Chapter 13.
Preparation and Review of Data Products

13.1 OVERVIEW

This chapter discusses the data products derived from the American Community Survey (ACS). ACS data products include the tables, reports, and files that contain estimates of population and housing characteristics. These products cover geographic areas within the United States and Puerto Rico. Tools such as the Public Use Microdata Sample (PUMS) files, which enable data users to create their own estimates, also are data products.

ACS data products will continue to meet the traditional needs of those who used the decennial census long-form sample estimates. However, as described in Chapter 14, Section 3, the ACS will provide more current data products than those available from the census long form, an especially important advantage toward the end of a decade.

Most surveys of the population provide sufficient samples to support the release of data products only for the nation, the states, and, possibly, a few substate areas. Because the ACS is a very large survey that collects data continuously in every county, products can be released for many types of geographic areas, including many smaller geographic areas such as counties, townships, and census tracts. For this reason, geography is an important topic for all ACS data products.

The first step in the preparation of a data product is defining the topics and characteristics it will cover. Once the initial characteristics are determined, they must be reviewed by the Census Bureau Disclosure Review Board (DRB) to ensure that individual responses will be kept confidential. Based on this review, the specifications of the products may be revised. The DRB also may require that the microdata files be altered in certain ways, and may restrict the population size of the geographic areas for which these estimates are published. These activities are collectively referred to as disclosure avoidance.

The actual processing of the data products cannot begin until all response records for a given year or years are edited and imputed in the data preparation and processing phases, the final weights are determined, and disclosure avoidance techniques are applied. Using the weights, the sample data are tabulated for a wide variety of characteristics according to the predetermined content. These tabulations are done for the geographic areas that have a sample size sufficient to support statistically reliable estimates, with the exception of 5-year period estimates, which will be available for small geographic areas down to the census tract and block group levels. The PUMS data files are created by different processes because the data are a subset of the full sample data.

After the estimates are produced and verified for correctness, Census Bureau subject matter analysts review them. When the estimates have passed the final review, they are released to the public. A similar process of review and public release is followed for PUMS data.

While the 2005 ACS sample was limited to the housing unit (HU) population for the United States and Puerto Rico, starting in sample year 2006, the ACS was expanded to include the group quarters (GQ) population. Therefore, the ACS sample is representative of the entire resident population in the United States and Puerto Rico. In 2007, 1-year period estimates for the total population and subgroups of the total population in both the United States and Puerto Rico were released for sample year 2006. Similarly, in 2008, 1-year period estimates were released for sample year 2007.

In 2008, the Census Bureau will, for the first time, release products based on 3 years of ACS sample, 2005 through 2007. In 2010, the Census Bureau plans to release the first products based on 5 years of consecutive ACS samples, 2005 through 2009. Since several years of samples form the basis of these multiyear products, reliable estimates can be released for much smaller geographic areas than is possible for products based on single-year data.

In addition to data products regularly released to the public, other data products may be requested by government agencies, private organizations and businesses, or individuals. To accommodate such requests, the Census Bureau operates a custom tabulations program for the ACS on a fee basis. These tabulation requests are reviewed by the DRB to assure protection of confidentiality before release.

Chapter 14 describes the dissemination of the data products discussed in this chapter, including display of products on the Census Bureau's Web site and topics related to data file formatting.

13.2 GEOGRAPHY

The Census Bureau strives to provide products for the geographic areas that are most useful to users of those data. For example, ACS data products are already disseminated for many of the nation's legal and administrative entities, including states, American Indian and Alaska Native (AIAN) areas, counties, minor civil divisions (MCDs), incorporated places, congressional districts, as well as data for a variety of other geographic entities. In cooperation with state and local agencies, the Census Bureau identifies and delineates geographic entities referred to as "statistical areas." These include regions, divisions, urban areas (UAs), census county divisions (CCDs), census designated places (CDPs), census tracts, and block groups. Data users then can select the geographic entity or set of entities that most closely represent their geographic areas of interest and needs.

