Paper 329-2013

Resources for getting the 2010 US Census Summary Files into SAS®

Rebecca Ottesen City of Hope and California Polytechnic State University

ABSTRACT

At first glance, accessing the 2010 US Census Summary Files data with SAS seems like a daunting task. The main limitation is that for the 2010 Summary Files it seems that the Census has gravitated toward supporting data access via Microsoft Access rather than SAS as they did in the past. However, there are several tactics that can be deployed to make accessing this data with SAS much easier. A thorough understanding of the Census Summary File data structure and documentation can be used to leverage both SAS code from programs that the Census previously supported and Census 2010 versioned SAS programming available through other public sources. Knowledge of the available resources can assist SAS analysts in taking advantage of this rich data set.

INTRODUCTION

Before launching into the Census 2010 data it is important to become familiar with the various products and resources that are available. There is a plethora of information about the Census data online, which can lead to hours of research in order to pinpoint the best options for the analysis at hand. This time sink is magnified by the fact that there are many 2010 Census data products available such as Redistricting Files, Demographic Profiles, the American Community Survey and the first two Summary Files (SF1 and SF2).

The best place to start when becoming acquainted with the Census 2010 data products is the main webpage itself. This page lists interactive resources for searching the data as well as a DATA tab that leads to the main page for the 2010 data. These data products were released on a product based schedule at the state and the national level. The Redistricting and Demographic Profiles were the first products to be released, and they provide a quick snapshot of select characteristics. The Summary Files, which are the main focus of this paper, contain population and housing data such as age, gender, race, ethnicity, and household information. The SF2 data is similar to the SF1 population and housing data, however SF2 contains more detail on certain race and ethnicity groups within a community. Both SF1 and SF2 are 100% count, rather than sample based data. The long form data of SF3 from Census 2000 is no longer available. The SF3 data has been replaced by the American Community Survey (ACS) which is sample based data and provides updated statistics every year on more in depth data such as income, education, occupation and more. Simply put the Summary Files are actual counts but on more general topics, while the ACS is sample data with margin of error but focused on more detailed topics. Now that the majority of these Census 2010 data products are available it is just a matter of understanding the difference between them and where to find the correct data as it pertains to an analysis.

The actual data for Census 2010 is available in many different places online and also in many different formats. Various state and academic agencies have provided the Census data for their region in SAS (and other) file formats. The Census site provides the data in its raw form, by state and nationally, and it also provides various options for reading the data depending on the data product. Most of the Census data products previously mentioned do have SAS programs available for reading in the raw data at their corresponding Census.gov site. However the SF data files for 2010 do not have SAS programs published by Census.gov as they did in the previous Census of 2000. As a result getting the SF data directly from the Census into SAS becomes somewhat of a burden in that the technical documentation needs to be deciphered and programming needs to be written. This burden can be relieved with an organized approach to researching the resources available. This includes not only the Census site, but also other institutions and the SAS community who have worked toward making this information more easily available others. This paper seeks to organize the information available and provide insights into the best resources for working with the Census 2010 Summary Files.

TIP #1: UNDERSTANDING THE SUMMARY FILE DATA SETS

There are several resources for acquainting oneself with the Census 2010 Summary Files. The first and most obvious would be the Census 2010 website (http://www.census.gov/2010census/), specifically the Census 2010 DATA page. This DATA site presents background and technical information as well as links to each Summary File 1 and Summary File 2 site. Each of the Summary File sites lists information about the Summary File itself, technical documentation, a crosswalk for comparison of 2000 to 2010, news releases, tips and more.

The most important source of information for working with the Summary Files is the technical documentation (found in the Background section) which is the masterpiece that explains it all. One of the main concepts to grasp is that the basic idea behind the segmented raw data file layout for the 2010 Summary Files is similar to what it was for the Census 2000. Therefore an understanding of the geo file and corresponding raw data files from 2000 will help considerably in reviewing the 2010 technical documentation. First time users of the Census Summary Files should be prepared to dig into the technical documentation with the assistance of a basic summary of the 'must read' chapters, shown below in Table 1.

