POLIDATA® Political Data Analysis
DATABASE DEVELOPMENT, ANALYSIS AND PUBLICATION;
POLITICAL AND CENSUS DATA; LITIGATION SUPPORT

CLARK BENSEN

POLIDATA · 1303 HAYWARD ROAD, P. O. BOX 530 · CORINTH, VT 05039
Tel: 703-690-4066 · Fax: 202-318-0793 · email: clark@polidata.org
PUBLISHER OF THE POLIDATA ® DEMOGRAPHIC AND POLITICAL GUIDES AND ATLASES

Notes on the American Community Survey Data

NCSL Redistricting Seminar #5; Washington, DC; January 2011

The American Community Survey (ACS) is the replacement for the so-called long form of previous censuses and is the new source for demographic characteristic information. In previous cycles this type of information was not generally available until after most plans had been drawn. Therefore, a threshold consideration for the redistricting cycle of 2011-2012 is what use can be made of the ACS data for plan preparation and review?

There are several reasons why ACS data might be useful for the redistricting community. Perhaps the most important reasons are: a) for the first time, we will be able to see demographic characteristics, other than race and Hispanic origin, that are relatively current for many levels of census geography, including the current districts; and b) the ACS forms the basis for the citizen voting age population data (CVAP) which may be relevant with respect to the Voting Rights Act (VRA).

ACS Data Collection and Release. The ACS program began following the 2000 census and the full-scale data collection has been on a continuous basis since the beginning of the 2005 calendar year. The ACS data do differ from the long-form data in the sense that they are not a snapshot in time but are based upon all persons in the ACS survey universe for selected periods of time. For example, the first release was in 2006: the 2005 1-year release which combined the information from all respondents collected during calendar year 2005.

The release of the 2009 5-year data in December 2010 was the 7th release of this demographic characteristic data collected from an approximate sample of 3 million addresses each year. This is also the first release of data representing a 5-year timeframe (all respondents from the 2005, 2006, 2007, 2008, and 2009 survey universes). As such, it represents the largest sample to date in the short history of the ACS. Because the sample is larger, estimates of the characteristics are reported for many more geographic areas, and summary levels, than any of the previous 1-year or 3-year releases.

The 5-year release is the first to provide characteristics for census tracts and block groups, though not all tables are released for block groups, and the geography is still that from the 2000 census. For these two low levels of census geography (only the census block is lower) the 5-year releases will be the only source of characteristics.

It is also the first release to provide complete coverage for the higher-levels of the census hierarchy, notably, counties. While it is not the first to contain data for congressional districts (CDs), it is the first to contain data for state legislative districts (SLDs). In addition, it is the first release to contain data for all 25,000 Places (cities, towns, and census designated places) and 21,000 minor civil divisions (MCDs) in selected states.

There is an important caveat for all releases through the 2009 collection year: they are generally based upon the 2000 census geography[1] (at least for the lower levels) and are controlled for the population from the 2000 census. For the vast majority of geographic areas at the higher level of geography this will not be much of a problem. At the lower levels, census tracts are designed to be more or less comparable over time but block groups are not.

ACS Release Options. The current plan is that for each subsequent year of the ACS, there will be three types of release: thus this 2009 5-year release completes the rollout of the three types (1-year, 3-year, and 5-year release).

Given the availability of three types of ACS releases through 2009 (the 2009 3-year release was released earlier this month), which of the three does the researcher use? Perhaps the most important considerations are a) the level of geography for which the information is needed; b) when it is needed; c) the level of accuracy required; and d) the currency of the data.

Generally, the 1-year release has been released earlier in the year than the 3-year or 5-year releases[2]. However, the 1-year release, while the most current, has the smallest sample size and is only available for geographic areas that have a base population of 65,000 or more. The 3-year release has a larger sample size but is available for more geographic areas: those with a base population of 20,000 or more. The 5-year release has the largest sample and is available for all geographic areas and most levels of census geography: it also covers the longest period of time in the pooled universe.

[1] "Census tracts and block groups used to tabulate and present 2005-2009 ACS 5-year data are those used for Census 2000 data products. Inadvertently, 26 counties use 2010 Census boundaries for tabulation and presentation of census tracts and block groups in the ACS 5-year data. These census tracts and block groups were included in the version of the Census Bureau's geographic database (TIGER) used to produce geographic area information for the 2005-2009 ACS 5-year data." http://www.census.gov/acs/www/data_documentation/geography_notes/

[2] For example the releases for the 2007 ACS were made available in September (1-year) and October (3-year) of 2008.