"Geographic summary level" is a term used by the Census Bureau to designate the different geographic levels or types of geographic areas for which data are summarized. Examples include the entities described above, such as states, counties, and places (the Census Bureau's term for entities such as for cities and towns, including unincorporated areas). Information on the types of geographic areas for which the Census Bureau publishes data is available at <http://www.census.gov/geo/www/garm.html>.

Single-year period estimates of ACS data are published annually for recognized legal, administrative, or statistical areas with populations of 65,000 or more (based on the latest Census Bureau population estimates). Three-year period estimates based on 3 successive years of ACS samples are published for areas of 20,000 or more. If a geographic area met the 1-year or 3-year threshold for a previous period but dropped below it for the current period, it will continue to be published as long as the population does not drop more than 5 percent below the threshold. Plans are to publish 5-year period estimates based on 5 successive years of ACS samples starting in 2010 with the 2005–2009 data. Multiyear period estimates based on 5 successive years of ACS samples will be published for all legal, administrative, and statistical areas down to the block-group level, regardless of population size. However, there are rules from the Census Bureau's DRB that must be applied.

The Puerto Rico Community Survey (PRCS) also provides estimates for legal, administrative, and statistical areas in Puerto Rico. The same rules as described above for the 1-year, 3-year, and 5-year period estimates for the U.S resident population apply for the PRCS as well.

The ACS publishes annual estimates for hundreds of substate areas, many of which will undergo boundary changes due to annexations, detachments, or mergers with other areas.[1] Each year, the Census Bureau's Geography Division, working with state and local governments, updates its files to reflect these boundary changes. Minor corrections to the location of boundaries also can occur as a result of the Census Bureau's ongoing Master Address File (MAF)/Topologically Integrated Geographic Encoding and Referencing (TIGER®) Enhancement Project. The ACS estimates must

[1]The Census Bureau conducts the Boundary and Annexation Survey (BAS) each year. This survey collects information on a voluntary basis from local governments and federally recognized American Indian areas. The information collected includes the correct legal place names, type of government, legal actions that resulted in boundary changes, and up-to-date boundary information. The BAS uses a fixed reference date of January 1 of the BAS year. In years ending in 8, 9, and 0, all incorporated places, all minor civil divisions, and all federally recognized tribal governments are included in the survey. In other years, only governments at or above various population thresholds are contacted. More detailed information on the BAS can be found at <http://www.census .gov/geo/www/bas/bashome.html>.

reflect these legal boundary changes, so all estimates are based on Geography Division files that show the geographic boundaries as they existed on January 1 of the sample year or, in the case of multiyear data products, at the beginning of the final year of data collection.

13.3 DEFINING THE DATA PRODUCTS

For the 1999 through 2002 sample years, the ACS detailed tables were designed to be comparable with Census 2000 Summary File 3 to allow comparisons between data from Census 2000 and the ACS. However, when Census 2000 data users indicated certain changes they wanted in many tables, ACS managers saw the years 2003 and 2004 as opportunities to define ACS products based on users' advice.

Once a preliminary version of the revised suite of products had been developed, the Census Bureau asked for feedback on the planned changes from data users (including other federal agencies) via a *Federal Register* Notice (Fed. Reg. #3510-07-P). The notice requested comments on current and proposed new products, particularly on the basic concept of the product and its usefulness to the data users. Data users provided a wide variety of comments, leading to modifications of planned products.

ACS managers determined the exact form of the new products in time for their use in 2005 for the ACS data release of sample year 2004. This schedule allowed users sufficient time to become familiar with the new products and to provide comments well in advance of the data release for the 2005 sample.

Similarly, a *Federal Register* Notice issued in August 2007 shared with the public plans for the data release schedule and products that would be available beginning in 2008. This notice was the first that described products for multiyear estimates. Improvements will continue when multiyear period estimates are available.

