Understanding and Using the American Community Survey Public Use Microdata Sample Files

What Data Users Need to Know

Issued February 2021

United States® Census Bureau

U.S. Department of Commerce
U.S. CENSUS BUREAU
census.gov

Acknowledgments

Linda A. Jacobsen, Vice President, U.S. Programs, Population Reference Bureau (PRB), and **Mark Mather**, Associate Vice President, U.S. Programs, PRB, drafted this handbook in partnership with the U.S. Census Bureau's American Community Survey Office. Other PRB staff who assisted in drafting and reviewing the handbook include: **Jean D'Amico**, **Lillian Kilduff**, **Kelvin Pollard**, **Paola Scommegna**, and **Alicia VanOrman**. Some of the material in this handbook was adapted from the Census Bureau's 2009 publication, *A Compass for Understanding and Using American Community Survey Data: What PUMS Data Users Need to Know*, drafted by **Leonard M. Gaines**.

Nicole Scanniello, **Gretchen Gooding**, and **Amanda Klimek**, Census Bureau, contributed to the planning and review of this handbook.

The American Community Survey program is under the direction of **Albert E. Fontenot Jr.**, Associate Director for Decennial Census Programs, **Deborah M. Stempowski**, Assistant Director for Decennial Census Programs, and **Donna M. Daily**, Chief, American Community Survey Office.

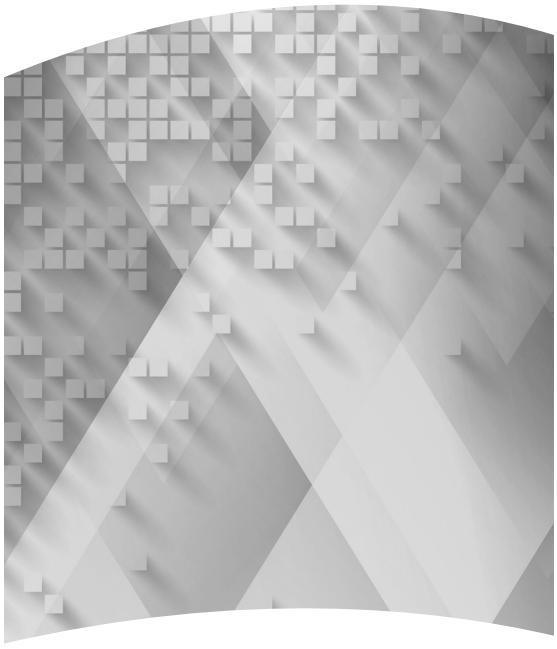
Other individuals from the Census Bureau who contributed to the review and release of these handbooks include **Lydia Anderson**, **Aaron Basler**, **Sirius Fuller**, **William Hazard**, **KaNin Reese**, **Camille Ryan**, **Janice Valdisera**, **Tyson Weister**, and **Kai Wu.**

Faye E. Brock, **Linda Chen**, and **Christine E. Geter** provided publication management, graphic design and composition, and editorial review for the print and electronic media under the direction of **Corey Beasley**, Acting Chief of the Graphic and Editorial Services Branch, Public Information Office.

Understanding and Using the American Community Survey Public Use Microdata Sample Files

What Data Users Need to Know

Issued February 2021



U.S. Department of Commerce
Wynn Coggins,
Acting Agency Head

U.S. CENSUS BUREAU
Dr. Ron Jarmin,
Acting Director

Suggested Citation

U.S. Census Bureau,
Understanding and Using the American Community Survey Public Use Microdata Sample Files: What Data Users Need to Know,
U.S. Government Printing Office,
Washington, DC, 2021.

U.S. CENSUS BUREAU
Dr. Ron Jarmin,
Acting Director

Dr. Ron Jarmin,
Deputy Director and Chief Operating Officer

Albert E. Fontenot Jr.,
Associate Director for Decennial Census Programs

Deborah M. Stempowski,
Assistant Director for Decennial Census Programs

Donna M. Daily,
Chief, American Community Survey Office

Contents

This page is intentionally blank.

UNDERSTANDING AND USING THE AMERICAN COMMUNITY SURVEY PUBLIC USE MICRODATA SAMPLE FILES: WHAT DATA USERS NEED TO KNOW

The U.S. Census Bureau produces a large number of data profiles, tables, maps, and other products based on American Community Survey (ACS) data. Even this abundance of pretabulated estimates and data products cannot meet the needs of every data user. The Census Bureau's ACS Public Use Microdata Sample (PUMS) files enable data users to create custom estimates and tables, free of charge, that are not available through pretabulated ACS data products.

This guide provides an overview of the ACS PUMS files and how they can be used to access data about America's communities.

What Is the ACS?

The ACS is a nationwide survey designed to provide communities with reliable and timely social, economic, housing, and demographic data every year. A separate annual survey, called the Puerto Rico Community Survey (PRCS), collects similar data about the population and housing units in Puerto Rico. The Census Bureau uses data collected in the ACS and the PRCS to provide estimates on a broad range of population, housing unit, and household characteristics for states, counties, cities, school districts, congressional districts, census tracts, block groups, and many other geographic areas.

The ACS has an annual sample size of about 3.5 million addresses, with survey information collected nearly every day of the year. Data are pooled across a calendar year to produce estimates for that year. As a result, ACS estimates reflect data that have been collected over a period of time rather than for a single point in time as in the decennial census, which is conducted every 10 years and provides population counts as of April 1 of the census year.

ACS 1-year estimates are data that have been collected over a 12-month period and are available for geographic areas with at least 65,000 people. Starting with the 2014 ACS, the Census Bureau is also producing "1-year Supplemental Estimates"—simplified versions of popular ACS tables—for geographic areas with at least 20,000 people. The Census Bureau combines 5 consecutive years of ACS data to produce multiyear estimates for geographic areas with fewer than 65,000 residents. These 5-year estimates represent data collected over a period of 60 months.

For more detailed information about the ACS—how to judge the accuracy of ACS estimates, understanding multiyear estimates, knowing which geographic areas are covered in the ACS, and how to access ACS data on the Census Bureau's Web site—see the Census Bureau's handbook on *Understanding and Using American Community Survey Data: What All Data Users Need to Know.*[1]

[1] U.S. Census Bureau, U*nderstanding and Using American Community Survey Data: What All Data Users Need to Know*, <www.census.gov/programs-surveys/acs/guidance/handbooks/general.html>.

1. ACS PUMS FILES: THE BASICS

The American Community Survey (ACS) microdata consist of individual records with information about the characteristics of each person and housing unit in the survey. The ACS Public Use Microdata Sample (PUMS) includes a subsample of the ACS microdata, devoid of personalized information. The PUMS represents about two-thirds of the responses collected in the ACS in a specific 1-year or 5-year period.

The U.S. Census Bureau produces ACS 1-year and 5-year PUMS files and typically releases these files 1 month after the release of the published ACS tables. The 5-year PUMS file is a combination of five 1-year PUMS files. The 1-year PUMS file includes records for about 1 percent of the total population and the 5-year file includes records for about 5 percent of the total population.

TIP: While PUMS data allow for more detailed and complex research techniques, the files are more difficult to work with than published tables. Data users need to use statistical software, such as SPSS, SAS, R, or Stata, to process PUMS data, and the responsibility for producing estimates from PUMS and judging their statistical significance is up to the data user.

There are two types of PUMS files, one for persons and one for housing units. The person-level file includes records for people, including those who live in group quarter facilities such as nursing homes or college dorms. The housing-level files include records pertaining to housing units, including vacant units.

The ACS PUMS is a weighted sample, and weighting variables must be used to generate estimates and standard errors that represent the population. The PUMS files include both population weights and household weights. Population weights should be used to generate statistics about individuals, and household weights should be used to generate statistics about housing units. (See the section on "Preparing ACS PUMS Data Files for Analysis" for more information.)

