AMERICAN COMMUNITY SURVEY
2009-2013 ACS 5-YEAR PUMS FILES

Prepared by
American Community Survey Office
U.S. Census Bureau
January 22, 2015

I.) Overview of the Public Use Microdata Sample (PUMS)

The Public Use Microdata Sample (PUMS) contains a sample of actual responses to the American Community Survey (ACS). The PUMS dataset includes variables for nearly every question on the survey, as well as many new variables that were derived after the fact from multiple survey responses (such as poverty status). Each record in the file represents a single person, or--in the household-level dataset--a single housing unit. In the person-level file, individuals are organized into households, making possible the study of people within the contexts of their families and other household members. PUMS files for an individual year, such as 2013, contain records of data from approximately one percent of the United States population. As such, PUMS files covering a three-year period, such as 2011-2013, contain records of data from approximately three percent of the United States population, while PUMS files covering a five-year period, such as 2009-2013, contain records of data from approximately five percent of the United States population.

The PUMS files are much more flexible than the aggregate data available on American FactFinder, though the PUMS also tend to be more complicated to use. Working with PUMS data generally involves downloading large datasets onto a local computer and analyzing the data using statistical software such as R, SPSS, Stata, or SAS.

Since all ACS responses are strictly confidential, many variables in the PUMS file have been modified in order to protect the confidentiality of survey respondents. For instance, particularly high incomes are "top-coded", uncommon birthplace or ancestry responses are grouped into broader categories, and the PUMS file provides a very limited set of geographic variables (explained more below).

II.) Public Use Microdata Areas (PUMA)

While PUMS files contain cases from nearly every town and county in the country, towns and counties (and other low-level geography) are not identified by any variables in the PUMS datasets. The most detailed unit of geography contained in the PUMS files is the Public Use Microdata Area (PUMA). PUMAs are special non-overlapping areas that partition each state into geographic units originally defined as containing no fewer than 100,000 people each.

Please note that there are two sets of PUMA geographies on this file. The 2009-2013 5-year ACS PUMS files use PUMA boundaries that were drawn by state governments at the time of Census 2000 and the 2010 Census. Due to disclosure avoidance procedures, the ACS PUMS does not release overlapping geographies for any record. Therefore the records from data years 2009 through 2011 still carry the older 2000-based PUMA codes. Only the 2012 and 2013 records display the newer 2010-based PUMA geography. If you choose to use PUMAs on this multiyear file, please read closely the discussion of the two PUMA codes in the Attachment, and also the section entitled "Variable Changes in the 2009-2013 5-year PUMS File".

III.) PUMS Documentation

The PUMS Documentation page (http://www.census.gov/acs/www/data_documentation/pums_documentation/) includes the following documents:

- **Subjects in the PUMS**
- **PUMS Code Lists**
- **PUMS Top Coded and Bottom Coded Values**
  This document contains tables that show the top code only or the top code and bottom code values for each of these housing and person variables by state.
- **PUMS Data Dictionary**
  Information on PUMS variables.
- **PUMS ReadMe**
- **PUMS Estimates for User Verification**
  PUMS estimates for selected housing and population characteristics are included on the ACS website to assist data users in determining that they are correctly using the weights to compute estimates. These estimates are referred to as PUMS Control Counts. When data users have doubts about the way they are computing estimates, they should attempt to reproduce the estimates that are provided in the files. The standard errors provided in this document were computed using the replicate weight method.
- **Accuracy of the PUMS**
  Detailed descriptions of the sampling methodology, weighting methodology, confidentiality, and standard errors for the PUMS.

IV.) Getting PUMS data

ACS Website

PUMS files can be accessed via the ACS website at http://www.census.gov/acs/www/data_documentation/pums_data/.

American FactFinder

PUMS Files are also accessible via American FactFinder at http://factfinder2.census.gov/.

Data Ferrett

It is also possible to get PUMS data from the Census Bureau's DataFerrett, which has the additional feature of being able to make tables and perform basic analysis online. This tool is particularly useful for researchers who need a quick statistic or do not have access to statistical software. DataFerrett is available at http://www.census.gov/acs/www/data_documentation/data_ferrett_for_pums/.

V.) PUMS file structure

The ACS questionnaire contains "household" items that are the same for all members of the household (such as the number of rooms in the home) and "person" items that are unique for each household member (such as age, sex, and race). The ACS PUMS files are made available in this

same structure. Researchers who are analyzing only household-level items can use the household files, whereas those using only person-level variables can use the person-level files.

Data users should note that PUMS files containing data for the entire United States (in contrast to individual state and state-equivalent files) are separated into multiple data files. These files must be concatenated in order to create a complete file. For example, users downloading the 2009-2013 ACS 5-year PUMS file of United States Population Records will notice an “a” file, “b” file, “c” file, and “d” file. Each file contains approximately one-fourth of the population records in the 2009-2013 5-year PUMS dataset of the United States. Below are instructions for concatenating the four PUMS person-level files, in the form of an italicized SAS program and pseudo-code.