13.4 DESCRIPTION OF AGGREGATED DATA PRODUCTS

ACS data products can be divided into two broad categories: aggregated data products, and the PUMS, which is described in Section 13.5 ("Public Use Microdata Sample").

Data for the ACS are collected from a sample of housing units (HUs), as well as the GQ population, and are used to produce estimates of the actual figures that would have been obtained by interviewing the entire population. The aggregated data products contain the estimates from the survey responses. Each estimate is created using the sample weights from respondent records that meet certain criteria. For example, the 2007 ACS estimate of people under the age of 18 in Chicago is calculated by adding the weights from all respondent records from interviews completed in 2007 in Chicago with residents under 18 years old.

This section provides a description of each aggregated product. Each product described is available as single-year period estimates; unless otherwise indicated, they will be available as 3-year estimates and are planned for the 5-year estimates. Chapter 14 provides more detail on the actual appearance and content of each product.

These data products contain all estimates planned for release each year, including those from multiple years of data, such as the 2005–2007 products. Data release rules will prevent certain single- and 3-year period estimates from being released if they do not meet ACS requirements for statistical reliability.

Detailed Tables

The detailed tables provide basic distributions of characteristics. They are the foundation upon which other data products are built. These tables display estimates and the associated lower and upper bounds of the 90 percent confidence interval. They include demographic, social, economic, and housing characteristics, and provide 1-, 3-, or 5-year period estimates for the nation and the states, as well as for counties, towns, and other small geographic entities, such as census tracts and block groups.

The Census Bureau's goal is to maintain a high degree of comparability between ACS detailed tables and Census 2000 sample-based data products. In addition, characteristics not measured in the Census 2000 tables will be included in the new ACS base tables. The 2007 detailed table products include more than almost 600 tables that cover a wide variety of characteristics, and another 380 race and Hispanic-origin iterations that cover 40 key characteristics. In addition to the tables on characteristics, approximately 80 tables summarize allocation rates from the data edits for many of the characteristics. These provide measures of data quality by showing the extent to which responses to various questionnaire items were complete. Altogether, over 1,300 separate detailed tables are provided.

Data Profiles

Data profiles are high-level reports containing estimates for demographic, social, economic, and housing characteristics. For a given geographic area, the data profiles include distributions for such characteristics as sex, age, type of household, race and Hispanic origin, school enrollment, educational attainment, disability status, veteran status, language spoken at home, ancestry, income, poverty, physical housing characteristics, occupancy and owner/renter status, and housing value. The data profiles include a 90 percent margin of error for each estimate. Beginning with the 2007 ACS, a comparison profile that compares the 2007 sample year's estimates with those of the 2006 ACS also will be published. These profile reports include the results of a statistical significance test for each previous year's estimate, compared to the current year. This test result indicates whether the previous year's estimate is significantly different (at a 90 percent confidence level) from that of the current year.

Narrative Profiles

Narrative profiles cover the current sample year only. These are easy-to-read, computer-produced profiles that describe main topics from the data profiles for the general-purpose user. These are the only ACS products with no standard errors accompanying the estimates.

Subject Tables

These tables are similar to the Census 2000 quick tables, and like them, are derived from detailed tables. Both quick tables and subject tables are predefined, covering frequently requested information on a single topic for a single geographic area. However, subject tables contain more detail than the Census 2000 quick tables or the ACS data profiles. In general, a subject table contains distributions for a few key universes, such as the race groups and people in various age groups, which are relevant to the topic of the table. The estimates for these universes are displayed as whole numbers. The distribution that follows is displayed in percentages. For example, subject table S1501 on educational attainment provides the estimates for two different age groups—18 to 24 years old and 25 years and older, as a whole number. For each age group, these estimates are followed by the percentages of people in different educational attainment categories (high school graduate, college undergraduate degree, etc.). Subject tables also contain other measures, such as medians, and they include the imputation rates for relevant characteristics. More than 40 topic-specific subject tables are released each year.