Protecting Confidentiality in the ACS PUMS

Title 13 legally requires the Census Bureau to keep all personal information strictly confidential.[2] Examples of measures taken to protect confidentiality in the PUMS include:

- Using only a subset of the full ACS sample to create the PUMS.
- Excluding names, addresses, and any information that could be used to identify a specific housing unit, group quarters unit, or person.
- "Swapping" (or exchanging) a small number of records with similar records from neighboring areas.
- Top-coding or bottom-coding answers to selected variables. Top-coding and bottom-coding involves truncating extreme values for certain variables. A list of top-coded and bottom-coded values is available on the Census Bureau's PUMS Documentation Web page.[3]
- Limiting the geographic areas that can be identified in the PUMS. Data are available for the nation, regions, divisions, states, and Public Use Microdata Areas (PUMAs). The section on "Public Use Microdata Areas" provides more information.

TIP: Due to confidentiality protections and the fact that PUMS files are based on only about two-thirds of the ACS sample, estimates using the ACS PUMS may differ from estimates provided through the ACS Summary File or other published Census Bureau tables and profiles. You can verify that you have correctly accessed and tabulated data from the ACS PUMS file by replicating the values presented in "PUMS Estimates for User Verification" in the PUMS Technical Documentation.[4]

[2] U.S. Census Bureau, Data Protection and Privacy, Title 13 - Protection of Confidential Information, <www.census.gov/about/policies/privacy/data_stewardship/title_13_-_protection_of_confidential_information.html>.

[3] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

[4] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

Summary Data Versus Microdata

Summary data (also called aggregate or pretabulated data) are predefined weighted tabulations of person or housing characteristics. ACS summary tabulations are typically presented in tables, data profiles, and maps. Figure 1.1 provides a portion of a summary table on the means of transportation to work using the ACS 5-year estimates. In 2012–2016, 24,183 workers in Pittsburgh, Pennsylvania, got to work by bus or trolley.

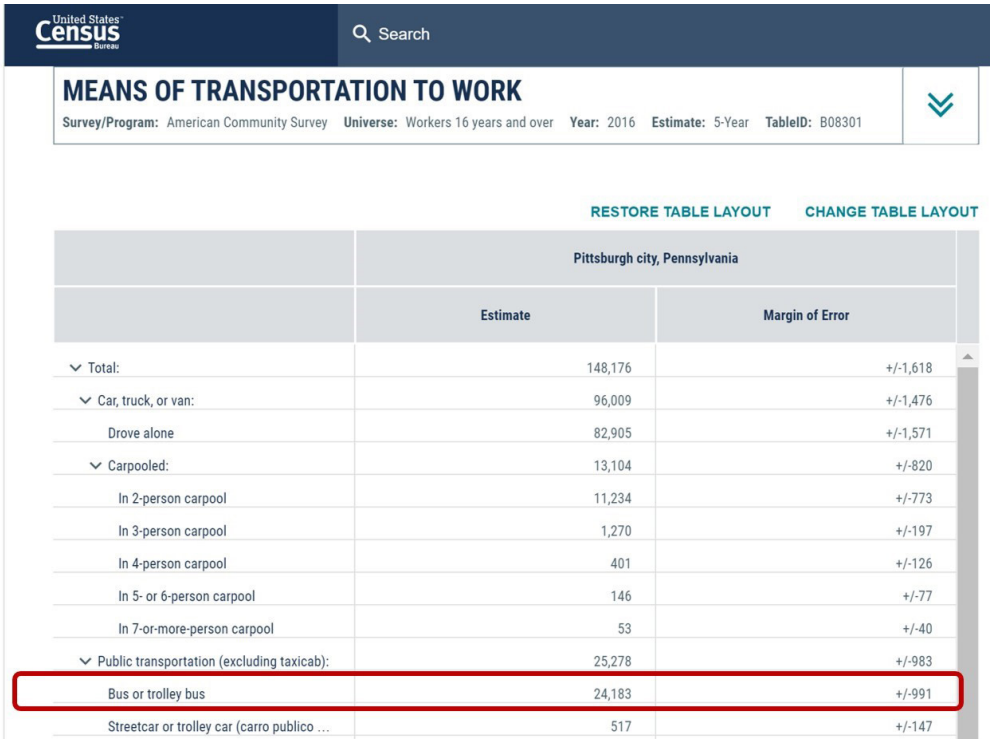
The basic unit of analysis for summary data is a specific geographic entity such as a state, county, or place (including cities and towns). Estimates in summary tables can be added (or subtracted), but users are constrained by the categories that are available.

The benefit of using summary data is that estimates of population and housing characteristics are provided for geographic areas as small as census tracts (small subdivisions of counties that typically have between 1,200 and 8,000 residents) and block groups (subdivisions of census tracts that typically have between 600 and 3,000 residents). Summary data in published tables are user-friendly because they are produced by Census Bureau analysts, informed by data user input, and usually include margins of error. A drawback to summary data is that the user is bound to a predetermined set of tabulations with fixed variable categories.

By contrast, the PUMS contains individual responses to the full range of topics on the ACS (without individually identifiable information). Table 1.1 provides an example of ACS 5-year PUMS microdata for a person in the state of Pennsylvania. The 2012–2016 ACS 5-year PUMS Data Dictionary provides information about each of the values in the table.[5]

[5] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

Figure 1.1. **Published Summary Data for Means of Transportation to Work in Pittsburgh, PA: 2012–2016**



	Pittsburgh city, Pennsylvania	
	Estimate	Margin of Error
Total:	148,176	+/-1,618
Car, truck, or van:	96,009	+/-1,476
Drove alone	82,905	+/-1,571
Carpooled:	13,104	+/-820
In 2-person carpool	11,234	+/-773
In 3-person carpool	1,270	+/-197
In 4-person carpool	401	+/-126
In 5- or 6-person carpool	146	+/-77
In 7-or-more-person carpool	53	+/-40
Public transportation (excluding taxicab):	25,278	+/-983
Bus or trolley bus	24,183	+/-991
Streetcar or trolley car (carro publico ...	517	+/-147

Source: U.S. Census Bureau, <https://data.census.gov>, Table B08301: Means of Transportation to Work.

Table 1.1. **Example of Microdata**

RT	SERIALNO	SPORDER	ST	AGEP	JWTR	SEX
P	2016000009344	02	42	49	.	2
P	2016000009344	03	42	15	.	2
P	2016000009578	02	42	27	01	1
P	2016000009578	03	42	28	01	1
P	2016000009578	04	42	29	01	1
P	2016000009578	05	42	25	01	1
P	2016000021254	02	42	25	10	2
P	2016000024874	02	42	72	.	2
P	2016000025941	02	42	22	02	2
P	2016000025941	03	42	14	.	1
P	2016000025941	04	42	71	.	2

Source: U.S. Census Bureau, 2012–2016 American Community Survey Estimates, 5-Year Public Use Microdata Sample File.

In the highlighted row in Table 1.1:

- The variable for Record Type (RT) is "P," indicating that this record comes from the PUMS person file.
- The SERIALNO ("2016000025941") is a unique identifier for the housing unit.
- SPORDER ("02") is a unique identifier of persons within a housing unit.
- ST is "42," indicating the individual lives in Pennsylvania.
- AGEP is "22," indicating the person was 22 years old at the time of the survey.
- JWTR, or Means of Transportation to Work, is "02" indicating the person commutes to work by bus or trolley.
- SEX is "2" indicating the person is female.

You can find values for all PUMS variables in the PUMS Data Dictionary.[6] Here, missing values are shown as periods. A missing numeric value or a blank character value may be displayed differently depending on the software used. A missing value indicates that the person recode was not in the universe for the relevant variable. For example, as shown in the data dictionary, JWTR values are not available for respondents who are not workers.

There are approximately 250 person-level variables and 200 housing-level variables available in the ACS PUMS. The PUMS files permit data users to analyze specific population groups and create custom variables that are not available through published tables. PUMS files allow more flexibility because data users can focus in on small population groups as well as investigate the relationship among survey questionnaire items by defining and creating unique combinations of person or household variables. Some examples include:

- Estimating the population living below a specified income-to-poverty ratio (for example, children living in families with income below 185 percent of the poverty threshold).
- Investigating the income and poverty status of Gulf War veterans.
- Comparing poverty and unemployment estimates for women and men working in different occupational categories.
- Tracking trends in state-to-state migration among baby boomers since the Great Recession.
- Analyzing data for more detailed languages than those available through published tables in <https://data.census.gov>.

