AMERICAN COMMUNITY SURVEY
2007-2011 ACS 5-YEAR PUMS FILES

Prepared by
American Community Survey Office
U.S. Census Bureau
January 29, 2013

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 2011, contain records of data from approximately one percent of the United States population. As such, PUMS files covering a three-year period, such as 2009-2011, contain records of data from approximately three percent of the United States population, while PUMS files covering a five-year period, such as 2007-2011, 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. The 2007-2011 ACS PUMS files rely on PUMA boundaries that were drawn by state governments at the time of Census 2000. PDF-format maps of PUMA boundaries are available from the Census Bureau's web site at http://www.census.gov/geo/www/maps/puma5pct.htm.

From this index page, choose a state. The first page of the PDF document for each state displays entities called "Super PUMAs." Super PUMAs differ from the PUMAs available in the ACS PUMS files. The PUMAs in the ACS PUMS are sometimes referred to as "5% PUMAs" because they were delineated for use in presenting PUMS data based on a 5% sample of decennial census data.

The key to using these maps is to understand that the PUMAs nest within the Super-PUMAs. Following the initial state-level Super-PUMA overview map, each PDF file has one or more inset maps for each Super-PUMA. These insets display the boundaries of the PUMAs. The maps show census tract and county boundaries to help you see what geographic areas correspond to the PUMAs, as well.

A listing of the detailed components of each PUMA is available within the directories at http://www2.census.gov/census_2000/datasets/PUMS/FivePercent/.

The Missouri Census Data Center's MABLE/Geocorr2K: Geographic Correspondence Engine with Census 2000 Geographies (http://mcdc2.missouri.edu/websas/geocorr2k.html) has a tool that allows you to enter the geography you are interested in and then it supplies you with the PUMA codes.

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 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.
- **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 2007-2011 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 2007-2011 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 2007-2011 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 weight 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 2007-2011 5-year PUMS file

The 2007-2011 ACS PUMS includes most of the variables that were included in the 1-year PUMS files from 2007-2011. A small number of variables were not included in the 5-year PUMS because they were added to the survey during the 5-year period, or changed so significantly that they cannot be considered comparable over the 5-year period. There were also a small number of variables with new codes, modified codes, or cosmetic changes to variable labels or value labels.

Variables with new or modified codes compared to the previous 5-year file: ADJHSG, ADJINC, DECADE, SERIALNO, YBL, and YOEP.

Variables with cosmetic changes to variable labels or value labels compared to the previous 5-year file: FINCP, FTELP, GCR, HINCP, HUGCL, INTP, LANP, MRGX, MSP, OCCP10, POWPUMA, PUMA, SEMP, SOCP10, TAXP, and TEL.

Major changes to the industry and occupation variables
Census occupation codes are a condensed list of codes based the Standard Occupational Classification (SOC). The SOC implemented a major update in 2010, resulting in the deletion of 38 Census codes, the addition of 68 Census codes, and modification of 1 Census code. All data products have been updated accordingly. This change also affects the occupation variables in the 2007-2011 ACS 5-year PUMS. As a result of the occupational changes, there was a net addition of 22 occupation codes to the PUMS occupation code list: 469 occupation categories in 2009 compared to 491 occupation categories in 2010.

Data previously available in OCCP and SOCP are now presented in 4 separate fields. OCCP10 and SOCP10 contain data for 2010 and 2011 cases only, using the 2010 occupational classification system. OCCP02 and SOCP00 contain data for 2007, 2008 and 2009 cases only, using the 2002 occupational classification system.

Census industry codes are a condensed list of codes based on the North American Industry Classification System (NAICS). In 2008, Census industry codes transitioned to the 2007 NAICS. The new classification system resulted in the addition of 1 industry code (6672), modification of 1 industry code (6670), and the deletion of 2 industry codes (6675, 6692), for a total of 268 industries. Data previously available in INDP and NAICSP are now presented in 4 separate fields. Industry data for 2008-2011 are available in INDP07 and NAICSP07 only. Data for 2007 are available in INDP02 and NAICSP02 only.

The Census Bureau is providing conversion rates to map across the different coding schemes for both industry and occupation variables. These conversion rates, along with additional technical documentation, are available on the ACS PUMS documentation website (listed under code lists), at http://www.census.gov/acs/www/data_documentation/pums_documentation/

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.