Ranking Products

Ranking products contain ranked results of many important measures across states. They are produced as 1-year products only, based on the current sample year. The ranked results among the states for each measure are displayed in three ways—charts, tables, and tabular displays that allow for testing statistical significance.

The rankings show approximately 80 selected measures. The data used in ranking products are pulled directly from a detailed table or a data profile for each state.

Geographic Comparison Tables (GCTs)

GCTs contain the same measures that appear in the ranking products. They are produced as both 1-year and multiyear products. GCTs are produced for states as well as for substate entities, such as congressional districts. The results among the geographic entities for each measure are displayed as tables and thematic maps (see next).

Thematic Maps

Thematic maps are similar to ranking tables. They show mapped values for geographic areas at a given geographic summary level. They have the added advantage of visually displaying the geographic variation of key characteristics (referred to as themes). An example of a thematic map would be a map showing the percentage of a population 65 years and older by state.

Selected Population Profiles (SPPs)

SPPs provide certain characteristics from the data profiles for a specific race or ethnic group (e.g., Alaska Natives) or some other selected population group (e.g., people aged 60 years and older). SPPs are provided every year for many of the Census 2000 Summary File 4 iteration groups. SPPs were introduced on a limited basis in the fall of 2005, using the 2004 sample. In 2008 (sample year 2007), this product was significantly expanded. The earlier SPP requirement was that a substate geographic area must have a population of at least 1,000,000 people. This threshold was reduced to 500,000, and congressional districts were added to the list of geographic types that can receive SPPs. Another change to SPPs in 2008 is the addition of many country-of-birth groups.

Groups too small to warrant an SPP for a geographic area based on 1 year of sample data may appear in an SPP based on the 3- or 5-year accumulations of sample data. More details on these profiles can be found in Hillmer (2005), which includes a list of selected race, Hispanic origin, and ancestry populations.

13.5 PUBLIC USE MICRODATA SAMPLE

Microdata are the individual records that contain information collected about each person and HU. PUMS files are extracts from the confidential microdata that avoid disclosure of information about households or individuals. These extracts cover all of the same characteristics contained in the full microdata sample files. Chapter 14 provides information on data and file organization for the PUMS.

The only geography other than state shown on a PUMS file is the Public Use Microdata Area (PUMA). PUMAs are special nonoverlapping areas that partition a state, each containing a population of about 100,000. State governments drew the PUMA boundaries at the time of Census 2000. They were used for the Census 2000 sample PUMS files and are known as the "5 percent PUMAs." (For more information on these geographic areas, go to <http://www.census.gov/prod/cen2000/doc/pums.pdf>.)

The Census Bureau has released a 1-year PUMS file from the ACS since the survey's inception. In addition to the 1-year ACS PUMS file, the Census Bureau plans to create multiyear PUMS files from the ACS sample, starting with the 2005–2007 3-year PUMS file. The multiyear PUMS files combine annual PUMS files to create larger samples in each PUMA, covering a longer period of time. This will allow users to create estimates that are more statistically reliable.

13.6 GENERATION OF DATA PRODUCTS

Following conversations with users of census data, the subject matter analysts in the Census Bureau's Housing and Household Economic Statistics Division and Population Division specify the organization of the ACS data products. These specifications include the logic used to calculate every estimate in each data product and the exact textual description associated with each estimate. Starting with the 2006 ACS data release, only limited changes to these specifications have occurred. Changes to the data product specifications must preserve the ability to compare estimates from one year to another and must be operationally feasible. Changes must be made no later than late winter of each year to ensure that the revised specifications are finalized by the spring of that year and ready for the data releases beginning in the late summer of the year.

After the edited data with the final weights are available (see Chapters 10 and 11), generation of the data products begins with the creation of the detailed tables data products with the 1-year period estimates. The programming teams of the American Community Survey Office (ACSO) generate these estimates. Another staff within ACSO verifies that the estimates comply with the specifications from subject matter analysts. Both the generation and the verification activities are automated.