[6] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

2. PUBLIC USE MICRODATA AREAS

The nation, regions, divisions, states, and Public Use Microdata Areas, or PUMAs, are the only geographic areas identified in the American Community Survey (ACS) Public Use Microdata Sample (PUMS). PUMAs are geographic areas defined specifically for the dissemination of PUMS data from the decennial census, ACS, and Puerto Rico Community Survey.

PUMAs were first delineated for the 1990 Census by the state data centers (SDCs) in cooperation with regional, state, local, and tribal organizations and agencies. They are redrawn following each decennial census. This means that PUMA identifiers change each decade with varying comparability across time.

In the PUMS files, PUMAs are identified by a five-digit code. However, the five-digit codes are not unique across states. State identifiers must be used in conjunction with PUMA codes when working with data for multiple states. For example, data users can combine the state FIPS code for California (06) with PUMA code 07701 to extract records for the San Joaquin County (Central)—Stockton City (North) PUMA. A complete list of 2010 PUMA codes and descriptions is available on the U.S. Census Bureau's Web site.[7]

PUMA boundaries are defined using three main criteria:

1. Each PUMA must have a population of 100,000 or more at the time of delineation, and that population threshold must be maintained throughout the decade. Areas that are experiencing substantial population decline at the time of delineation (or where decline is anticipated) are delineated to include a population greater than 100,000 persons. If the population falls substantially below 100,000 in a given PUMA, the Census Bureau may combine that PUMA with one or more adjacent PUMAs to ensure data confidentiality.

2. PUMAs are based only on aggregations of counties and census tracts and cannot cross state boundaries.

3. The building blocks for PUMAs must be contiguous—or share a common border—unless the features of the counties or census tracts used as building blocks are noncontiguous (for example, islands). As long as population criteria are met, one county may be designated as a PUMA. Contiguous census tracts may be aggregated to create a PUMA, as can two or more contiguous counties. Tract-based PUMAs may cross county boundaries, provided each PUMA-county part meets a minimum population of 2,400.

In addition to the required criteria, the Census Bureau strongly encourages the SDCs to incorporate the following guidelines in their PUMA definitions:

- Wherever possible, each PUMA should comprise an area either entirely inside or entirely outside metropolitan or micropolitan statistical areas.

- 2010 place definitions, 2010 urban/rural definitions, and local knowledge should inform PUMA delineations.

- PUMAs should not contain more than 200,000 people, unless identified as an area that is likely to undergo substantial population decline over the decade.

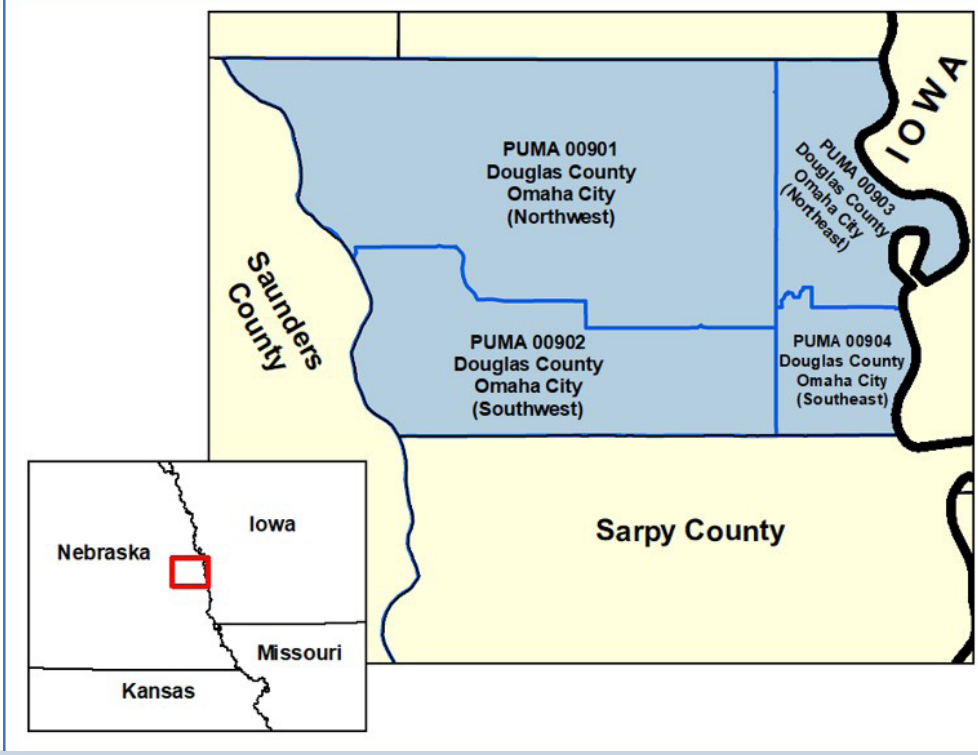
- PUMAs should avoid unnecessarily splitting American Indian reservations (AIRs) and/or off-reservation trust lands (ORTLs), and separating American Indian populations particularly if large numbers of American Indians are included within all parts of the split AIRs/ORTLs.

Given the various criteria and guidelines for defining PUMA boundaries, counties with large populations typically are subdivided into multiple PUMAs, while PUMAs in more rural areas usually comprise two or more adjacent counties.[8] Figure 2.1 shows four PUMAs located within Douglas County, Nebraska. In contrast, Figure 2.2 shows a single PUMA in Alabama (PUMA 00100) that spans four counties. Three of these counties (Colbert, Franklin, and Lauderdale) are fully contained within the PUMA, while one county (Marion) is only partially contained within the PUMA.

[7] U.S. Census Bureau, 2010 PUMA Names, <https://www2.census.gov/geo/pdfs/reference/puma/2010_PUMA_Names.pdf>.

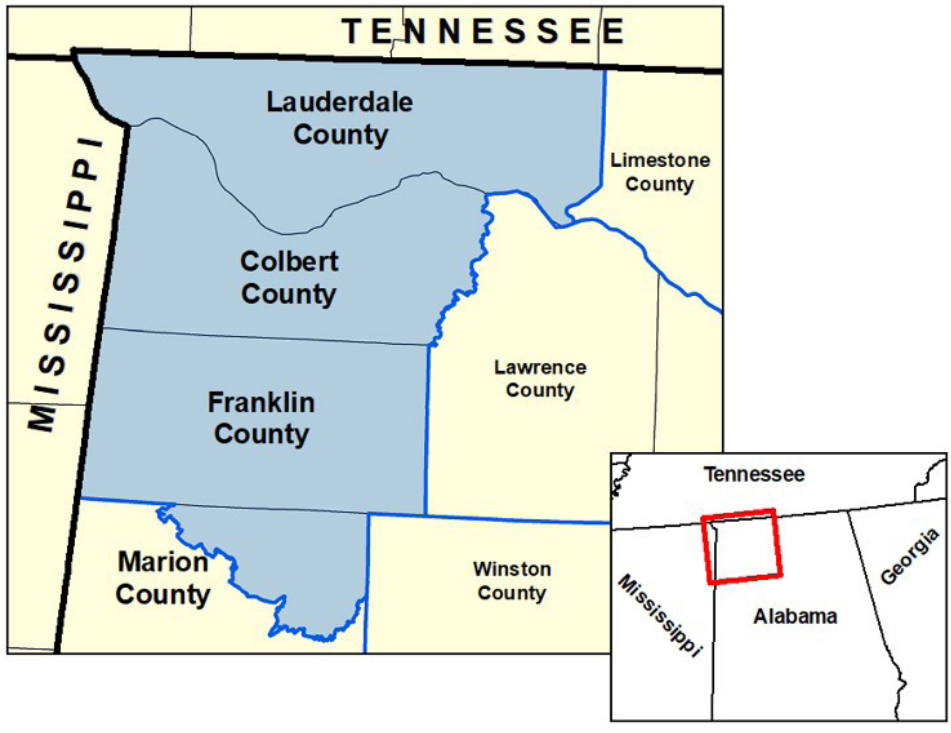
[8] For more information, see the Census Bureau's Final Public Use Microdata Area (PUMA) Criteria and Guidelines for the 2010 Census and the American Community Survey, <https://www2.census.gov/geo/pdfs/reference/puma/2010_puma_guidelines.pdf>.

