AMERICAN HOUSING SURVEY

Disclosure Avoidance Techniques: 1985 to 2013

LAST UPDATED: MARCH 2020

U.S. DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT
U.S. CENSUS BUREAU
U.S. DEPARTMENT OF COMMERCE

Contents

List of Exhibits

1. Overview

The purpose of this document is to explain how the Department of Housing and Urban Development (HUD) and the U.S. Census Bureau (Census Bureau) applied disclosure avoidance techniques to the American Housing Survey (AHS) for 1985 to 2013.

Title 13, Section 9 of the United States Code (U.S.C.) requires the Census Bureau to keep confidential the information collected from the public under the authority of Title 13, under which the AHS data are collected. Disclosure avoidance is the process of protecting the confidentiality of data. A disclosure of data occurs when someone can use published statistical information to identify an individual who has provided confidential information.

All AHS data products released to the public are first reviewed by the Census Bureau Disclosure Review Board (DRB) to ensure that no identifiable Title 13 data are or may be disclosed. If the DRB determines that the requested statistical product does or reasonably could result in such disclosure, then the data product will be modified prior to approval for the public. Increased prevalence of administrative records and disclosure research in recent years has led HUD and the Census Bureau to take increasingly strict measures to protect the data from re-identification.

For more information on disclosure avoidance techniques for 2015 and later years of the AHS, see *Disclosure Avoidance Techniques: 2015 and Beyond.*

2. The American Housing Survey's Vulnerability to Disclosure

For each year of the AHS, HUD and the Census Bureau produce two microdata products that contain individual responses to survey questions: the internal use file (IUF) and the public use file (PUF). The IUF contains all the individual responses as provided by the respondent and detailed geographic information (for example, census block and parcel number).

The purpose of producing the PUF is so users can conduct their own statistical analysis, including summary statistics and regression modeling. Compared with the IUF, the PUF is altered in numerous ways to avoid the disclosure of a respondent's name or address. Generally speaking, there are three types of disclosure we try to avoid:

1. **Direct disclosure of a respondent's name or address:** Including a respondent's name or address would be a clear violation of confidentiality.
2. **Indirect disclosure of a respondent's address through disclosure of detailed spatial information:** Including precise spatial information such as census block, or inherently spatial information such as distance to water, could result in an indirect disclosure of the respondent's address.
3. **Re-identification of a respondent's name or address via a re-identification attack:** A re-identification attack occurs when an attacker matches an external data source with precise name or address information to the individual PUF responses using information common to both datasets. The AHS is vulnerable to this type of attack due to the large number of housing attributes included in the survey.

Re-identification can also occur through what is referred to as a "database reconstruction-abetted re-identification attack." This type of attack occurs when an attacker is able to reconstruct individual IUF records using the summary table estimates, which themselves are derived from the IUF. The feasibility of this type of attack increases as the number of published summary table estimates increases.

From 1985–2009, HUD and the Census Bureau published only a limited set of summary tables (approximately 50 per survey). For these years, the likelihood of reconstruction of the IUFs using summary table estimates is de minimis.

For 2011 and 2013, HUD and the Census Bureau published summary table estimates through the AHS Table Creator. The number of published summary tables increased to approximately 150,000 per survey year. Moreover, the AHS Table Creator summary table estimates for smaller geographic areas than what was available from 1985–2009. To guard against database reconstruction attacks, numerous disclosure avoidance techniques are applied to the summary table estimates in the AHS Table Creator.

AHS users should note that summary estimates derived using the PUF may not match summary table estimates. This mismatch is because of the disclosure avoidance techniques applied to the PUF.

Section 3 of this document details the disclosure avoidance technique applied to the summary table estimates for 1985–2009. Section 4 of this document details the disclosure avoidance technique applied to the summary table estimates for 2011 and 2013. Section 5 of this document details the disclosure avoidance techniques applied to the PUF.

3. Disclosure Avoidance Techniques Applied to Summary Table Estimates for 1985–2009

Summary table estimates for 1985–2009 are available in print publications, PDF documents, and Excel files (2005–2009). Three types of disclosure avoidance techniques are applied to summary table estimates: population thresholds and rounding.

3.1. Geographic Area Population Threshold

For 2009 and earlier, summary table estimates were not created for geographic areas with less than 100,000 persons. Historically, the Census Bureau published summary table estimates for small geographic areas (cities, counties, tract) for other surveys, including the American Community Survey. Despite it not being clear as to why the Census Bureau decided not to publish AHS summary table estimates for geographic areas with less than 100,000 persons, this was a “rule” for the AHS.

3.2. Rounding

There are three types of summary table estimates found in the summary tables: housing unit counts (in thousands), means, and medians. For published tables for years 2009 and earlier, rounding occurred based on what was being measured.

- Medians in dollars are rounded to the nearest dollar.
- Medians in feet are rounded to the nearest foot.
- Medians in years are rounded to the nearest year.
- Medians for percentages, ratios, and rates are rounded to the nearest tenth.