The 1-year release has a median currency of approximately 6 months plus the number of months before it is released in the following year; the 3-year release would have a median currency of approximately 1 year and 6 months plus those preceding its release; the 5-year release would have a median currency of approximately 2 years and 6 months plus those preceding its release. For the sake of simplicity, if we assume that all releases were made in December, the 1-year would have a median 'age' of about 17 months (6+11); the 3-year would have a median 'age' of 29 months (6+12+11); and the 5-year release would have a median 'age' of 41 months (6+12+12+11).

If currency isn't the biggest concern and geography is more paramount, your choices are merely what is included in each release. If you want to compare states, you could use any release but to compare all counties, you would need the 5-year release. The choice of which to use is thus a balancing test between factors.

Accuracy of the Data. As with most data collections program of the Bureau of the Census, the ACS data are the result of estimations from survey responses and are thus subject to both sampling error[3] and non-sampling error[4]. Due to the nature of sampling, the point estimate provided in one ACS release may differ greatly from previous, or subsequent, releases. The data releases include the margin of error with each data release[5]. Understanding these is one reason that the technical documentation is an important part of the research. For example, a 5 percentage point increase in a demographic characteristic may, or may not, mean there was an actual increase in the variable for the geographic area of interest. The Bureau documentation provides a discussion of how a comparison of estimates can be tested to determine if the change is statistically significant[6].

[3] "The data in the ACS products are estimates of the actual figures that would have been obtained by interviewing the entire population using the same methodology." Accuracy of the Data (2005)

[4] "For example, operations such as data entry from questionnaires and editing may introduce error into the estimates." Accuracy of the Data (2005)

[5] "Margin of Error – Instead of providing the upper and lower confidence bounds in published ACS tables, the margin of error is provided instead. The margin of error is the difference between an estimate and its upper or lower confidence bound. Both the confidence bounds and the standard error can easily be computed from the margin of error. All ACS published margins of error are based on a 90 percent confidence level. Standard Error = Margin of Error / 1.65. Lower Confidence Bound = Estimate - Margin of Error. Upper Confidence Bound = Estimate + Margin of Error." Accuracy of the Data (2005)

[6] "Significant differences – Users may conduct a statistical test to see if the difference between an ACS estimate and any other chosen estimates is statistically significant at a given confidence level. 'Statistically significant' means that the difference is not likely due to random chance alone." [The only items needed to determine this are the two estimates and the two standard errors (which can be calculated from the margin of error).] "Any estimate can be compared to an ACS estimate using this method, including other ACS estimates from the current year, the ACS estimate for the same characteristic and geographic area but from a previous year, Census 2000 100% counts and long form estimates, estimates from other Census Bureau surveys, and estimates from other sources. Not all estimates have sampling error — Census 2000 100% counts do not, for example, although Census 2000 long form estimates do — but they should be used if they exist to give the most accurate result of the test."Accuracy of the Data (2005)

Means of Access. There are two main means of access to the ACS data.

1) The primary means is interactively via the Census web site, either generically via www.census.gov or directly via factfinder.census.gov. There is also a main page just for the ACS at www.census.gov/acs. This allows for interactive selection of a) dataset (year and period for the survey, e.g., 2009 5-year); b) geography (from the nation down to the lower, but not the lowest, levels of the census hierarchy); and c) subject matter by choosing a single table or multiple tables. After selection the data may be viewed, printed, or downloaded for further use. This is a probably a good way to review customized searches for a handful of tables for a limited set of geographic units.
2) The secondary means is by downloading the raw data files so that they might be imported into a spreadsheet or into a relational database system. This requires a bit more effort but if the researcher only needs a few tables, but for multiple geographic units, this is the preferred option. There are two ways to import the data into readily-available software: a) via the Bureau's new Excel-based import tool[7] or b) by using either statistical or relational database software (such as SPSS or SAS or FoxPro or Oracle). Using Excel is a simple choice and fills the middle of the technical spectrum. Using database software involves both knowledge of that specialized software and considerably more effort to manipulate the data files.

The important differences between these two means are, at least, the following: a) the interactive tool is preferable for ease of use and customized searches but the database download is clearly preferable for experienced users who will need to integrate data for many areas of geography or many subject tables; and b) not all tables are available via the interactive American Fact Finder. Some tables, and some levels of geography are only available by downloading the summary files and this is what will be discussed in the following section.