Figure 2.1. **PUMAs in Douglas County, Nebraska**



Source: U.S. Census Bureau.

Figure 2.2. **Alabama PUMA 00100 (Lauderdale, Colbert, Franklin, and Northeast Marion Counties)**



Source: U.S. Census Bureau.

Place of Work and Migration PUMAs

Place of work (POW) PUMAs and Migration (MIG) PUMAs are used in the publication of ACS PUMS files to provide data on place of work, in- and out-migration flows, and demographic characteristics of workers and migrants. POWPUMAs identify the location of a respondent's primary place of work, while MIGPUMAs identify a respondent's place of residence 1 year ago.

POWPUMAs and MIGPUMAs follow the same sets of boundaries and codes, but are not always aligned with standard PUMAs. Because POWPUMAs/MIGPUMAs are county-based, they may contain multiple standard PUMAs that have been aggregated together to create larger areas in order to protect the confidentiality of respondents. In these instances, the POWPUMAs/MIGPUMAs are assigned unique codes that do not match the codes of any standard PUMAs.

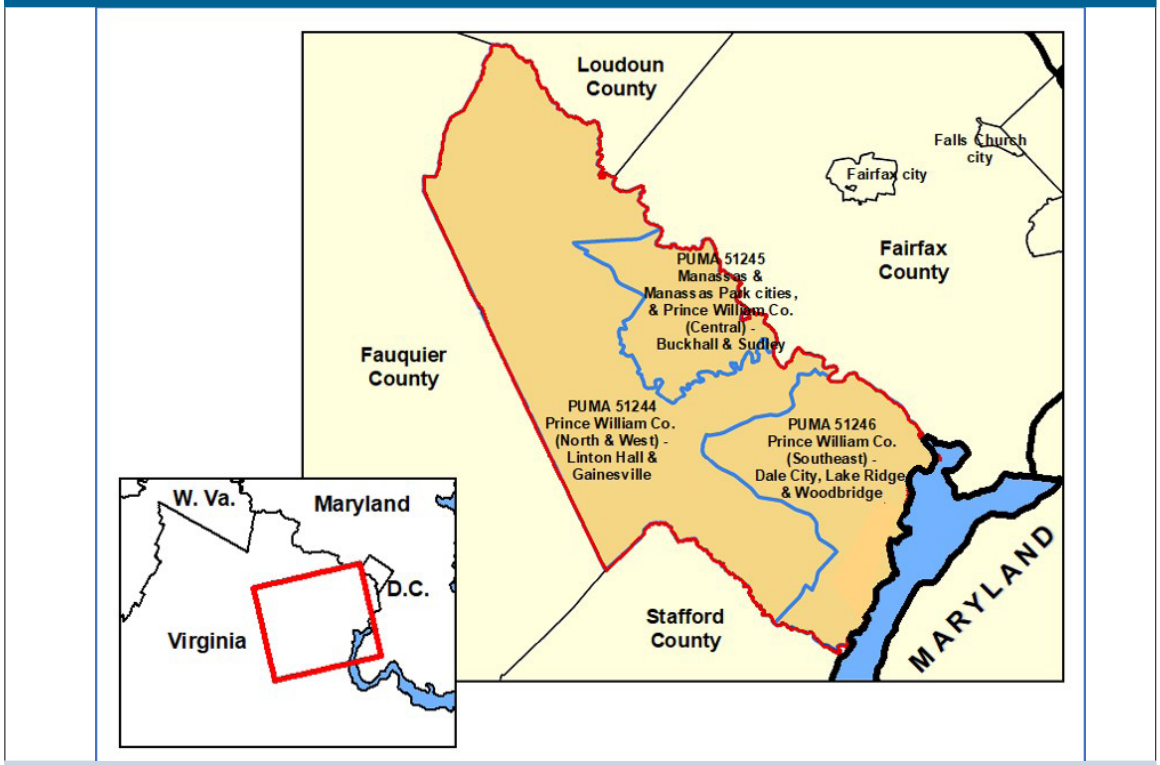
For example:

- POWPUMA/MIGPUMA 51115 in Virginia represents an area that covers Stafford County and is exactly aligned with Virginia PUMA 51115.
- POWPUMA/MIGPUMA 51256 in Virginia represents an area that completely covers Manassas and Prince William County. It is an aggregate of Virginia PUMAs 51244, 51245, and 51246 (see Figure 2.3).

Values and value labels for the POWPUMA and MIGPUMA variables are maintained in an Excel file under the "Code Lists" heading on the PUMS Documentation Web page.[9]

[9] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

Figure 2.3. **Virginia POWPUMA/MIGPUMA 51256 (Manassas and Prince William Counties)**



Note: POWPUMA/MIGPUMA 51256 is the area shaded orange in the figure and fully contains PUMAs 51244, 51245, and 51246.
Source: U.S. Census Bureau.

Changes to PUMA Boundaries

The Census Bureau redraws PUMA boundaries every 10 years based on new population data from the decennial census. The 2012 ACS data files were the first to include PUMAs defined using the 2010 Census data. ACS data files from 2011 and earlier years used the PUMAs defined after the 2000 Census. This means that the ACS 5-year PUMS files from 2008–2012 through 2011–2015 include a mix of PUMAs that were drawn after the 2000 and 2010 Censuses. The records from data years 2008 through 2011 still carry the older 2000-based PUMA codes, while the records from 2012 and later years display the newer 2010-based PUMA geography.

Guidance for working with these dual-vintage PUMA codes is available in the ACS 5-year PUMS Documentation ("PUMS ReadMe" files) for 2009–2013 through 2011–2015.[10]

Data users can also use PUMA Maps and equivalence files, described below, to visualize PUMA boundaries over time or calculate the proportion of a population from a 2010-based PUMA that lies within a 2000-based PUMA.

PUMA Maps and Equivalency Files

There are several tools available to help users understand and visualize PUMA definitions and boundaries.

TIGERweb is a Web-based application that provides a simple way to visualize TIGER (Topologically Integrated Geographic Encoding and Referencing database) data for PUMAs and other geographic areas.[11] For example, ACS data users can use this tool to see how 2000- and 2010-based PUMAs nest within or across counties and cities.

The 2010 Census PUMA Reference Maps can be used to display maps for each of the 2,378 2010-based PUMAs.[12] In addition to the boundaries and codes for PUMAs, these maps provide the names of federal AIRs, ORTLs, states and state equivalents, counties and county equivalents, county subdivisions (in states where they function as governmental units), places, and census tracts.

The 2010 PUMA Equivalency Files show the relationship between 2010 PUMAs and 2010 counties, and some standard 2010 Census geographic entities such as governmental minor civil divisions, places, and census tracts.[13] Because 2010 PUMAs nest within states, there is a separate geographic equivalency file for each state.

The Tract to PUMA Relationship File allows users to identify which census tracts are contained within each PUMA.[14]

The Missouri Census Data Center's Geocorr 2018 Geographic Correspondence Engine enables users to produce a geographic correspondence file, or crosswalk, between PUMAs and dozens of geographic layers.[15] The database includes both 2000- and 2010-based PUMAs. Geocorr 2018 instructs the user to identify a source geography (such as PUMAs) and a target geography (for example, counties). The user is provided an allocation factor, which indicates the portion of the source area that falls within the target geography.

[10] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

[11] U.S. Census Bureau, TIGERweb, <https://tigerweb.geo.census.gov/tigerwebmain/TIGERweb_apps.html>.

[12] U.S. Census Bureau, Geography Reference Maps, <www.census.gov/geo/maps-data/maps/reference.html>.

[13] U.S. Census Bureau, Geography Program, Public Use Microdata Areas (PUMAs), <www.census.gov/programs-surveys/geography/guidance/geo-areas/pumas.html>.

