AMERICAN COMMUNITY SURVEY
2008-2012 ACS 5-YEAR PUMS FILES

Prepared by
American Community Survey Office
U.S. Census Bureau
March 6, 2014

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 2012, contain records of data from approximately one percent of the United States population. As such, PUMS files covering a three-year period, such as 2010-2012, contain records of data from approximately three percent of the United States population, while PUMS files covering a five-year period, such as 2008-2012, 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, which is different from all previously published ACS PUMS files. The 2008-2012 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 2008 through 2011 still carry the older 2000-based PUMA codes. Only the 2012 records display the newer 2010-based PUMA geography. If you choose to use PUMAs on this multiyear file, please read closely the description of the two PUMA codes, and also the section entitled "Variable Changes in the 2008-2012 5-year PUMS 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. Click the “TIGERweb Decennial” link on the left side of the webpage. A new window will open. To view PUMA boundaries used for the year 2012 in the 2008-2012 ACS PUMS, first select the “Census 2010” option in the “Map

Layers" overlay. Second, 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 can now close or minimize the "Map Layers" overlay and zoom in on the map to view the 2010 PUMA boundaries and numbers. To view PUMA boundaries used for the years 2008 through 2011 in the 2008-2012 ACS PUMS, select the "Census 2000" option in the "Map Layers" overlay.

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.

From this index page, click on "Public Use Microdata Area (PUMAs)". You will then 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.

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/.

The Missouri Census Data Center's MABLE/Geocorr2K: Geographic Correspondence Engine with Census 2000 Geographies (http://mcdc2.missouri.edu/websas/geocorr2k.html) and 2010 Census Geographies (http://mcdc.missouri.edu/websas/geocorr12.html) has a tool that allows you to enter the geography you are interested in and then it supplies you with the PUMA codes. MABLE can calculate the proportion of a PUMA's population that is within a county or other geography that you select.

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.

- **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 2008-2012 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 2008-2012 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 2008-2012 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 2008-2012 5-year PUMS file

The 2008-2012 ACS PUMS includes most of the variables that were included in the 1-year PUMS files from 2008-2012. There were a number of variables with new codes, modified codes, or cosmetic changes to variable labels or value labels. See the 2008-2012 5-year PUMS Data Dictionary for a complete listing of the variables and values contained in the 2008-2012 5-year PUMS data file.

Variables new to the 2008-2012 5-year PUMS file: BATH, BDSP (replaces BDS), CITWP05, CITWP12, DDRS, DEAR, DEYE, DIS, DOUT, DPHY, DRAT, DRATX, DREM, FBATHP, FCITWP, FDDRSP, FDEARP, FDEYEP, FDOUTP, FDPHYP, FDRATP, FDRATXP, FDREMP, FHINS1P, FHINS2P, FHINS3C, FHINS3P, FHINS4C, FHINS4P, FHINS5C, FHINS5P, FHINS6P, FHINS7P, FMARHDP, FMARHMP, FMARHTP, FMARHWP, FMARHYP, FREFRP, FRWATP, FSINKP, FSTOVP, FTOILP, HICOV, HINS1, HINS2, HINS3, HINS4, HINS5, HINS6, HINS7, INDP (replaces INDP07), MARHD, MARHM, MARHT, MARHW, MARHYP05, MARHYP12, MULTG, NAICSP (replaces NAICSP07), PRIVCOV, PUBCOV, REFR, RWAT, RMSP (replaces RMS), SINK, STOV, TOIL, VALP (replaces VAL)

Variables with new or modified codes: ADJHSG, ADJINC, ANC1P05, ANC2P05, RELP, SCHG, SCHL, SERIALNO, YBL

Variables with cosmetic changes to variable labels or value labels: LANP05, POBP05, RAC1P, SCH, WKEXREL, WORKSTAT

Multiple vintage variables: As a result of data disclosure requirements, a number of variables were recollapsed into new categories for data year 2012. As a result, there are multiple vintages for many variables. In order to obtain data for the entire PUMS sample, all of these vintages must be used for a given variable. 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.

ANC1P05, ANC1P12
ANC2P05, ANC2P12
CITWP05, CITWP12
LANP05, LANP12
MARHYP05, MARHYP12
MIGSP05, MIGSP12
OCCP02, OCCP10, OCCP12
POBP05, POBP12
POWSP05, POWSP12
RAC2P05, RAC2P12
RAC3P05, RAC3P12

SOCP00, SOCP10, SOCP12
YOEP05, YOEP12

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/

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 2008-2011 PUMS. Tiger maps show the boundaries for both sets of PUMAs. See: http://www.census.gov/acs/www/data_documentation/geography/ , select "Other Mapping Resources", select "2010 Census Public Use Microdata Area (PUMA) Maps", and then select "TIGERweb". On the TIGERweb site, select "TIGERweb Decennial." There you will have the option to map 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.

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

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.

Variables removed from the 5-year file: BDS (replaced by BDSP), INDP02, INDP07 (replaced by INDP), NAICSP02, NAICSP07 (replaced by NAICSP), RMS (replaced by RMSP), VAL (replaced by VALP)

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.