General Notes on the Summary Files. The ACS database for each release is delivered in formats similar to those used in previous censuses to deliver the Summary (Tape) Files (formerly STF, now SF). These files create a virtual record/row for each summary level of geography with every field/column being a discrete value, e.g., number of males from age 35 to 44. In order to make file manipulation a bit more understandable and to accommodate readily-available legacy database software, the virtual record is broken into separate files of record segments, with each file containing no more than 256 fields/columns horizontally, though there is no limit on the number of records/rows that are in each record segment (aside from the levels of geography available). The tables are more or less arranged by subject matter so some researchers may be lucky

[7] There are actually two Excel-based tools: a retrieval tool and an import tool. The retrieval tool downloads the data files and allows for some minimal options. The import tool provides the 'headers', or field definitions of each raw data file that the researcher has downloaded previously.

enough to only need a few files to cover the appropriate record segments: if this is so, consider using the Excel-import tool first.

The summary files can be downloaded directly from the web via your web browser (e.g., Internet Explorer) or via FTP. These include a geography file that contains the basic information for the geographic area and a relational link (LOGRECNO) into the other record segment files. All files are plain text files and all record segment files are delimited by commas. However, to account for missing data and a few other aspects of the data, not all fields are actually numbers.

The geography file is also a text file but it is a 'flat' non-delimited file and thus requires a data structure indicating the field lengths. Unfortunately, this appears to be available only as a listing in the documentation. Fortunately, the only recent change in the geography structure from previous releases relates to a change in one variable (SUBMCD: the length has increased from 2 to 5) and a few other fields that have been designated as "Reserved or Blank" for now. Each record segment file of the characteristic data has a separate data structure.

Documentation. The Bureau has compiled quite a bit of documentation, ranging from quite technical to more database oriented. For the ACS the Bureau has prepared the so-called Compass guides[8] that give a higher-level focus to the ACS and uses for the data. Appendix material gives a readable summary of the statistical concepts involved without overwhelming the researcher. The technical documentation is useful for both describing the Excel-import tool and the database structure.

The first step is to obtain the technical documentation which is a 20-page document with a 51-page appendix[9]. Aside from the 71-page printed format, there is an Excel version of the data tables and cells included in the database. This is what used to be called the Merge_5_6 file that now has the more understandable name of the Sequence and Table Number Lookup file[10]. N.B., that the file names for some files may not include any designation for the year/period release, e.g., a Merge_5_6 may be provided in each folder.

The Bureau does provide some detailed documentation in the nature of "Product Changes" or so-called crosswalk tables. None have been provided for the 2009 5-year release because it is a new product.

[8] Compass guides: http://www.census.gov/acs/www/guidance_for_data_users/handbooks/

[9] Technical documentation: http://www2.census.gov/acs2009_5yr/summaryfile/. The year/period may be edited for easy access via the URL address window.

[10] Excel table: Sequence_Number_and_Table_Number_Lookup.xls. Note the lack of a year/period indicator in the file name.

Miscellaneous Notes on the Structure of the Summary Files.

1. The files are all text files (i.e., visible in any text editor) and, with the exception of the geography file, are all delimited by a comma between each field/variable. Therefore, they may be easily imported to database software once the structure for each file segment is designated.

2. The file structure can be, with some minor modifications, created from the record layout file provided by the Bureau (the so-called Sequence Number and Table Number Lookup file, hereinafter generally referred to as the Data Dictionary.

3. There are two components to the field name: a TABLEID and an ORDER identifier. The concatenation of these two fields may result in a unique identifier that is longer than 10 characters; such field names are not unique at only 10 characters, as required for some legacy software that include a 10-character uniqueness rule.

4. The TABLEID may consist of the following parts: a) prefix for type of table, i.e., "B" for base table or "C" for collapsed table; b) core table number, e.g., 07101, in which the first two identify the primary subject of the table; c) suffix for race/Hispanic Origin breakouts; and d) an alternative geo-suffix if the table contains responses only for Puerto Rico. (See attachment.)

5. The ORDER identifier is generally blank for the table name or indication of the universe. This field may also include several values with a decimal point: such records do not represent fields and are for documentation purposes only. Most tables contain from a handful of cells to several dozen, though one set of tables (B24121-B24126 relating to OCCUPATION) contains 499 cells in each table.

6. Due to the sampled nature of the ACS, margins of error exist, and are provided for each table, though in a separate data file. The ACS estimates are in the "e" files; the margins are in the "m" files. The old Standard Error information of the "s" files has been eliminated. The format for the "e" and "m" files is the same so either parallel files must be maintained or field names must be modified to merge the data into one dataset.