Chapter	***Summary***	***Noteworthy tables, figures and sections***
2. How to Use This Product	The basic guide for how to use the Summary File product and the documentation.	Figure 2-1 displays an example of the file structure and the fundamental idea behind merging the geo file with the data files. Figure is 2-2 shows how many raw data sets (data file segments) are available in the Summary File, as well as the starting and ending variables names in each file. Figure 2-3 provides a visual of how the data are collected and summarized at varying levels: nation, region, division, state, county, Census tract, etc…
3. Subject Locator	The subject locator provides an index of what variables can be accessed from the Summary File data. The chapter is organized by subject and provides table numbers which can be used as a roadmap to find corresponding data table number (variable names) as listed later in chapters 5 and 6.	
4. Summary Level Sequence Chart	This chapter explains the summary levels that correspond to Figure 2-3, the hierarchical relationships of how the data was collected.	The first set of tables lists the summary levels available for the state files. The second set of tables corresponds to the summary levels for the national files.
5. List of Tables (Matrices)	This chapter lists the different tables/matrices available in the raw data sets. A table/matrix can be thought of as a cluster of variables that relate to a Census table number from Chapter 3.	The table number corresponds to the first part of the variable name, while the total number of data cells indicates how many variables exist for that table. The variable names can be found in chapter 6.
6. Data Dictionary	The data dictionary which outlines the various responses that were recorded for each table/matrix and lists the universe associated with the table/matrix.	The identification section lists all of the data dictionary reference names (variable names), field size (number of columns), and starting positions for the data in the geo file. The table/matrix section helps identify which raw data file to use depending on the subject of interest. Important identification information includes the data dictionary reference name column (variable names), the segment column (raw data file). For example P0020001 corresponds to the first variable in table P2 which is found in the raw data file (segment) 02, P0020002 is the second variable in table P2. Note that the first variable in each file tells what the universe is for the data collected.

Table 1: Chapter Highlights in the Summary File Technical Documentation.

TIP #2: PREVIEWING THE SUMMARY FILE DATA

There are several interactive data sources that can be found on the 2010 Census site which provide a quick peek at the Census 2010 data. A few of them include:

1) The Population Finder (http://www.census.gov/2010census/) provides a way to select and view Demographic Profile statistics one state at a time. There is a link to 'download summary files', however this data is complicated and structured in a similar way to the SF files. These Demographic Profiles were early release and only provide the frequently used data elements from SF1.

2) The Interactive Population Map (http://www.census.gov/2010census/popmap/) provides a very cool interactive way to access certain statistics at various geographic levels. This data is also based on the Demographic Profiles.

3) The American Fact Finder (http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml) comes a step closer to the actual loading of the 2010 Summary File data in the form of a data set. The Fact Finder contains Summary File data, as well as other Census data products such as Redistricting and ACS. It allows the user to select variables from the various data products, view the results as a table, manipulate the table and also store the data as CSV, or as a presentation ready table in Adobe, Excel or Rich Text Format.

The Population Finder and Interactive Population Map are very easy to use. Using the Fact Finder requires a basic understanding of the content in the data products such as the tables/matrices and data files mentioned previously. This tool would be useful for creating Census based data tables for a user who does not have access or the ability to use higher end software. However, it is not the best solution for a user with the skills to read the data in to SAS and manipulate it as appropriate for analysis.

TIP #3: FINDING AVAILABLE SUMMARY FILE DATA

There are primarily two websites for locating the complete raw Summary File data sets so that they can be read into SAS.

1) The source of the raw data for SF 1 and SF 2 is the Census 2010 DATA site and the data can be downloaded there by navigating to the respective Summary File main page and finding the 'FTP site' link (in the Background section). The data sets are zipped and organized in folders by state and at the national level (called National/ in SF1 and United States/ in SF2).

2) The Missouri Census Data Center (MCDC: http://mcdc.missouri.edu/) has read in all of the SF1 and SF2 data (and even more) and created SAS data sets for public consumption. These are posted for download by state and nation, and by summary level. The Summary File SAS data sets can be found in the 'MCDC Data Archive (Uexplore/Dexter)' link, by navigating to the 'sf12010/Summary File 1, Complete Tables' and the 'sf22010/Summary File 2' links. This site is also noted as a 'SAS data sets' resource on the Summary File 1 webpage at Census.gov.

Both of these sites are very detailed as there is much information to be covered. The Missouri Census Data Center has a lot of information within the directories of the previously mentioned SF1 and SF2 pages. At MCDC there are many more folders for such things as technical documentation, tools and metadata available for perusal. In addition, they also have a data extraction application for Summary File 1 (see 'SF1 Data extract app' in the quick links section on their main webpage). This tool provides the ability to create a data set based on the desired summary level and SF1 variables. The tool will also save the data file as delimited (csv, tab) file, a report (text, pdf, html), and/or a database (dbf, SAS) file.

TIP #4: FINDING AVAILABLE SAS PROGRAMS

Why reinvent the wheel? The SF data is complicated and others have graciously posted their work to allow users to work with the Census 2010 Summary Files more easily in SAS. There are several resources that can be utilized to leverage existing SAS programming when a user would like to work directly with the raw data.