Estimates of monthly housing costs as a percentage of current income were computed separately for each unit and rounded to the nearest percentage.

4. Disclosure Avoidance Techniques Applied to Summary Table Estimates: 2011 and 2013

Summary table estimates for 2011 and 2013 are available in the AHS Table Creator. The AHS Table Creator allows for customized tables, which are presented in a format that is easy to understand and read. The Table Creator dramatically expanded the number of summary tables available, from 50 per survey to more than 150,000 per survey. Because of the dramatic expansion of summary estimate availability, additional suppression was integrated into the Table Creator to help prevent database reconstruction attacks. At the same time, the Table Creator enabled HUD and the Census Bureau to easily produce estimates for geographic areas of less than 100,000 persons.

There are two types of disclosure avoidance techniques applied to summary table estimates: cell suppression and rounding.

4.1. Cell Suppression

Within Table Creator, suppressed cells are displayed with an “S.” Suppression rules apply when an estimate is based on less than three unweighted AHS observations while one of the two conditions below is present:

- When an AHS Table Creator estimate is based on a variable available only on the internal use file (IUF) (that is, including numerous geographic indicators).
- When an AHS Table Creator estimate is based on a variable available from the public use file (PUF) but is cross-tabulated with a column variable by-group that is based on a variable available only on the IUF.

When suppression rules apply, they apply to more than just a single estimate; they also apply to any other estimate that has a “parent” or “child” relationship to the suppressed estimate. Parent indicators are row (column) indicators that have rows (columns) indented under them. Child indicators are the indented rows (columns) that, when added together, sum up to the parent row (column). To do this, mutually exclusive indicators within each table stub were grouped according to parent/child relationships to identify which rows (columns) were “related” to one another. From this, related cells were flagged as requiring suppression to prevent multidimensional disclosure (by subtraction) of any other cells within the group where at least one of the cells had an unweighted count of less than 3.

Additionally, for all means and medians except interpolated medians (for example, Year Structure Built), when a mean or median cell count is less than 10, the cell is suppressed, and any replicated indicators are suppressed as well. Interpolated medians and means have a suppression threshold of 3.

4.2. Rounding

Within Table Creator, all housing unit count estimates, including margins of error, are rounded to the nearest thousand. All means and medians, including margins of error for means and medians, are rounded to four significant digits.

5. Disclosure Avoidance Techniques Applied to the Internal Use File to Create the Public Use File, 1985–2013

As noted in Section 2, the public use file (PUF) is altered in numerous ways to protect against disclosure. PUFs can be used to create custom tabulations, enabling users to delve further into the rich detail collected in the AHS. To help AHS users navigate the PUF, the Census Bureau posted the *AHS Codebook* online.

Disclosure avoidance techniques are applied to numerous variables in the PUF. The *AHS Codebook* includes the “Disclosure” field that lists the specific disclosure technique applied to the PUF variable.

The subsections below describe the disclosure avoidance techniques applied to the PUF.

5.1. Removal of Personally Identifiable Information Variables

Variables that directly identify a housing unit or person are withheld from the PUF. These variables include name, address, phone number, latitude, longitude, and parcel number.

5.2. Pseudocoding, Alteration, and Suppression of Political or Census Geographic Variables[1]

Pseudocoding, alteration, and suppression were applied to geographic variables to ensure that areas with a population of less than 100,000 cannot be identified on the PUF. This rule is commonly referred to as the “100,000 persons” rule and was applicable to microdata for all Census Bureau surveys.

Exhibit 5.2.1 below lists the impacted geographic indicators and the corresponding actions taken for the national longitudinal sample PUF. Exhibit 5.2.2 does the same for the independent metropolitan area longitudinal oversample (herinafter reffered to as the metropolitan area sample).

Exhibit 5.2.1. National Longitudinal Sample Public Use File Disclosure Avoidance Techniques for Geographic Variables

Geographic Indicator	Disclosure Avoidance Technique
SMSA	Suppression: All sample cases in *SMSAs* where the population was less than 100,000 or outside of *SMSAs* (nonmetro) were given a value of 9999.
SMSA/CMSA	Suppression: Cases in *SMSAs* where the rural population was less than 100,000 were given a value of 9999. Suppression: Cases in *SMSAs* where the non-central city population was less than 100,000 were given a value of 9999.
SMSA	Pseudocode: Some cases in *SMSAs* in the Chicago, New York, and northern New Jersey areas were pseudocoded to reflect their location within the general metropolitan area, but not within a specific PMSA. These have *SMSA* values of 9991, 9992, or 9993.
METRO	Alteration: For these *SMSAs*, cases where *METRO* = 3 or 4 have been altered to *METRO* = 2. In some SMSAs, all cases were coded to *METRO* = 1 or *METRO* = 2.
METRO3	Alteration: For these *SMSAs*, cases that are truly rural (*METRO3* = 3) have been altered to (METRO3 = 2).
DEGREE	Alteration: Some cases had their *DEGREE* value altered by replacing the true value (1–6) with a value that is as close to the true value as possible without violating confidentiality restrictions.

