American Housing Survey (AHS): Using the Internal User File (IUF)

The American Housing Survey (AHS) is sponsored by the U.S. Department of Housing and Urban Development (HUD) and the data are collected by the Census Bureau. The first national survey was conducted in 1973 and the first metro area was surveyed in 1974. To date surveys have been conducted in 48 metro areas.

Accessing the AHS

The AHS data are available to the general public in SAS and ASCII format. The Public Use Files (PUFs) are available at http://www.huduser.org/portal/datasets/ahs.html. At this site users can download:

1) Datasets of all AHS surveys dating back to 1997;
2) Order datasets of AHS surveys conducted prior to 1997 on CD-ROM (*till 1989 – years previous are available at the National Archives*);
3) The codebook that provides the name and detailed information on each variable for each survey year (*to access each year's codebook, click on that year's survey link on the webpage mentioned above and then click on the "Codebook to the American Housing Survey" link*);
4) The survey questionnaire/booklet in both English and Spanish; and
5) Information on topcoded and bottomcoded variables and the methodology used in the coding.

Why Use the Internal User File (IUF)?

The AHS PUFs include most of the data collected on the AHS, but in order to maintain the privacy of the respondents, many items are not released on the PUF, are recoded and/or topcoded or bottomcoded (i.e., really high and really low values are coded to a single value) and the level of geography available is limited and some of it is masked. The following table shows the differences between the PUF and the IUF.

Public Use File (PUF)	Internal Use File (IUF)
Topcoded/Bottomcoded and Perturbed (Age) data to protect the privacy and confidentiality of respondents	Non-topcoded/bottomcoded data and age is not perturbed
Some variables, such as amenities available in the community, may be recoded to a single Yes/No variable	All individual variables collected in the survey are available
Other variables, such as some variables tracking details on why an unit was not interviewed, are not available	All variables in the interview are available
Merged/Masked geography	Detailed geography (including 1980 census tract)
Recoded/Collapsed information on home improvement	Detailed information on home improvement at the job-level
No information on sampling frames	Detailed information on sampling frames
In some metro areas some cases may be dropped for disclosure purposes, as the population count in certain areas may no longer meet Census' threshold for disclosure (post-Hurricane Katrina New Orleans, for	All cases are available

Public Use File (PUF)	Internal Use File (IUF)
example)	
Not recoded/computed variables that are used to create the publication tables	A file including the recoded/computed variables used to create tables are included
No information from MASTER file	MASTER file which includes detailed information on sampling frame, geography, etc. is included

Using the IUF

The AHS internal user files (IUFs) contain geographic identifiers (e.g., census tract) that do not appear on the public use file (PUF) and also includes data that has not been topcoded, or aggregated for confidentiality reasons. In addition, they contain variables that can be used to calculate the correct standard errors for a complex random sample. The data, however, has been edited and imputed. The internal IUFs available at the CES do not include names and addresses or unedited data.[1]

The main documentation for the AHS file is the PUF codebook, available on the HUD User web site (http://www.huduser.org/portal/datasets/ahs.html). In general, the AHS internal variable names and file structure corresponds to that of the PUF. There are some exceptions, some of which are described in this document.

About the Data Files

Note, that all IUF files consist of ALL variables on the PUF, including computed variables that are released on the PUF (i.e., variables that are created by combining or recoding other variables). The only difference between variables on the PUF and IUF are that all IUF variables are the actual data, whereas some variables on the PUF have been topcoded or bottomcoded and have been otherwise adjusted to protect the privacy and maintain the confidentiality of the respondent.

All data files described below are in SAS format. The internal use AHS datasets consist of 11 different SAS files. They are:

> **HOUSHLD/NEWHOUSE**: This file used to be called HOUSHLD (until 2005) and is now called NEWHOUSE. The file has the same name in the PUF and IUF and contains most of the survey data. It has one record per household and the IUF includes more variables than the PUF.
>
> **PERSON:** This file includes a single record for each member of the household. This means there can be multiple records for each occupied household and there are no records for vacant, usual-residence-elsewhere and noninterview units. This file includes all person-level data including race/ethnicity, income, age, disability status, etc. of each person in the household. Like on the PUF, the variable MVG on the person

[1] The unedited data is maintained mainly for the purposes of developing the AHS processing system. An example is VOTHER (non-wage income), which is available on the internal file in a form analogous to the PUF variable. The unedited data contains VOTHER1 (responses to the original question about non-wage income) and VOTHER2 (responses to the follow-up “are you sure?” question).

file can be used with MOVGRP (mover group), which is available in the RMOV (recent mover) data file to determine to which mover group each person belongs.

HOMIMP: This file includes information on the home improvement projects conducted on each housing unit. In the PUF, the data are one record for each home improvement job reported, whereas on the IUF the data are one record per household with each job shown in their own variables.