The 1-year data products are released on a phased schedule starting in the summer. Currently, the Census Bureau plans to release the multiyear data products late each year, after the release of the 1-year products.

One distinguishing feature of the ACS data products system is that standard errors are calculated for all estimates and are released with the latter in tables. Subject matter analysts also use the standard errors in their internal reviews of estimates.

Disclosure Avoidance

Once plans are finalized for the ACS data products, the DRB reviews them to assure that confidentiality of respondents has been protected.

Title 13 of the United States Code (U.S.C.) is the basis for the Census Bureau's policies on disclosure avoidance. Title 13 says, "Neither the Secretary, nor any other officer or employee of the Department of Commerce may make any publication whereby the data furnished by any particular establishment or individual under this title can be identified . . ." The DRB reviews all data products planned for public release to ensure adherence to Title 13 requirements, and may insist on applying disclosure avoidance rules that could result in the suppression of certain measures for small geographic areas. (More information about the DRB and its policies can be found at <http://www.factfinder.census.gov/jsp/saff/SAFFInfo.jsp?_pageId=su5_confidentiality>.

To satisfy Title 13 U.S.C., the Census Bureau uses several statistical methodologies during tabulation and data review to ensure that individually identifiable data will not be released.

Swapping. The main procedure used for protecting Census 2000 tabulations was data swapping. It was applied to both short-form (100 percent) and long-form (sample) data independently. Currently, it also is used to protect ACS tabulations. In each case, a small percentage of household records is swapped. Pairs of households in different geographic regions are swapped. The selection process for deciding which households should be swapped is highly targeted to affect the records with the most disclosure risk. Pairs of households that are swapped match on a minimal set of demographic variables. All data products (tables and microdata) are created from the swapped data files.

For PUMS data the following techniques are employed in addition to swapping:

Top-coding is a method of disclosure avoidance in which all cases in or above a certain percentage of the distribution are placed into a single category.

Geographic population thresholds prohibit the disclosure of data for individuals or HUs for geographic units with population counts below a specified level.

Age perturbation (modifying the age of household members) is required for large households containing 10 people or more due to concerns about confidentiality.

Detail for categorical variables is collapsed if the number of occurrences in each category does not meet a specified national minimum threshold.

For more information on disclosure avoidance techniques, see Section 5, "Current disclosure avoidance practices" at <http://www.census.gov/srd/papers/pdf/rrs2005-06.pdf>.

The DRB also may determine that certain tables are so detailed that other restrictions are required to ensure that there is sufficient sample to avoid revealing information on individual respondents. In such instances, a restriction may be placed on the size of the geographic area for which the table can be published. Current DRB rules require that detailed tables containing more than 100 detailed cells may not be released below the census tract level.

The data products released in the summer of 2006 for the 2005 sample covered the HU population of the United States and Puerto Rico only. In January 2006, data collection began for the population living in GQ facilities. Thus, the data products released in summer 2007 (and each year

thereafter) covered the entire resident population of the United States and Puerto Rico. Most estimates for person characteristics covered in the data products were affected by this expansion. For the most part, the actual characteristics remained the same, and only the description of the population group changed from HU to resident population.

Data Release Rules

Even with the population size thresholds described earlier, in certain geographic areas some very detailed tables might include estimates with unacceptable reliability. Data release rules, based on the statistical reliability of the survey estimates, were first applied in the 2005 ACS. These release rules apply only to the 1- and 3-year data products.

The main data release rule for the ACS tables works as follows. Every detailed table consists of a series of estimates. Each estimate is subject to sampling variability that can be summarized by its standard error. If more than half of the estimates in the table are not statistically different from 0 (at a 90 percent confidence level), then the table fails. Dividing the standard error by the estimate yields the coefficient of variation (CV) for each estimate. (If the estimate is 0, a CV of 100 percent is assigned.) To implement this requirement for each table at a given geographic area, CVs are calculated for each table's estimates, and the median CV value is determined. If the median CV value for the table is less than or equal to 61 percent, the table passes for that geographic area and is published; if it is greater than 61 percent, the table fails and is not published.