Concatenate the four **person-level** files using the set statement:

data population;
set psam_pusa psam_pusb psam_pusc psam_pusd;
run;

The 2009-2013 ACS 5-year PUMS file of the United States Housing Records also contains an “a” file, “b” file, “c” file, and “d” file. To create a complete housing-level file, the four files must be concatenated. Below are instructions for concatenating the four PUMS household-level files, in the form of an italicized SAS program and pseudo-code.

Concatenate the four **household-level** files using the set statement:

data housing;
set psam_husa psam_husb psam_husc psam_husd;
run;

Some data users will need to use household and person items together--for instance, to analyze how the number of rooms in a home varies by a person’s age. This type of analysis will require the merging of the household and person files. This merger must rely on the SERIALNO variable, which is the same in the household and person files. Below are instructions for merging the housing and population PUMS files, in the form of an italicized SAS program and pseudo-code.

Use the variable SERIALNO to merge population and housing files.

1. First make sure the files are sorted by SERIALNO:
    proc sort data=population;
    by serialno;
    run;
    proc sort data=housing;
    by serialno;
    run;

2. Then merge the two files together using SERIALNO as a merge key.
    data combined;
    merge population (in=pop) housing;

/*In SAS, the 'in=' option will allow you to keep only those housing units that have people*/

```
    by serialno;
```

/*This SAS statement keeps only those housing units that were in the population file*/

```
    if pop;
    run;
```

You should not merge the files unless the estimates you want require a merge. Note that there are many estimates that can be tabulated from the person file and from the household file without any merging. The suggested merge will create a person level file, so that the estimate of persons can be tallied within categories from the household file and the person weights should be used for such tallies.

Please note that housing characteristics cannot be tallied from this merged file without extra steps to ensure that each housing weight is counted only once per household.

VI.) Weights in the PUMS

The ACS PUMS is a weighted sample, and weighting variables must be used to generate accurate estimates and standard errors. The PUMS file includes 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. The weighting variables are described briefly below.

PWGTP: Person's weight for generating statistics on individuals (such as age).

WGTP: Household weight for generating statistics on housing units and households (such as average household income).

WGTP1-WGTP80 and PWGTP1-PWGTP80: Replicate weighting variables, used for generating the most accurate standard errors for households or individuals.

PWGTP and WGTP can be used both to generate the point estimates and to generate standard errors when using a generalized formula. Replicate weights can be used just to calculate "direct standard errors." Direct standard errors are expected to be more accurate than generalized standard errors, although they may be more inconvenient for some users to calculate. Both generalized and direct standard errors are explained in more detail in the Accuracy of the PUMS document (http://www.census.gov/acs/www/data_documentation/pums_documentation/).

Each housing unit and person record contains 80 replicate weights. To use the replicate weights to calculate an estimate of the direct standard error, first form the estimate using the full PUMS weight, then form the estimate using each of the 80 replicate weights--providing both the full PUMS estimate

and 80 replicate estimates. These should then be entered into the following formula, which is explained in more detail in the Accuracy of the PUMS document:

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

Where X_r is a replicate estimate from X_1 to X_{80}, and X is the full PUMS weighted estimate.

The technical explanation of the ACS replicate weights is in Chapter 12 of the Design and Methodology document found at: http://www.census.gov/acs/www/methodology/methodology_main/. For more information on the theoretical basis, please reference Fay, R. and Train, G. (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," Proceedings of the Section on Government Statistics, American Statistical Association, pp. 154-159, 1995."

Please note that many estimates generated with PUMS will be slightly different from estimates for the same characteristics published in American FactFinder. These differences are due to the fact that the PUMS files include only about two-thirds of the cases that were used to produce estimates on American FactFinder, as well as additional PUMS edits. More information on the PUMS sample design is available in the "Accuracy of the PUMS" document (http://www.census.gov/acs/www/data_documentation/pums_documentation/).

VII.) Variable changes in the 2009-2013 5-year PUMS file

The 2009-2013 ACS PUMS includes most of the variables that were included in the 1-year PUMS files from 2009 through 2013. It does not include the new computer and internet related variables which are only in the 2013 1-year PUMS. See the 2009-2013 5-year PUMS Data Dictionary for a complete listing of the variables and values contained in this PUMS data file.

New variables that were not in the previous 5-year PUMS file: FFODP, FOD1P, FOD2P, FRWATPRP, FWRKP, MLPCD, MLPFG, RWATPR, SCIENGP, SCIENGRLP, WRK.

Variables dropped from the previous 5-year PUMS: MLPC, MLPD, MLPF, MLPG.