[14] U.S. Census Bureau, Geographies, Relationship Files, <www.census.gov/geographies/reference-files/2010/geo/relationship-files.html>.

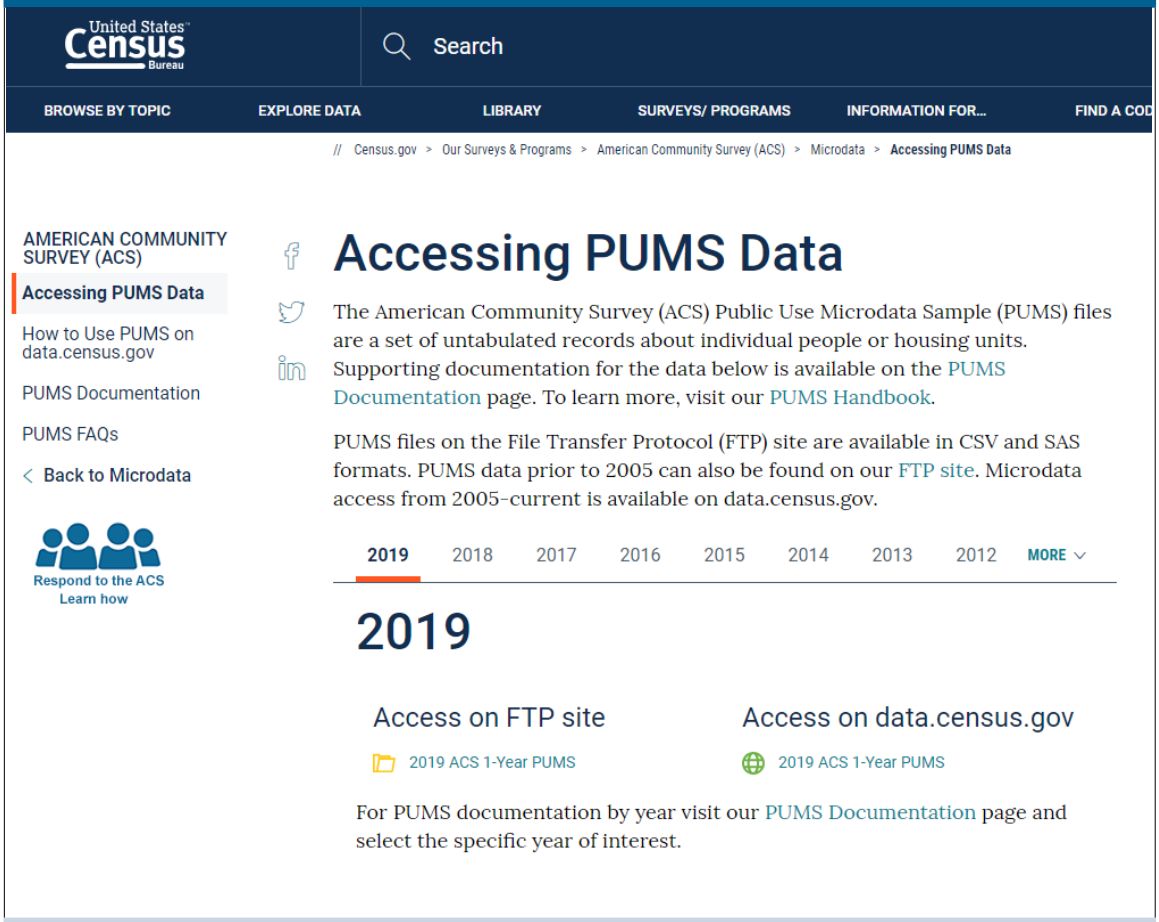
[15] Missouri Census Data Center, Geocorr 2018: Geographic Correspondence Engine, <http://mcdc.missouri.edu/applications/geocorr2018.html>.

3. ACCESSING ACS PUMS DATA

Data users can access the American Community Survey (ACS) Public Use Microdata Sample (PUMS) files through the U.S. Census Bureau's "Accessing PUMS Data" Web page (see Figure 3.1).[16] First, select a year of interest in the year tabs. From there, you can click on the link under "Access on FTP site" to navigate to the PUMS files on the Census Bureau's FTP site. You can also click on a link under "Access on data.census.gov" to access the PUMS data files using the microdata access tool on data.census.gov, the Census Bureau's dissemination tool.

[16] U.S. Census Bureau, American Community Survey (ACS), Accessing PUMS Data, <www.census.gov/programs-surveys/acs/microdata/access.html>.

Figure 3.1. **Accessing ACS PUMS Data**



Source: U.S. Census Bureau, American Community Survey (ACS), Accessing PUMS Data, <www.census.gov/programs-surveys/acs/microdata/access.html>.

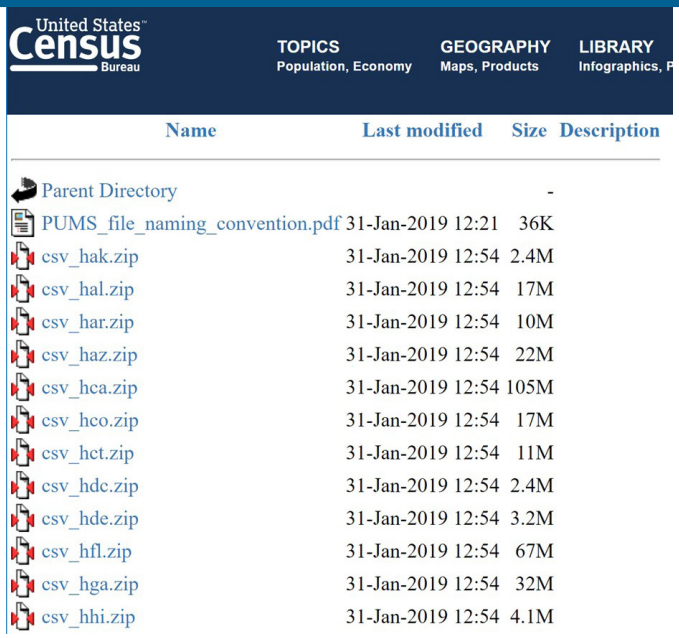
From the FTP site, you have already selected your desired year and data set of interest (see Figure 3.1).[17] ACS 1-year PUMS data sets for 2005 and later years contain data on approximately 1 percent of the U.S. population, while ACS 5-year PUMS files for 2005–2009 and later years contain data on approximately 5 percent of the population.

The next step is to download the data set(s) of interest (see Figure 3.2). PUMS files on the Census Bureau's FTP site are stored as ZIP files. The naming convention for PUMS files on the FTP server is based on three file features: the file format, the record type, and the state abbreviation.

- File formats are comma separated value files (CSV) and SAS data sets for UNIX.
- Record types are housing files (h) or person files (p).
- State (or state equivalent) abbreviations are two letter labels such as "tx" for Texas and "dc" for District of Columbia. The abbreviation for the file containing all records in the United States is "us."

[17] The Census Bureau previously released 3-year estimates based on 36 months of data collection. In 2015, the 3-year products were discontinued. The 2011–2013 ACS 3-year estimates, released in 2014, are the last release of this product.

Figure 3.2. **Downloading ACS 5-Year PUMS Files From the Census Bureau's FTP Server**



Source: U.S. Census Bureau, FTP server, <https://www2.census.gov/programs-surveys/acs/data/pums/2017/5-Year/>.

PUMS data sets can be downloaded for the entire nation, individual states, the District of Columbia, or for Puerto Rico. Data for Puerto Rico are not included in the national files. To create a complete national data set, you need to combine two or more files together. The 1-year data are divided into two files—an "a" and a "b" file—while the 5-year PUMS is divided into "a," "b," "c," and "d" files.

TIP: If you only need data for a single state or a few states, you can save hard disk space and computer-processing time by downloading data only for those areas rather than the entire nation. For example, the 2016 ACS 1-year PUMS SAS data set for the nation as a whole—including all person and housing records—is 4.0 GB in size, while the 2012–2016 ACS 5-year data set is 18.8 GB in size.