Whenever a table fails, a simpler table that collapses some of the detailed lines together can be substituted for the original. If the simpler table passes, it is released. If it fails, none of the estimates for that table and geographic area are released. These release rules are applied to single- and multiyear period estimates based on 3 years of sample data. Current plans are not to apply data release rules to the estimates based on 5 years of sample data.

13.7 DATA REVIEW AND ACCEPTANCE

After the editing, imputation, data products generation, disclosure avoidance, and application of the release rules have been completed, subject matter analysts perform a final review of the ACS data and estimates before release. This final data review and acceptance process helps to ensure that there are no missing values, obvious errors, or other data anomalies.

Each year, the ACS staff and subject matter analysts generate, review, and provide clearance of all ACS estimates. At a minimum, the analysts subject their data to a specific multistep review process before they are cleared and released to the public. Because of the short time available to review such a large amount of data, an automated review tool (ART) has been developed to facilitate the process.

ART is a computer application that enables subject matter analysts to detect statistically significant differences in estimates from one year to the next using several statistical tests. The initial version of ART was used to review 2003 and 2004 data. It featured predesigned reports as well as ad hoc, user-defined queries for hundreds of estimates and for 350 geographic areas. An ART workgroup defined a new version of ART to address several issues that emerged. The improved version has been used by the analysts since June 2005; it is designed to work on much larger data sets and a wider range of capabilities, with faster response time to user commands. A team of programmers, analysts, and statisticians then developed an automated tool to assist analysts in their review of the multiyear estimates. This tool was used in 2008 for the review of the 2005–2007 estimates.

The ACSO staff, together with the subject matter analysts, also have developed two other automated tools to facilitate documentation and clearance for required data review process steps: the edit management and messaging application (EMMA), and the PUMS management and messaging application (PMMA). Both are used to track the progress of analysts' review activities and both enable analysts and managers to see the current status of files under review and determine which review steps can be initiated.

13.8 IMPORTANT NOTES ON MULTIYEAR ESTIMATES

While the types of data products for the multiyear estimates are almost entirely identical to those used for the 1-year estimates, there are several distinctive features of the multiyear estimates that data users must bear in mind.

First, the geographic boundaries that are used for multiyear estimates are always the boundary as of January 1 of the final year of the period. Therefore, if a geographic area has gained or lost territory during the multiyear period, this practice can have a bearing on the user's interpretation of the estimates for that geographic area.

Secondly, for multiyear period estimates based on monetary characteristics (for example, median earnings), inflation factors are applied to the data to create estimates that reflect the dollar values in the final year of the multiyear period.

Finally, although the Census Bureau tries to minimize the changes to the ACS questionnaire, these changes will occur from time to time. Changes to a question can result in the inability to build certain estimates for a multiyear period containing the year in which the question was changed. In addition, if a new question is introduced during the multiyear period, it may be impossible to make estimates of characteristics related to the new question for the multiyear period.

13.9 CUSTOM DATA PRODUCTS

The Census Bureau offers a wide variety of general-purpose data products from the ACS designed to meet the needs of the majority of data users. They contain predefined sets of data for standard census geographic areas. For users whose data needs are not met by the general-purpose products, the Census Bureau offers customized special tabulations on a cost-reimbursable basis through the ACS custom tabulation program. Custom tabulations are created by tabulating data from ACS edited and weighted data files. These projects vary in size, complexity, and cost, depending on the needs of the sponsoring client.

Each custom tabulation request is reviewed in advance by the DRB to ensure that confidentiality is protected. The requestor may be required to modify the original request to meet disclosure avoidance requirements. For more detailed information on the ACS Custom Tabulations program, go to <http://www.census.gov/acs/www/Products/spec_tabs/index.htm>.