Variables with new or modified codes since the previous 5-year PUMS: ADJHSG, ADJINC, CITWP12, INDP, MARHYP12, MIL, NAICSP, PERNP, PINCP, RESMODE, RWAT, SERIALNO, SRNT, SVAL, VALP, YBL, YOEP12 .

Variables with cosmetic changes to variable labels or value labels: ANC1P05, ANC1P12, ANC2P05, ANC2P12, CITWP05**,** CITWP12**,** DRAT, DRATX, FSTOVP, FTAXP, HHT, LANP05, LANP12, LNGI, MARHYP05, MARHYP12, MIGPUMA00, MIGPUMA10, MIGSP05, MIGSP12, NAICSP, OCCP02, OCCP10, OCCP12, PAOC, PLM, POBP05, POBP12, POWPUMA00, POWPUMA10, POWSP05, POWSP12, PUMA00, PUMA10, RAC2P05, RAC2P12, RAC3P05, RAC3P12, SOCP00, SOCP10, SOCP12, SRNT, SVAL, YOEP05, YOEP12.

Multiple vintage variables:

There is a new set of dual vintage variables that applies to Puerto Rico only: RWAT and RWATPR Beginning with the 2013 data year, the hot and cold running water variable RWAT is set to '9' for not applicable in Puerto Rico, while a new variable RWATPR (running water) replaces it. For years previous to 2013, use RWAT.

As a result of data disclosure requirements, a number of variables were recollapsed into new categories for data year 2012. A dual vintage variable set was formed whenever both the original and new categories were preserved on the multi-year PUMS file. To distinguish between the old and new values, the variable from the 1-year PUMS was renamed to form two variables on the 3-year PUMS. In each set of dual vintage variables, one variable presents values used in years before 2012, and the other presents values used in 2012 and later data years, (except for RWAT/RWATPR mentioned above). In order to obtain data for the entire PUMS sample, both variables must be used. A value of -9, -09, -009, or -0009 (depending on the variable's length) is assigned to cases for which the variable is not applicable due to the data year. The exceptions are OCCP02, OCCP10, OCCP12, SOCP00, SOCP10, and SOCP12, for which not applicable codes N.A. and N.A.// are used. See the 2009-2013 5-year PUMS Data Dictionary to find the following dual vintage variables contained in this 5-year PUMS data file.

ANC1P05, ANC1P12
ANC2P05, ANC2P12
CITWP05, CITWP12
LANP05, LANP12
MARHYP05, MARHYP12
MIGSP05, MIGSP12
OCCP02, OCCP10, OCCP12
POBP05, POBP12
POWSP05, POWSP12
RAC2P05, RAC2P12
RAC3P05, RAC3P12
RWAT, RWATPR (the dual vintage applies to Puerto Rico only. Elsewhere, use RWAT)
SOCP00, SOCP10, SOCP12
YOEP05, YOEP12

NAICSP and INDP are not listed as dual vintage variables since the older categories were crosswalked to the new values used for the 2013 1-year PUMS. For additional information on changes in industry and occupation codes over time, see the crosswalk file under "code lists" at http://www.census.gov/acs/www/data_documentation/pums_documentation/

Dual Vintage Geography Variables

PUMA boundaries were redrawn based on the 2010 Census data. MIGPUMA and POWPUMA are composed of sets of the new PUMAs. Most of the 2010-based PUMAs cannot be mapped directly to the 2000-based PUMAs used for the 2009-2011 PUMS. The Tiger maps discussed in the Attachment can be used to show the boundaries for both sets of PUMAs.

POWPUMA00, POWPUMA10
MIGPUMA00, MIGPUMA10

PUMA00, PUMA10

Variables with suppressed values:
FER - Problems in the collection of data on women who gave birth in the past year (FER) led to suppressing this variable in 59 PUMAs within states Florida, Georgia, Kansas, Montana, North Carolina, Ohio and Texas for data year 2012. A code of 8 was applied to these cases for data year 2012 only.

TEL - Problems in the collection of data on the availability of telephone service (TEL) led to suppressing this variable in six PUMAs in Georgia for data year 2012. A code of 8 was applied to these cases for data year 2012 only.

PLM - Problems in data collection of complete plumbing facilities (PLM) led to the suppression of this variable for Puerto Rico for data years 2012 and 2013. A code of 9 was applied to these cases.

VIII.) Additional Information

The Census Bureau occasionally provides corrections or updates to PUMS files. We notify users of these updates via the Census Bureau's E-mail Updates system (https://service.govdelivery.com/service/subscribe.html?code=USCENSUS_C12) and on the ACS errata page (http://www.census.gov/acs/www/data_documentation/errata/).

Please contact acso.users.support@census.gov with any PUMS-related questions.