Housing and Population Records

The ACS PUMS data sets are hierarchical, with separate records for housing units and population. The housing records include information about the housing unit—such as the number of rooms, whether the unit is vacant, access to the Internet, and mortgage payments—and household characteristics such as household income and the presence of children. The housing data set contains one record for each sampled household.

The basic unit in the ACS PUMS is an individual housing unit. The PUMS also includes a sample of people living in group quarters (GQ). The population sample is defined as all persons living in households selected in the housing unit sample, plus the persons selected from the GQ sample. Each GQ record is assigned a unique SERIALNO, and data users cannot determine whether two individuals reside in the same GQ.

The population records contain information about individuals such as age, marital status, and educational attainment. The population data set includes one record for each individual in each sampled household or group quarters. Because PUMS data sets are large, you should only download the records that you need. If you need both housing and population records, then you need to download both files and merge them together. (See the section on "Preparing ACS PUMS Data Files for Analysis" for more information.)

Click on the selected data set to download the PUMS data. The data set is provided as a compressed ZIP file along with a README document that provides information about how to start using PUMS data.

Microdata Access on Data.census.gov

The Census Bureau recently launched a new microdata access tool on data.census.gov that may be used to generate ACS estimates online without the use of statistical software. This tool is still under development and in beta form on the Census Bureau's Web site.[18] The Census Bureau created a step-by-step guide on how to use this tool to produce custom estimates from the ACS 1-year PUMS file.[19]

[18] U.S. Census Bureau, Microdata Access Tool, <https://data.census.gov/mdat/>.

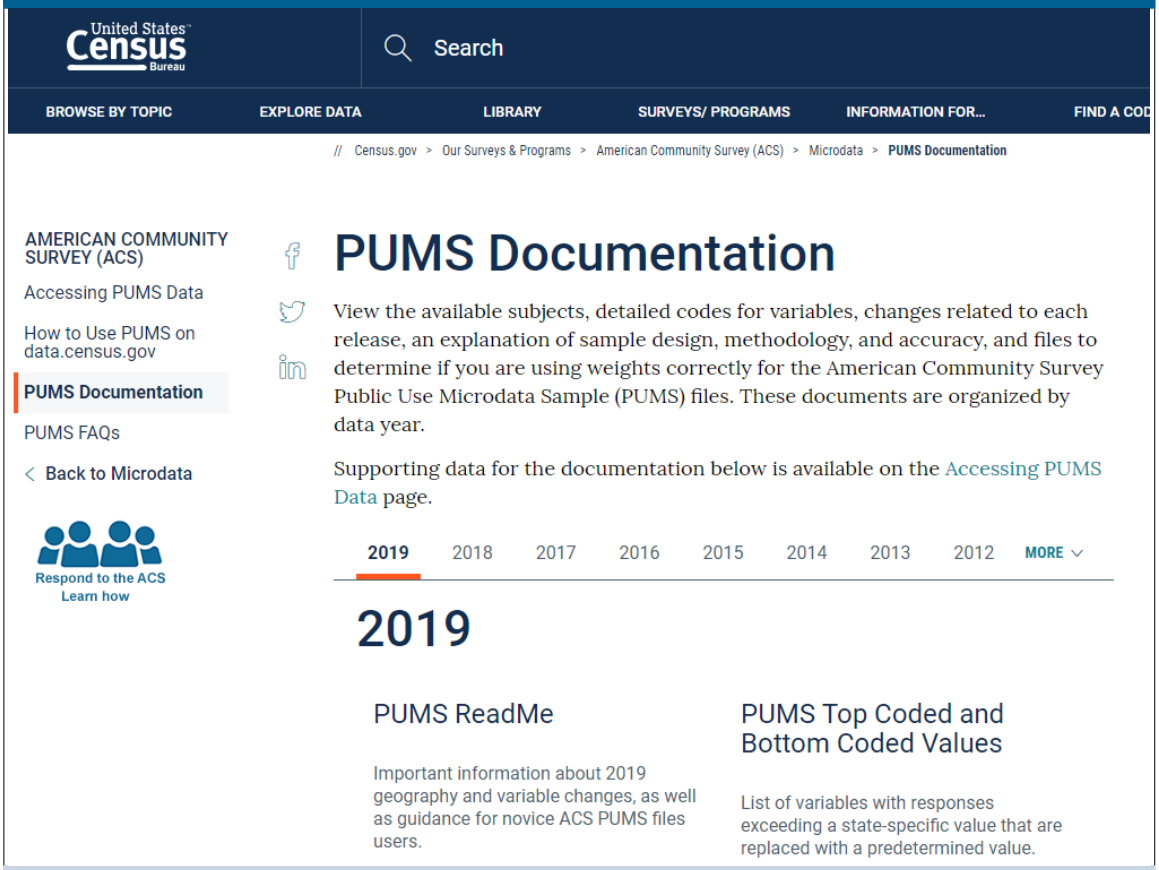
[19] U.S. Census Bureau, Using Microdata Access, <https://www2.census.gov/data/api-documentation/using-microdata-access/microdata-access-1-year-acs-pums.pdf>.

4. PREPARING ACS PUMS DATA FILES FOR ANALYSIS

The data dictionary available on the PUMS Documentation Web page includes the complete list of variables in the American Community Survey (ACS) Public Use Microdata Sample (PUMS) data sets and additional documentation to help data users work with the files (see Figure 4.1).[20]

[20] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

Figure 4.1. **ACS PUMS Documentation**



U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

The PUMS Documentation is organized by year of data release and includes several useful resources:

- PUMS ReadMe is updated with each data release and includes information about geography and variable changes, as well as guidance for getting started in using the ACS PUMS files.
- Subjects in the PUMS provides a list of the subjects that are included in the ACS PUMS files.
- PUMS Data Dictionary includes a list of all the variables in the ACS PUMS files, their values, and descriptions of what those values mean. The ACS PUMS data set includes variables for nearly every topic included in the ACS, as well as variables that were created by combining multiple survey responses.
- Code Lists provides detailed codes for variables with a long list of coded responses (for example, occupation or ancestry).
- PUMS Top Coded and Bottom Coded Values is a list of all the variables that have been top- and/or bottom-coded, and a list of state-specific values for the top- and bottom-coded variables. (More information about top-coding and bottom-coding is provided below.)
- Accuracy of the PUMS explains the sample design, estimation methodology, and the accuracy of the data. This section also includes instructions for calculating standard errors and margins of error for ACS PUMS estimates. (See the section on "Data Quality in the ACS PUMS" for more information.)
- PUMS Estimates for User Verification are weighted national- and state-level estimates, standard errors, and margins of error for several characteristics that data users can use to confirm that PUMS data files have been set up correctly.

PUMS data can be analyzed using a variety of statistical software packages (for example, SPSS, SAS, R, or Stata). In the examples below, SAS programming code is used to prepare the ACS PUMS data for analysis.

Working with Data for the Entire United States

The national ACS PUMS data sets contain millions of records and are very large. As a result, data sets for the entire United States are split into multiple files—"a" and "b" files for the ACS 1-year PUMS, and "a," "b," "c," and "d" files for the ACS 5-year PUMS. Data users must combine these component data files together to create a complete data set for the United States.

Here is a SAS statement that concatenates the "a" and "b" files for the ACS 1-year PUMS population records:

```
/*Concatenate a and b population records to obtain
all U.S. PUMS person records*/
data us_pums_person_data;
set psam_pusa psam_pusb;
run;
```

Data for individual states, the District of Columbia, and for Puerto Rico are also available as separate population data sets. When working with data for a single state, there are no "a" and "b" files, so data users can skip this step.

Combining Housing and Population Data

While some analyses can be conducted using only the housing data or population data, there are other instances where data users need to link housing unit records to population records. For example, housing unit records include data on the number of vehicles available, but to identify the characteristics of people who may have access to those vehicles, you need to link the housing unit records to the population records.

To combine the housing unit and population records, merge the two data sets together by matching on the SERIALNO variable. The SERIALNO variable is unique for each housing unit record across the nation. Matching on this variable merges the housing unit variables onto the population records.

