the sheer volume of data and documentation. However, the U.S. Census Bureau has organized the data into a manageable set of distinct components. The advantage of having the data in a de-centralized data base means that it is very flexible and can be adapted to many different uses. This flexibility comes at a cost, though, of having thousands of variables and data files from which to extract information. Nonetheless, after going through this entire process for one set of variables, we have found the process to be very manageable.

REFERENCES

1. Melissa Data, www.melissadata.com
2. U.S. Census Bureau, www.census.gov
3. U.S. Census Bureau, Summary file 3 documentation, www.census.gov/prod/cen2000/doc/sf3.pdf

CONTACT INFORMATION

For more documentation about how these programs work or to request a copy of the code, please contact Robert Matthews using one of the following addresses:

Robert Matthews
University of Alabama at Birmingham
Department of Epidemiology
1665 University Blvd. RPHB 517
Birmingham, AL 35294-0022

Email: rsm@uab.edu
Web: www.epi.soph.uab.edu/rsm/

Table 1. Summary Level Sequence Chart (*partial listing*)

Geographic component	Summary level
00, 01-49, 52-95	040 State
00, 01, 43, 49	050 State-County
00	060 State-County-County Subdivision
00	070 State-County-County Subdivision-Place/Remainder
00	080 State-County-County Subdivision-Place/Remainder-Census Tract
00	085 State-County-County Subdivision-Place/Remainder-Census Tract-Urban/Rural
00	090 State-County-County Subdivision-Place/Remainder-Census Tract-Urban/Rural-Block Group
00	067 State [Puerto Rico Only]-County-County Subdivision-Subbarrio
00	140 State-County-Census Tract
00	144 State-County-Census Tract-American Indian Area/Alaska Native Area/Hawaiian Home Land
00	**150 State-County-Census Tract-Block Group**
00	154 State-County-Census Tract-Block Group-American Indian Area/Alaska Native Area/Hawaiian Home Land
… more levels …	

Table 2. Subject Locator index (*partial listing*)

Subject description	**Subject Table numbers**
Median Income (Dollars)	
Families	P77
by Family Type by Presence of Own Children Under 18 Years	PCT40
by Presence of Own Children Under 18 Years	PCT39
Households	**P53**
by Age of Householder	P56
Nonfamily Households	P80
by Sex of Householder by Living Alone by Age of Householder	PCT42
Occupied Housing Units	
by Tenure	HCT12
Population 15 Years and Over With Income	
by Sex by Work Experience	PCT45
… more subjects …	