RMOV: This file includes information on Recent Movers and is at the person level and is the same for both the IUF and PUF.

JTW: This file has information on an individual's journey to work or and is at the person level. It is the same for both the IUF and PUF.

MORTG: The information on mortgages is available on this file. It is at the household level and includes information on all owner-occupied homes with a mortgage. The only major differences between the PUF and the IUF for this file are topcodes.

RATIOV: This is a household level file that includes the ratio verification variables. There is no difference between the IUF and the PUF.

OWNER: This is a household level file with information on the on-site owner or manager and there is no difference between the PUF and the IUF.

TBLRCD: Many recodes are computed to create the tables released by the Census Bureau on the AHS data. These recoded variables are available in this file. It is a household level file. There is no PUF equivalent for it, as this file is only released with the IUF.

MASTER: This is a household level file, which is also unique to the IUF (no PUF equivalent). This file has sampling frame information on each housing unit and includes fields identifying detailed geography (down to the 1980 census tract level) for each unit.

NONTOPPUF: In years prior to 2007, this file included the recodes that are included on the PUF. Unlike on the PUF, these variables have not been topcoded and do not have values suppressed or masked. It should be noted that many of these variables or near equivalents appear on other IUF files. These variables have been incorporated in the NEWHOUSE file starting with the 2007 National IUF.

Geography Variables

The key geography variables, all on the MASTER file, are:

CBNCOD90	1990 central city/balance code
CBUR80	1980 centrl city/balnce/urban/rural code
CENSTATE	1960 Census state code
CMSA80	1980 consolidated MSA code
CMSA90	1990 consolidated MSA code
COOLDAY	Cooling degree days

COUNTY80	1980 FIPS county code
COUNTY90	1990 FIPS county code
FIPSTATE	FIPS state code
HEATDAY	Heating degree days
MCDCOD80	1980 design MCD/CCD code
MCDCOD90	1990 design MCD/CCD code
MSASTA80	MSA status – 80 def.
MSASTA90	MSA status – 90 def.
PLCODE80	1980 design Census place code
PLCODE90	1990 design Census place code
PMSA80	1980 design MSA/PMSA code
PMSA90	1990 design MSA/PMSA code
REGION	Census region
TRACT80	1980 tract code
TRCTSF80	1980 tract suffix
UACODE80	1980 design urbanized area code
UACODE90	1990 design urbanized area code
UASIZE90	1990 design urbanized area size
URBRUR80	1980 design urban/rural code
URBRUR90	1990 design urban/rural code
ZONE	Zone code (metro only)

Federal Information Processing Standards (FIPS) codes can be found at:
http://www.census.gov/geo/www/fips/fips.html
Metropolitan Statistical Area (MSA) codes can be found at:
http://www.census.gov/population/www/estimates/pastmetro.html

Race and Nativity

Nativity (country of birth) information has been collected since the 2001 national file, but appears on the PUF in aggregated form (e.g., Portugal and the Azores are grouped into a single category).

Coding of NATVTY (Country of birth) on the IUF

(57) United States
(72) Puerto Rico
(96) Outlying Area of the U.S. (American Samoa, Guam, U.S. Virgin Islands, Northern Marianas, Other U.S. Territory.)

(200) Afghanistan
(375) Argentina
(185) Armenia
(102) Austria
(501) Australia
(130) Azores
(333) Bahamas
(202) Bangladesh
(334) Barbados
(301) Canada
(206) Cambodia
(378) Chile
(311) Costa Rica
(207) China
(379) Colombia
(337) Cuba
(155) Czech Republic
(105) Czechoslovakia
(139) England
(417) Ethiopia
(507) Fiji
(108) Finland
(109) France
(110) Germany
(421) Ghana
(138) Great Britain
(116) Greece
(209) Hong Kong
(117) Hungary
(211) Indonesia
(210) India
(212) Iran
(213) Iraq
(119) Ireland/Eire
(214) Israel
(120) Italy

(310) Belize (106) Denmark (340) Grenada (343) Jamaica
(103) Belgium (338) Dominica (313) Guatemala (215) Japan
(300) Bermuda (339) Dominican Rep. (383) Guyana (216) Jordan
(376) Bolivia (380) Ecuador (342) Haiti
(377) Brazil (415) Egypt (126) Holland
(205) Burma (312) El Salvador (314) Honduras
(427) Kenya (127) Norway (449) South Africa (242) Vietnam
(218) Korea/S. Korea(229) Pakistan (134) Spain (147) Yugoslavia
(221) Laos (253) Palestine (136) Sweden (353) Caribbean
(183) Latvia (317) Panama (137) Switzerland (318) Central America
(222) Lebanon (385) Peru (237) Syria (389) South America
(184) Lithuania (231) Philippines (238) Taiwan (304) North America
(224) Malaysia (128) Poland (239) Thailand (148) Europe
(315) Mexico (129) Portugal (351) Trinidad/Tobago (252) Middle East
(436) Morocco (132) Romania (240) Turkey (468) North Africa
(126) Netherlands (192) Russia (195) Ukraine (462) Other Africa
(514) New Zealand (233) Saudi Arabia (387) Uruguay (245) Asia
(316) Nicaragua (140) Scotland (180) USSR (527) Pacific Islands
(440) Nigeria (234) Singapore (388) Venezuela (555) Elsewhere
(142) Northern Ireland (156) Slovakia/Slovak Rep.