Here is a portion of a SAS program that merges housing and population records:

```
/*Concatenate a and b housing records to obtain all
US PUMS housing records*/
data us_pums_housing_data;
 set psam_husa psam_husb;
run;

/*Sort the housing and population records by
SERIALNO*/
proc sort data=us_pums_housing_data;
by SERIALNO;
run;
proc sort data=us_pums_person_data;
by SERIALNO;
run;

/*merge the housing and population records
together*/
data merged;
merge us_pums_person_data us_pums_housing_data;
by SERIALNO;
run;
```

Merging the housing and population records yields a population-level data set that can be used to estimate the number or share of people with various household-based characteristics.[21]

Within households, data for each household member are included in separate person-level records. However, data users may be interested in producing estimates that combine data from multiple household members. For example, you may want to know the proportion of spouses who have the same level of educational attainment. The steps involved in joining these records are shown in Appendix A.

Selecting Appropriate Weights

Each housing and person record is assigned a weight, because the records in the PUMS files represent a sample of the population. The weight is a numeric variable expressing the number of housing units or people that an individual microdata record represents. The sum of the housing unit and person weights for a geographic area is equal to the estimate of the total number of housing units and people in that area. To generate estimates based on the PUMS records, data users must correctly apply weights.

TIP: To generate statistics for housing units or households (for example, data on average household income), data users should apply the PUMS household weights (WGTP). To generate statistics for individuals (such as age or educational attainment), data users should apply the PUMS person weights (PWGTP).

When working with a merged file that includes both housing and person records, person weights should be used to produce estimates for person characteristics. Housing characteristics cannot be tallied from this merged file without taking extra steps to ensure that each housing weight is counted only once per household.

There are two additional sets of weights, one for households ranging from WGTP1 to WGTP80, and one for individuals ranging from PWGTP1 to PWGTP80. These "replicate weights" are used to calculate the error associated with each estimate. For more information about replicate weights, see the section below on "Data Quality in the ACS PUMS."

[21] Note that the housing unit data contain vacant housing units. These records will not have any corresponding person records.

Selecting Variables

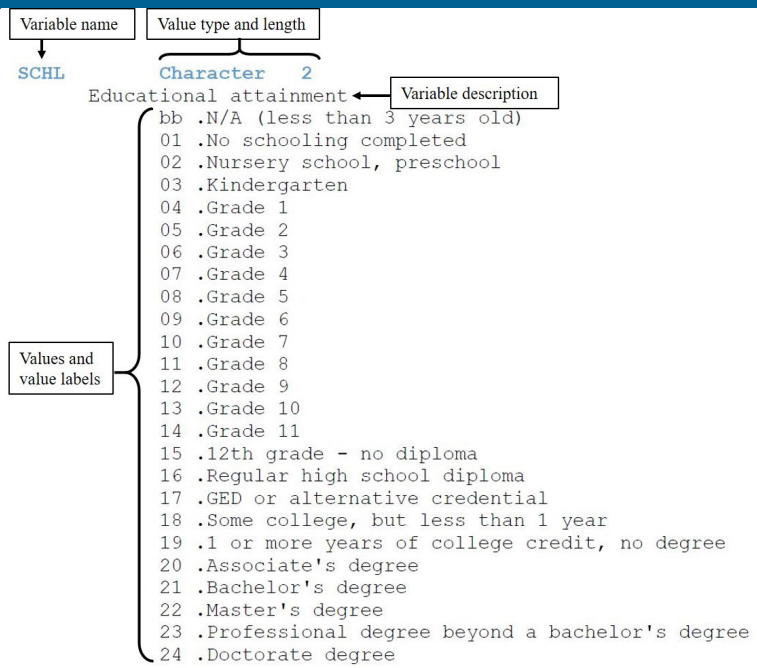
The PUMS Data Dictionary is available for each ACS PUMS data set and includes a complete list of the variables in the PUMS data files.[22] For each variable, the data dictionary includes the variable name, value type and length, variable description, values, and value labels (see Figure 4.2). The "bb" values indicate that data on educational attainment are not available for persons under the age of 3.

Some variables, such as ancestry, have a large number of coded responses. In addition to the data dictionary, values and value labels for these variables are maintained in a separate Excel file listed under "Code Lists" on the PUMS Documentation Web page.[23] The Code Lists show the detailed ACS codes that are included in each value for a PUMS variable. For example, the "German" ancestry code in the 2017 ACS PUMS includes a range of related responses, such as "German," "Saxon," and "West German" (see Table 4.3).

[22] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

[23] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

Figure 4.2. **Elements of a Sample Variable in the ACS PUMS Data Dictionary**



Source: U.S. Census Bureau, 2017 ACS PUMS Data Dictionary, <www.census.gov/programs-surveys/acs/microdata/documentation.2017.html>.

Table 4.3. **German Ancestry Codes in the 2017 ACS 1-Year PUMS**

PUMS code	Ancestry description	Ancestry code	Corresponding detailed ancestry code
032	German	032	German
		033	Bavaria
		034	Berlin
		035	Hamburg
		036	Hannover
		037	Hessian
		038	Lubecker
		039	Pomeranian
		041	Saxon
		042	Sudetenlander
		043	Westphalian
		044	East German
		045	West German

Source: U.S. Census Bureau, American Community Survey, PUMS Documentation, 2017 ACS 1-year PUMS Code Lists, <www.census.gov/programs-surveys/acs/microdata/documentation.2017.html>.

While the PUMS data dictionary provides the coded responses for each variable, the ACS Subject Definitions provide information about the meaning of each ACS variable or subject.[24] The subject definitions include information about how variables are defined and measured, the population universe, the survey questions used to derive each variable, how variables may have changed over time, and the comparability with ACS and decennial census variables from previous years.[25] These definitions apply to both ACS PUMS data, as well as pretabulated ACS data on the U.S. Census Bureau's Web site.

Top-Coded and Bottom-Coded Variables

ACS responses are strictly confidential. In addition to removing all identifying information, responses to open-ended questions, such as age, income, and housing unit value—where an extreme value might identify an individual—are top-coded and/or bottom-coded. Top-coding is the process of taking any response exceeding a particular value and replacing it with a predetermined value. These predetermined values differ by state. For example, for 2017, if someone in New York reports their age as 103, it will be reported in the ACS PUMS file as 95 (the top-coded value shown for New York). Refer to PUMS Top Coded and Bottom Coded Values in the PUMS Documentation for the list of impacted variables and the predetermined values in each state.[26]

PUMS Estimates for User Verification

PUMS data users are responsible for the accuracy of their estimates. Because PUMS data consist of a subset of the full ACS sample, tabulations from the ACS PUMS will not match those from published tables of ACS data. You can verify that you have correctly accessed and tabulated data from the ACS PUMS file by replicating the values presented in "PUMS Estimates for User Verification" in the PUMS Documentation.[27]

The PUMS estimates for user verification include weighted estimates for selected characteristics and associated standard errors and margins of error. The standard errors and margins of errors were calculated using the Successive Difference Replicate method (described in the next section). The estimates are produced for the United States and for each state, the District of Columbia, and for Puerto Rico. They are available in SAS and comma-separated value (CSV) formats.

Beginning in 2017, the user verification files for the ACS PUMS also include a CSV file with unweighted counts of the number of records in each PUMS file. Data users may verify that the number of records in their PUMS person or housing file match the number given in this file.[28]

[24] U.S. Census Bureau, American Community Survey (ACS), Code Lists, Definitions, and Accuracy, <www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html>.

[25] The Census Bureau's subject definitions were created primarily for use with published tables, and some of these definitions are not applicable to the variables in the PUMS.

[26] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

[27] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

[28] Previously, the number of records for the United States and Puerto Rico combined was provided at the beginning of the PUMS "Accuracy of the PUMS" document.

5. DATA QUALITY IN THE ACS PUMS

The weighted estimates calculated using the Public Use Microdata Sample (PUMS) files are estimates of the entire population. The degree of uncertainty associated with American Community Survey (ACS) estimates, known as sampling error, tends to be large when the sample size is small. Indeed, PUMS estimates are subject to additional sampling error because the PUMS data consist of a subsample of the full ACS sample. Researchers using the PUMS files must calculate their own measures of uncertainty in addition to producing their own estimates.