Attachment

Further Explanation of the Dual Vintage PUMA Codes

The two sets of PUMA codes on this file include the set of PUMA boundaries that were drawn by state governments at the time of Census 2000 and the new set based on the 2010 Census. As mentioned previously, due to disclosure avoidance procedures, the ACS PUMS does not release overlapping geographies for any record. Therefore the records from data years 2009 - 2011 still carry the older 2000-based PUMA codes. Only the 2012 and 2013 records display the newer 2010-based PUMA geography.

There are limitations to using dual vintage PUMAs since the boundaries of the 2010 PUMAs were formed independently from the 2000 PUMAs. PUMS users who have previously used PUMA codes in PUMS files, may attempt to use dual vintage PUMAs in this 2009-2013 PUMS.

PUMS users who are new to forming substate geography in PUMS files should be warned that even in our PUMS files which had a single PUMA code, it is not always possible to precisely construct specific substate geographies. For example, in the 2007-2011 5-year PUMS file, there was a single PUMA code. Using that code, some PUMS users constructed fuzzy geography. that included additional areas not a part of their target area. For example, a user interested in a large county found that it contained several PUMAs, but one of the PUMAs crosses the county boundary and includes part of an adjacent county or counties. In this case, the user chose to combine all of these PUMAs to form the area of analysis, even though a small portion of the combined area lies outside the county of interest.

To construct the same county using the dual vintage PUMAs in this 2009-2013 5-year PUMS, the user will have to examine both sets of PUMAs. For each set the user must determine if he/she can live with the fuzzy boundaries for the geography of interest. Then the user must use both the 2010 based PUMA code and the 2000-based PUMA codes in order to form estimates from this multi-year file.

The following information is provided to those who want to further investigate the use of the PUMA codes and boundaries on this file.

An interactive mapping application, TIGERweb, can be used to view PUMA boundaries from both Census 2000 and the 2010 Census. TIGERweb is available from the Census Bureau's web site at http://tigerweb.geo.census.gov/tigerwebmain/tigerweb_main.html. To access the maps:

- Click on “TIGERweb applications”.
- Click the “TIGERweb Decennial” link on the left side of the webpage. A new window will open.
- Find the “Layers” tab on the upper left. If it did not open automatically, click on the Layers icon and a drop down menu should open.
- In the “Layers” drop-down menu, find the “Select Vintage” box and select “Census 2000” to view PUMA boundaries used for data years 2009, 2010 and 2011. Select “Census 2010” to view PUMA boundaries used for data years 2012 and 2013.

- Check the box next to the "PUMAs, UGAs, and ZCTAs" and expand this section to check only the box for "Public Use Microdata Areas." You may have to zoom in on the map before this choice is allowed.
- You can now zoom in on the map to view the PUMA boundaries and numbers.
- You will need to adjust the level of zoom until you see the blue numbers that are the PUMA codes.

There are two additional resources that may help PUMS users understand and use PUMAs. They are the software MABLE, developed by the Missouri Census Data Center, and static maps published by the Census Bureau.

The proportion of a PUMA's population that is within a county or other geography can be calculated by using the software MABLE. Found at http://mcdc.missouri.edu/websas/geocorr12.html, it allows you to enter the geography you are interested in and then supplies you with the PUMA codes. This resource also allows you to calculate the proportion of the population in a 2010-based PUMA that lies in a 2000-based PUMA (or vice versa), which can help you make decisions about which PUMAs to analyze.

PDF-format maps of PUMA boundaries drawn at the time of Census 2000 and PUMA boundaries drawn at the time of the 2010 Census are also available from the Census Bureau's web site at http://www.census.gov/geo/maps-data/maps/reference.html. To use:

- From the link above, click on "Public Use Microdata Area (PUMAs)".
- Choose either "2010 Census Public Use Microdata Area (PUMA) Reference Maps", or "Census 2000 Public Use Microdata Area (PUMA) Maps (5-percent sample). For the Census 2010-based PUMAs, each map will show a single PUMA. For the Census 2000-based PUMAs (5 percent), the first page of the PDF document for each state displays entities called "Super PUMAs." (Note that "Super PUMAs" were not formed using Census 2010 data, and are not on the PUMS file.)
- Following the initial state-level Super-PUMA overview map, each PDF file has one or more inset maps which display the boundaries of the PUMAs within each Super PUMA. The maps also show census tract and county boundaries to help you see what geographic areas correspond to the PUMAs.
- Use the zoom feature to read the fine print that identifies geography features.

A listing of the detailed components of each of the Census 2000-based PUMAs is available within the directories at http://www2.census.gov/census_2000/datasets/PUMS/FivePercent/. Information about the 2010 PUMAs can be found at http://www.census.gov/geo/reference/puma.html. Scroll down to the drop-down heading "Reference Information" in order to find information about the 2010 PUMAs, including the 2010 PUMA equivalency files.