1) There are well documented and straightforward SAS programs and references available from the Census for the 2000 Summary Files. These programs can be found at the respective 'Summary File' link in the ASCII text data files (ASCII overview section) of the Census 2000 Tech Talk web page (http://www.census.gov/support/cen2000_extract.html) and then navigating to the 'Using SAS' section.

Another option is to review the code for the 2010 Redistricting SAS programs posted at the Tech Tips Census 2010 redistricting data site (http://www.census.gov/rdo/tech_tips/). This Redistricting data includes the same fundamental idea of segmented files, mainly a geo file and supporting raw data files, which is analogous to the segmentation in the Summary Files. While these none of the SAS programs mentioned here will automatically read in the 2010 SF1 and SF2 files, the programming concepts in these programs are simple and they can be used together with the technical documentation to map out a SAS program to read in the 2010 Summary Files.

2) The Missouri Census Data Center provides not only SAS data but also the SAS programs used to make the SF data sets. In each of the 'sf12010/' and 'sf22010/' pages there is a folder called 'Tools' which contains SAS programs (and logs) that can be used to read in the data. The cnvtsf1.sas and cnvtsf2.sas files are the main programs used to convert the raw SF data into SAS data sets. The code is graciously available for use but to understand it requires a thorough understanding of SAS macros.

3) SAScommunity.org also provides information about working with the Census 2010 products. In particular coding for reading in the SF1 into SAS data sets can be found (http://www.sascommunity.org/) in a post based on "Converting 2010 Census Summary File 1 (SF1) Data into SAS Data Sets". This page provides a basic description of the SF1 data files, SAS code that utilizes macros to read in the data, and also programming examples. The SAS community site for the SF1 data is well organized and gets to the point quickly.

TIP #5: LINKING THE CENSUS SUMMARY FILES

Once the raw data has been read into SAS it is ready to be analyzed, but first care should be taken to link it appropriately to the analytic data set. Figure 2-3 of the Summary File technical documentation shows how the census data is collected. If the analytic data set is strictly coded to the Census supplied summary levels such as FIPS coding, Census tracts or blocks, then the merge will be easier. If the data is loosely linkable such as zip code (a zip code is not exactly the same as the census zcta) or metropolitan statistical area (the character city MSA name may not exactly match, and these definitions change over time) then more care will need to be taken. In some cases a crosswalk data set may need to be used to bring both data the census SF data and the analytic data set together. The MCDC provides detailed information about the census geography and summary levels (http://mcdc.missouri.edu/allabout/sumlevs/) which can be used to research the appropriate summary level for the analysis.

In addition, there are many other easily obtainable resources available online such as SAS user group papers that explain how to read and understand the Census 2000 Summary Files and the other Census data products. Suggested topics to review are how to: combine the geo file to the raw data sets; identify the variables of interest and corresponding data product; how to utilize the appropriate stratification level based on the summary level variable. The fundamental ideas of these papers can be used to facilitate working with the 2010 Summary Files in SAS.

CONCLUSION

The first step to working with the Census 2010 Summary Files is to become familiar with the technical documentation. Upon initial review attention should be taken to understand how the Summary File data is structured, especially with respect to the geofile and the raw data files. Once the segmented structure of the SF data sets is thoroughly digested a review of the summary levels and topics available for analysis can take place. Knowing how this data fits together and what to access is the key to working with Summary Files, no matter the tool.

In terms of which data source to use the answer depends on what the ability to work with the data will be. The interactive applications provided by the Census are nice for quick snapshots of the data. However to get the data into SAS and link it to the analytic cohort, the quickest solution is to utilize the SAS data sets available at the Missouri Census Data Center (posted or via the extract tool). However to take full control, the raw data available at the Census data FTP sites can be downloaded. This route would require leveraging existing SAS programming. A tinkerer with minimal macro experience might choose to create code using the programming concepts from other Census supplied SAS programs. An advanced macro user might choose the programming sources from the Missouri State Data Center or sascommunity.org.

With any of the methods listed above the key is to have a well organized plan of attack and an understanding of the best resources available. This will result in countless hours, that would have been spent on online research, which can be used to understand the data that is contained in the 2010 Census Summary Files and then carry out the analysis.

REFERENCES

Census 2010. http://www.census.gov/2010census/

Missouri Census Data Center. http://mcdc.missouri.edu/

Mike Zdeb, “Converting 2010 Census Summary File 1 (SF1) Data into SAS Data Sets”. http://www.sascommunity.org/wiki/Converting_2010_Census_Summary_File_1_(SF1)_Data_into_SAS_Data_Sets

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Rebecca Ottesen
Cal Poly State University, San Luis Obispo
Department of Statistics
San Luis Obispo, Ca, 93407
rottesen@calpoly.edu