One way to quickly evaluate your results from the ACS PUMS is to reproduce the estimates without using weights. By examining the unweighted counts, you can see if there are any cells based on just a few sample cases, which are less likely to yield statistically reliable weighted estimates.

However, most researchers want to calculate more formal measures of sampling error such as standard errors, margins of error, and confidence intervals. Both the margin of error and the confidence interval can be calculated based on the standard error. For more information about sampling error in the ACS, see the section on "Understanding Error and Determining Statistical Significance" in the U.S. Census Bureau's handbook on *Understanding and Using American Community Survey Data: What All Data Users Need to Know*.[29]

There are two ways to calculate standard errors for ACS PUMS estimates. The first is a generalized variance formula (GVF) that uses design factors. The second is a successive difference replicate (SDR) method that uses the replicate weights. The Census Bureau uses the SDR method to calculate margins of error for published ACS tables.

Generalized Standard Error Formula

Generalized standard errors are model-based. They are, therefore, considered less accurate than direct standard errors calculated with the replicate weight method, although they may be easier to calculate for some data users.

The Census Bureau provides formulas to approximate GVF standard errors in the Accuracy of the PUMS document provided with each ACS PUMS data release.[30] To calculate GVF standard errors, design factors are applied to reflect the effects of the actual ACS sample design and estimation procedures.[31] Prior to 2017, the design factors for PUMS subject groups and state are available in the PUMS accuracy document. Beginning with 2017 data, the design factors are provided in a comma separated value (CSV) file.

ACS estimates (and their corresponding standard errors) may be generated from two or more variables representing different subject areas. When more than one subject is involved in the analysis, use the largest design factor for the factors being considered. For example, an estimate of the population under the age of 21 living below poverty is derived from ACS variables on age and poverty status. The design factor for "Poverty Status (Person)," at 1.9, is larger than the design factor for "Age," at 1.0, so the design factor for poverty should be used to calculate the standard error for this estimate. The only exception to this rule is for items cross-tabulated by race or Hispanic origin. For those items, use the largest design factor not including the race or Hispanic origin design factor.

Successive Difference Replicate Formula (Replicate Weights)

One benefit of using the SDR method is that a single formula can be used to calculate standard errors for many different types of ACS estimates, such as counts, aggregates, percentages, and ratios. However, the SDR method may be inconvenient for some data users because of the computational requirements.

The SDR method uses the 80 replicate weights to calculate 80 replicate estimates. These replicate estimates use the replicate weights. For the PUMS person files, they are called PWGTP1 through PWGTP80. For the housing files, the replicate weights are WGTP1 through WGTP80. Note that these weights are used solely for calculating uncertainty. They should never be used to calculate an estimate.

The first step is to generate an ACS estimate of interest using the PUMS weight (PWGTP or WGTP). Next,

[29] U.S. Census Bureau, *Understanding and Using American Community Survey Data: What All Data Users Need to Know*, <www.census.gov/programs-surveys/acs/guidance/handbooks/general.html>.

[30] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

[31] U.S. Census Bureau, American Community Survey (ACS), Code Lists, Definitions, and Accuracy, 2016 ACS 1-year Accuracy of the Data, <www.census.gov/programs-surveys/acs/technical-documentation/code-lists.2016.html>.

generate this estimate 80 times, using each of the 80 different replicate weights. Once you have these 81 estimated values, you can calculate the standard error by using the following formula:

$$SE(X) = \sqrt{\frac{4}{80}\sum_{r=1}^{80}(X_r - X)^2}$$

where:

X = the estimate based on the PUMS weight (PWGTP or WGTP).

Xr = the 80 individual estimates based on each of the replicate weights.

More details about calculating standard errors using the replicate weight method can be found in the Accuracy of the PUMS documentation.[32] The PUMS Estimates for User Verification include examples of standard errors that were calculated based on the replicate weight method.[33]

Data users interested in worked examples based on the replicate weight method can also consult the documentation for the ACS Variance Replicate Tables.[34] These tables are intended for advanced users who are aggregating pretabulated ACS data and want to calculate exact margins of error.

[32] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

[33] U.S. Census Bureau, American Community Survey (ACS), PUMS Documentation, <www.census.gov/programs-surveys/acs/microdata/documentation.html>.

[34] U.S. Census Bureau, American Community Survey (ACS), Variance Replicate Tables Documentation, <www.census.gov/programs-surveys/acs/technical-documentation/variance-tables.html>.

6. ADDITIONAL RESOURCES

ACS PUMS Documentation
<www.census.gov/programs-surveys/acs/microdata/documentation.html>

This Web page includes information about the American Community Survey (ACS) Public Use Microdata Sample (PUMS), as well as the PUMS technical documentation.

U.S. Census Bureau, Understanding and Using American Community Survey Data: What All Data Users Need to Know
<www.census.gov/programs-surveys/acs/guidance/handbooks/general.html>

This handbook provides an overview of the ACS to help data users understand the basics of the survey, how the data can be used, how to judge the accuracy of ACS estimates, and how to access ACS data.

Webinar: Calculating Margins of Error the ACS Way
<www.census.gov/data/academy/webinars/2020/calculating-margins-of-error-acs.html>

In this Webinar, U.S. Census Bureau staff discuss how to calculate variances, standard errors, and margins of error using the Successive Difference Replicate method.

Webinar: Introduction to the Public Use Microdata Sample (PUMS) File
<www.census.gov/data/academy/webinars/2020/introduction-to-american-community-survey-public-use-microdata-sample-pums-files.html>

In this Webinar, U.S Census Bureau staff discuss foundational aspects of working with the ACS PUMS files, including the organization of the files, the confidentiality of the files, accessing the data, geographic availability, and the PUMS documentation. This Webinar also explores how to use new features on data.census.gov to create custom PUMS tables.

Geography and ACS
<www.census.gov/programs-surveys/acs/geography-acs.html>

This Web page includes information about changes in geographic boundaries in the ACS, key concepts and definitions, and reference maps.

Using Microdata Access
<https://www2.census.gov/data/api-documentation/using-microdata-access/microdata-access-1-year-acs-pums.pdf>

This document provides a step-by-step guide for using microdata access on data.census.gov to create your own tables using the ACS PUMS files.

ACS Online Community
<https://acsdatacommunity.prb.org/>

The ACS Online Community is a site where people can share messages, materials, and announcements related to ACS data, methods, and events.

APPENDIX A: LINKING HOUSEHOLD MEMBERS TOGETHER

To link records for married couples that include the householder, first identify the relevant variables for the analysis. In this example, the variables HHT (household type), RELP (relationship to householder), and MAR (marital status) can be used to identify married householders and their spouses. SCHL (educational attainment) can be used to measure educational attainment.

Here is a portion of a SAS program that identifies married householders and their spouses.

```
/*Create a person file of householder, which is the
reference person*/

data householder;
 keep serialno reference h_schl h_mar h_pwgtp;
  set psam_pus;
if hht in (1,2,3) and relp='00' and mar='1';

/*define householder and set it equal to 1 for the
householder */;
reference=1;
/*Create variables for the householder/reference
person;
   These variables will start with 'h_'*/;
h_schl=schl;
h_mar=mar;
h_pwgtp=pwgtp;
run;

/*Create a data file of spouses*/
data household_spouse;
 keep serialno spouse schl s_mar s_pwgtp;
 set persons;
if hht in (1,2,3) and relp='01' and mar='1';

spouse=1;
s_schl=schl;
s_mar=mar;
s_pwgtp=pwgtp;
run;

/*Sort the data before merging */;
proc sort data = householder;
 by SERIALNO;
proc sort data = household_spouse;
 by SERIALNO;
run;
```

Next, create separate data files for each person that will be linked together. For each of these data files, rename the variables so they are unique for each person. Keep only the identifiers and person-specific variables that will be used in the analysis.

Following these steps produces a data file of married couples based only on the householder and his or her spouse. There may be additional married couples within a household—called subfamilies—that can be identified with the subfamily relationship variable. To create a data file that includes all married couples, data users will need to follow similar steps to identify married couples in subfamilies and link spouses together.