[1] For more details on geography in the PUFs, see the document: AHS PUF Geography: 1985-2013.

Exhibit 5.2.2. Independent Metropolitan Area Sample Public Use File Disclosure Avoidance Techniques for Geographic Variables2

Geographic Indicator	Disclosure Avoidance Technique
METRO	Alteration or Suppression: Some values of *METRO* have been altered or suppressed.
ZONE[3]	Alteration or Suppression: Some values of *ZONE* have been altered or suppressed.
STATE	Suppression: Cases where a *ZONE* spans multiple states have a suppressed STATE code of 99.
COUNTY	Pseudocode: When a specific county cannot be disclosed, it is combined with other counties to form a pseudocounty. If a *COUNTY* code is above 840, it is pseudocoded. The full list of *COUNTY* pseudocodes and what they represent can be found in the AHS Codebook.

Caution should be taken when using standard metropolitan statistical area (SMSA). Due to how suppression and pseudocoding were applied, areas within an *SMSA* but in an area with fewer than 100,000 are flagged as "9999." Thus, cases are not missing at random from the *SMSAs*.

The geographic vintages applied to the metropolitan area have changed over time as OMB issued new geographic definitions (for example, 1983 metropolitan areas versus 2003 metropolitan areas.

5.3. Topcoding and Bottomcoding

Topcoding is a disclosure limitation technique that involves limiting the maximum value of a variable allowed on the file to prevent disclosure of units with extreme values in a distribution (for example, outliers).

Top and bottom coded variables were edited up or down to a point determined by the top-coding rules. These rules vary by variable and year. Top code levels are provided in spreadsheet documents for survey years 2003 to 2013 on their respective landing page (Exhibit 5.3.1). Information on topcodes for earlier years are available in the appropriate codebook.

Exhibit 5.3.1. Sample Topcoding Information Available for 2015 American Housing Survey National Longitudinal Sample

Name	Topcode Level	Bottomcode	Number with Values	Value Next to Maximum	Maximum	Frequency of Maximum
LPRICE	97th Percentile	NO	30,256	773,000	1,314,181	426

To preserve confidentiality, the Census Bureau Disclosure Review Board policy is that there must be at least three cases included in the calculation of a mean. It is not unusual in the AHS PUF, particularly in the AHS Metropolitan Sample PUF, for a variable's universe of cases to be so small that there is not a minimum of three cases greater than or equal to the topcode predetermined or calculated for that variable. In these instances, the value of the topcode is lowered until there are at least three cases that

[2] For detailed information regarding each AHS metropolitan area and the geographic boundaries that were used during these years, see Metropolitan Area Oversample Histories: 1973 to 2013.

[3] AHS Zones are concept functionally like the U.S. Census Bureau's Public Use Micro Area (or PUMA). They were created by HUD to identify smaller geographic areas within each metropolitan area that complied with the "100,000 persons" rule.

can be included in the calculation of the mean. In the rare instances where there are not three eligible cases in the universe for a variable, all applicable values are set to a not reported code.

5.4. Rounding

Rounding reduces the number of unique values in the data and protects against rare-event situations. Exhibit 5.4.1 lists the rounding rules applied to the AHS and the internal use file (IUF) variables subject to rounding in the PUF. All variables that are created from or edited against variables that are rounded are re-calculated or re-edited following rounding.

Exhibit 5.4.1. Variables Rounded on the Public Use File

PUF Variable	Years	Rounding Rule
PMT, PMT2-4	2013	Nearest $25 value
AMMORT, AMMORT2-4, CPRICE, LPRICE, UNPBAL, UNPBAL2-4, VALUE	2013	Nearest $10,000 and values rounded to $0 were set to $1
AMTX	1997 to 2013	Nearest integer in the sequence 5, 15, 25, 35, 45 before being top coded at the 95.5th percentile
CONFEE	1997 to 2013	Nearest integer in the sequence 50, 150, 250, 350 45 before being top coded at the 97th percentile

5.5. Collapsing

Collapsing of categorical variables into more generalized categories was done to protect against rare-event situations. Collapsing was done by reducing the number of categories in the variable on the PUF. Collapsing occurred for four PUF variables: *NATVTY*, *HHNATVTY*, *RACE*, and *HHRACE*.

5.6. Perturbation

Some AHS PUF variables that represent a year or number of years are perturbed, or slightly altered in a non-random way, to protect against rare-event situations. *AGE* is perturbed and other variables are re-edited around it to preserve confidentiality and consistency with other demographic variables.

U.S. Department of Housing and Urban Development

Office of Policy Development and Research

Washington, DC 20410-6000

AMERICAN HOUSING SURVEY

March 2020