Race variables.

From N1997-M2002, RACE is coded identically on both the IUF and the PUF. Beginning in 2003, respondents were allowed to choose “one or more” races. The IUF stores these answers in RACE1 (first race mentioned) through RACE5 (last race mentioned). The PUF includes only the variables RACE, which recodes these answers into 21 categories such as “White only.”

Coding of RACE1-RACE5 on the IUF

(1) White
(2) Black or African American
(3) American Indian or Alaska Native
(4) Asian
(5) Native Hawaiian or Other Pacific Islander
(6) Other - DO NOT READ

Recodes Files

The TBLRCD file contains the recodes used internally to produce the AHS publication tables. Most of these variables are fairly simple recodes, but may be of use to users seeking to reproduce the AHS publication tables. These variables may allow users to save some effort and avoid “reinventing the wheel.” Three variables that are based on especially complex calculations, and hence probably of the most interest to users, are:

OTPINR	Outstanding principal and interest (rnd)
POORR	Hhld income as % of povery level (rnd)
POVLVL	Poverty value that corr. to lookup table
ZSMHCM	Monthly housing costs w/ maintenance

ZSMHCN Monthly housing costs w/o maintenance

Sampling Variables - National Files

The AHS is a complex random sample. Standard errors calculated from the data will be larger than those calculated using formulas that assume a simple random sample. This section discusses the key features of the AHS sample design, and discusses the IUF variables that correspond to those features.

The AHS sample is stratified in two ways: by Primary Sampling Unit (PSU) and by sample frame. A PSU is a county or a group of counties. The AHS sample was drawn by dividing the country into PSUs and then randomly sampling PSUs. More specifically, large PSUs (called “self-representing” PSUs) were drawn with certainty. Smaller PSUs (“non-self-representing”) were divided into strata based on geography and characteristics from the 1980 census.[2] Finally one PSU was randomly chosen from each strata.

A difficulty with this scheme is that there is only one PSU per stratum, while at least two PSUs per stratum are required to calculate the standard errors. Hence, the Census Bureau has combined pairs (sometimes triplets) of PSUs into “pseudo-strata” of similar PSUs for the purpose of calculating SEs.

Within each PSU, housing units were randomly chosen from four sample frames (lists of housing units). The 1980 census frame contains housing units constructed before 1980. The permit frame contains housing units built since 1980 in areas where a building permit is required to authorize construction. The special areas frame contains housing units built since 1980 outside of permit-issuing areas, a small number of rural areas. Finally the group quarters frame contains units in group quarters. These group quarters units are not considered housing units and are not in the interviewed sample. A sample is drawn from these units nonetheless, because some of these units may later be converted to housing units.

The key variables are:

STRPSU80 1980 design stratification PSU

SEGMTYPE Can be used to identify which sample frame unit comes from

Frame defined using SEGMTYPE ==>

Permit - SEGMTYPE = 4,13
GQ - SEGMTYPE = 12
Special Area - SEGMTYPE = 6,10
Unit - SEGMTYPE = 1,2,3,7,8,9,11

SEGMNT80 The last 3 digits can also be used to identify which sample frame unit comes from

Frame defined using SEGMNT80 ==>

Permit - 001-249
GQ - 250-274

[2] For a more detailed description of the AHS sample design and formation of the strata, see “American Housing Survey, A Quality Profile,” Current Housing Reports H121/95-1, available at http://www.census.gov/prod/www/abs/cons-hou.html .

Special Area - 275-299
Unit - 300-999

STRPSU80	Pseudo-strata
PSUTYP80	1980 design PSU type (self-representing/non-self-representing). Has codes 1 to 5 (Note: Toni is checking to see why there are 5 codes and what they mean.)

In a statistical procedure such as SAS proc means, a user would specify strata as selfrepresenting*frame*pseudostrata, and specify cluster (within strata) as PSU80.

Note that the AHS national files from 1985-present (currently 2009) are based on a 1980 sample design. 1990 and 2000 sample design variables are also included on the file, for internal use (i.e., for field staff), but played no role in drawing the sample.

Contact Information

Census Bureau AHS staff: (301) 763-3235 or 1-888-518-7365 or by email at mailto:ahsn@census.gov.
HUD AHS staff: (202) 708-3178 or 1-800-245-2691.