7. Due to the sampling, the estimation value provided may be non-numeric. These values generally relate to either data that are either missing or suppressed for privacy concerns though a few other reasons are discussed in the documentation.

8. For the inclusion of the data for the levels of Tract and Block Group (only available in the 5-year releases), there are additional raw data files. N.B., these files have the same name as the files for the upper-level geography: be sure to unzip them to a separate folder/directory. They must be appended separately.

9. Some tables are included in the data files even though they contain data only collected for Puerto Rico; such fields are still included in the data structure even though they will be blank for areas other the Puerto Rico. These Puerto Rico-only tables appear in record segments 108 to 117. N.B., that the TABLEID in the Data Dictionary does not always contain the geo-suffix of "PR" (e.g., B05003 in segment 0017 and B05003 in segment 0110).

10. As the virtual record is broken into segments, if there is a problem in a dataset structure, it will only affect the tables in that file segment.

11. The subject tables are of two basic types: "B" and "C". The "B" tables are what most researchers would normally think of: they contain each discrete data cell for a table and may contain over 100 cells. The "C" tables are collapsed versions, e.g., combining several "B" cells into one "C" cell. There are no "A" tables and there need not be a "B" table if there is a "C" table or vice-versa.

12. The tables included in each ACS release may differ, both by the type of release (e.g., 1-year versus 5-year) and because the questions asked on the survey are different: questions may be added or deleted and cells within tables may be modified.

13. For the 2009 5-year release there are 117 file segments, substantially fewer than the 158 provided for the 2008 3-year release and the 153 provided for the 2007 3-year release. While there has been some revision in the tables since that time, the main reason is due the absence of many of the "C" tables from earlier releases. It is unclear whether the researcher can rely upon the existence of a "C" table in one release being available in subsequent releases.

Attachments:

1. Table Numbers Explained

FactFinder Help Census Data Information Glossary Tutorials

Current Topic Table of Contents

American Community Survey (ACS) Table Numbers Explained

An ACS Detailed Table number consists of:

- An initial character which is usually B, but sometimes C.

  B is used for the basic or base tables that provide the most detailed estimates on all topics and for all geographies. These tables are the source for many of the other tables such as Data Profiles, Subject Tables, etc.

  C is used for a collapsed version of a B table. A C table is very similar to a B table with the same number (e.g., C07001 and B07001), but two or more lines from the B table have been collapsed to a single line in the C table. For example, the lines "75 to 79 years", "80 to 84 years" and "85 years and over" from a B table may be collapsed to a single line of "75 years and over" in a C table. Not every B table has a collapsed version.
- The next two characters identify the primary subject of the table.

  01 = Age and Sex
  02 = Race
  03 = Hispanic or Latino Origin
  04 = Ancestry
  05 = Foreign Born; Citizenship; Year of Entry; Nativity
  06 = Place of Birth
  07 = Residence 1 Year Ago; Migration
  08 = Journey to Work; Workers' Characteristics; Commuting
  09 = Children; Household Relationship
  10 = Grandparents; Grandchildren
  11 = Household Type; Family Type; Subfamilies
  12 = Marital Status
  13 = Fertility
  14 = School Enrollment
  15 = Educational Attainment
  16 = Language Spoken at Home and Ability to Speak English
  17 = Poverty
  18 = Disability
  19 = Income (Households and Families)
  20 = Earnings (Individuals)
  21 = Veteran Status
  22 = Food Stamps
  23 = Employment Status; Work Experience
  24 = Industry; Occupation; Class of Worker
  25 = Housing Characteristics
  26 = Group Quarters Population
  27 = Health Insurance Coverage
  98 = Quality Measures
  99 = Imputation table for any subject
- The next three digits are a sequential number, such as 001 or 002, to uniquely identify the table within a given subject.
- For select tables, an alphabetic suffix follows to indicate that a table is repeated for the nine major race and Hispanic or Latino groups:

  A = White Alone
  B = Black or African American Alone
  C = American Indian and Alaska Native Alone
  D = Asian Alone
  E = Native Hawaiian and Other Pacific Islander Alone
  F = Some Other Race Alone
  G = Two or More Races

H = White Alone, Not Hispanic or Latino

I = Hispanic or Latino

- For select tables, a final alphabetic suffix "PR" follows to indicate a table used for Puerto Rico geographies only. These Puerto Rico-specific tables exist because for some geography-based subjects, the wording of the Puerto Rico Community Survey questionnaire differs slightly but significantly from the American Community Survey questionnaire. The matching table used for United States geographies has the same ID but without the trailing "PR" (e.g., B06014 and B06014-PR).