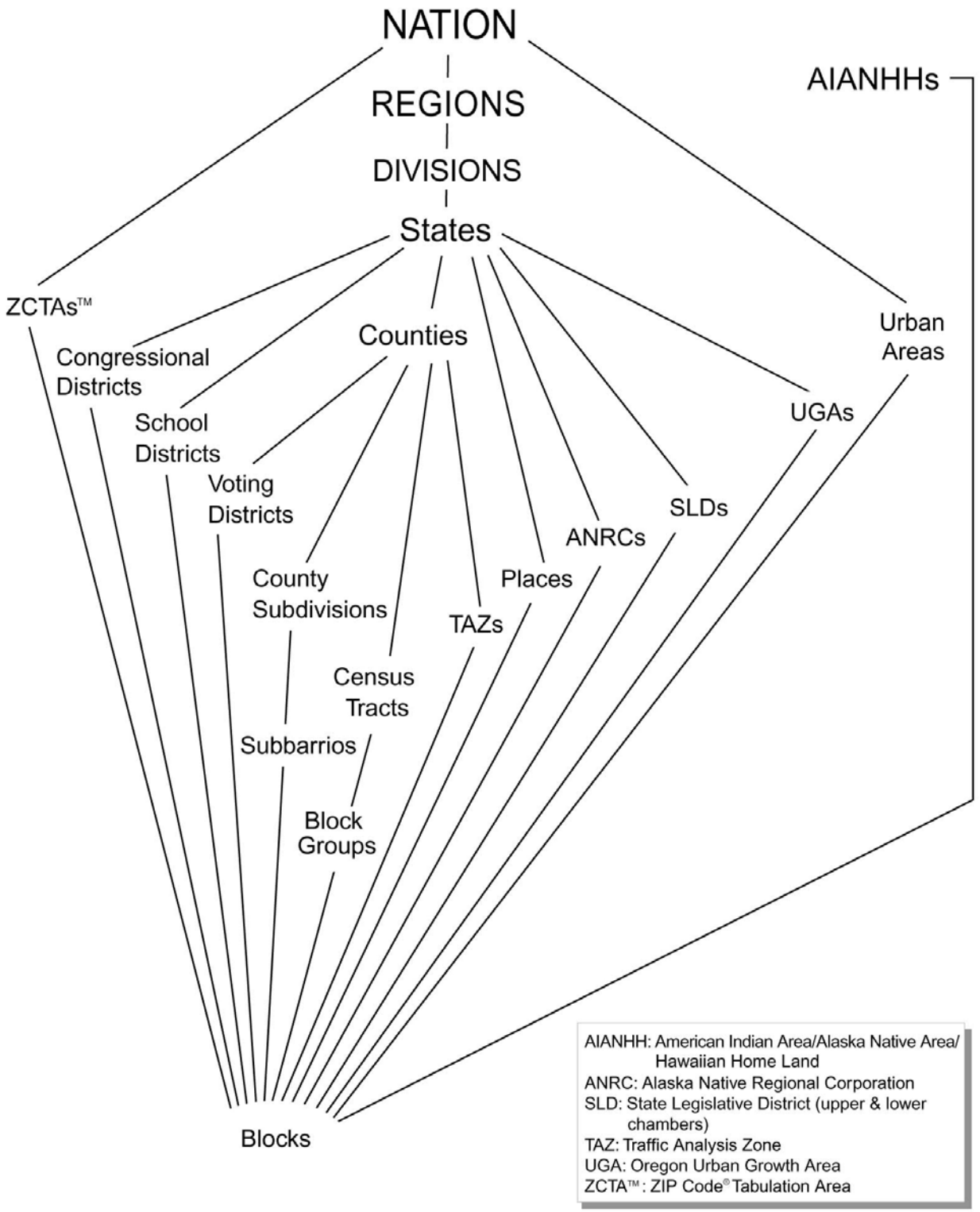
Table 3. File Segmentation Table

File name	Number of data items	Starting table number	Ending table number
Stgeo.uf3			
st00001.uf3	248	P1	P14
st00002.uf3	218	P15	P24
st00003.uf3	241	P25	P37
st00004.uf3	227	P38	P46
st00005.uf3	220	P47	P50
st00006.uf3	250	P51	P67
st00007.uf3	213	P68	P91
st00008.uf3	245	P92	P138
st00009.uf3	203	P139	P145C
st00010.uf3	245	P145D	P145H
st00011.uf3	235	P145I	P146F
st00012.uf3	246	P146G	P147I
st00013.uf3	241	P148A	P149D
st00014.uf3	245	P149E	P150I
st00015.uf3	239	P151A	P154D
st00016.uf3	240	P154E	P159G
st00017.uf3	239	P159H	P160E
st00018.uf3	164	P160F	P160I
st00019.uf3	247	PCT1	PCT8
st00020.uf3	204	PCT9	PCT15
st00021.uf3	222	PCT16	PCT17
st00022.uf3	235	PCT18	PCT19
st00023.uf3	233	PCT20	PCT24
st00024.uf3	233	PCT25	PCT27
st00025.uf3	221	PCT28	PCT32
st00026.uf3	106	PCT33	PCT34
st00027.uf3	221	PCT35	PCT37
st00028.uf3	162	PCT38	PCT43
st00029.uf3	205	PCT44	PCT48
st00030.uf3	224	PCT49	PCT51
st00031.uf3	205	PCT52	PCT56
st00032.uf3	243	PCT57	PCT61
st00033.uf3	243	PCT62A	PCT63C
st00034.uf3	234	PCT63D	PCT64H
st00035.uf3	231	PCT64I	PCT66C
st00036.uf3	233	PCT66D	PCT67E
st00037.uf3	223	PCT67F	PCT68C
st00038.uf3	245	PCT68D	PCT68H

st00039.uf3	247	PCT68I	PCT69I
st00040.uf3	243	PCT70A	PCT70I
st00041.uf3	245	PCT71A	PCT71E
st00042.uf3	196	PCT71F	PCT71I
st00043.uf3	240	PCT72A	PCT72B
st00044.uf3	240	PCT72C	PCT72D
st00045.uf3	240	PCT72E	PCT72F
st00046.uf3	240	PCT72G	PCT72H
st00047.uf3	215	PCT72I	PCT73A
st00048.uf3	190	PCT73B	PCT73C
st00049.uf3	190	PCT73D	PCT73E
st00050.uf3	190	PCT73F	PCT73G
st00051.uf3	190	PCT73H	PCT73I
st00052.uf3	231	PCT74A	PCT75C
st00053.uf3	236	PCT75D	PCT75G
st00054.uf3	234	PCT75H	PCT76D
st00055.uf3	145	PCT76E	PCT76I
st00056.uf3	127	H1	H18
st00057.uf3	249	H19	H26
st00058.uf3	216	H27	H44
st00059.uf3	250	H45	H68
st00060.uf3	248	H69	H86
st00061.uf3	250	H87	H104
st00062.uf3	59	H105	H121
st00063.uf3	171	HCT1	HCT3
st00064.uf3	115	HCT4	HCT4
st00065.uf3	143	HCT5	HCT5
st00066.uf3	248	HCT6	HCT7
st00067.uf3	219	HCT8	HCT14
st00068.uf3	214	HCT15	HCT17
st00069.uf3	220	HCT18	HCT23
st00070.uf3	248	HCT24	HCT31C
st00071.uf3	246	HCT31D	HCT36D
st00072.uf3	246	HCT36E	HCT40I
st00073.uf3	243	HCT41A	HCT43I
st00074.uf3	224	HCT44A	HCT44G
st00075.uf3	247	HCT44H	HCT47F
st00076.uf3	96	HCT47G	HCT48I

Note: *st* represents the United States Postal Service two-character alphabetic state abbreviation.

Figure 1. Hierarchical Relationships of Census Geographic Structures



Source: U.S. Census Bureau, Summary file 3 documentation