

Research And Methodology Directorate

Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples

By Laura McKenna
Issued April 2019



CONTENTS

| | |
|---|----|
| Introduction | 3 |
| Microdata | 4 |
| Disclosure Avoidance Methods for Microdata | 4 |
| Removing Information to Protect Microdata | 4 |
| Altering Information to Protect Microdata | 5 |
| 1960 | 6 |
| PUMS Data | 6 |
| DA Techniques | 6 |
| 1970 | 6 |
| PUMS Data | 6 |
| DA Techniques | 6 |
| 1980 | 6 |
| PUMS Data | 6 |
| DA Techniques | 7 |
| 1990 | 7 |
| PUMS Data | 7 |
| DA Techniques | 7 |
| 2000 | 8 |
| PUMS Data | 8 |
| DA Techniques | 8 |
| 2010 | 9 |
| PUMS Data | 9 |
| DA Techniques | 9 |
| Conclusion | 9 |
| References | 9 |
| Appendix A | 11 |
| Appendix B | 13 |

INTRODUCTION¹

The U.S. Census Bureau conducts the decennial censuses under Title 13, U.S. Code, Section 9 mandate to not “use the information furnished under the provisions of this title for any purpose other than the statistical purposes for which it is supplied; or make any publication whereby the data furnished by any particular establishment or individual under this title can be identified; or permit anyone other than the sworn officers and employees of the Department or bureau or agency thereof to examine the individual reports (13 U.S.C. § 9 (2007)).” The Census Bureau applies disclosure avoidance (DA) techniques to its publicly released statistical products in order to protect the confidentiality of its respondents and their data.

Different DA procedures were used for the 1960, 1970, 1980, 1990, 2000, and 2010 decennial censuses’ Public Use Microdata Samples (PUMS). This paper summarizes these historical methods in order to put the ongoing DA modernization effort in context. This history of decennial census

¹ This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

disclosure avoidance methods discusses only publicly acknowledged confidentiality edits as noted in official documentation. All of the information in this summary was taken from historical public sources, except as noted. None of the information in this paper is confidential.

There is minimal public documentation of the disclosure avoidance methods used in the 1960 Census. There is no discussion of disclosure avoidance for group quarters (GQ) data in public or internal documents for the 1960, 1970, 1980, 1990, and 2000 Censuses, but the 2010 Census has an additional subsection for that purpose.² This paper is focused on microdata files from the censuses. The American Community Survey (ACS) is out of scope.

This history gleans procedures from various types of PUMS that differed in terms of sample size, geographic thresholds, short-form (100 percent) data vs. long-form (sample data), and universe. All publications were based on both people in households and people in GQ.

² GQ data include information about people living in nursing homes, prisons, college dormitories, and military barracks (somewhere other than a household).

MICRODATA

Statisticians use the term microdata to refer to any record-level data. At the Census Bureau, the term microdata has a narrower definition: it refers to collected data that have been cleaned, edited, and sometimes imputed so that they can be used to produce statistical tabulations and analyses. These data are presented at the record level. A microdata file consists of data at the respondent level, as opposed to aggregate counts or magnitudes. Each record represents one respondent, such as a person or household, and consists of values of characteristic variables for this respondent. Typical variables for a person-level demographic microdata file are age, race, sex, and income, and a household-level file might include mortgage payment/rent, year house built, and source of heat. Microdata files may include hundreds of such variables for each respondent.

The Census Bureau publicly releases microdata files from the decennial census and from many of its demographic and economic surveys. This paper focuses on those from previous decennial censuses. The PUMS from a decennial census is different from that of most surveys (with the exception of the ACS). The difference is due to the fact that the PUMS from the decennial census and the ACS do not contain records from each respondent. They contain records from a sample of their respondents that can be released with an underlying layer of uncertainty. The uncertainty exists from the inability to discern whether or not an individual respondent is captured in the PUMS files. This creates a scenario where a record with a unique combination of certain variables in the PUMS may not necessarily represent a unique person or household in the population (decennial census) or full sample (ACS). Microdata files from other demographic surveys contain records for all respondents.

First steps in minimizing the risk of unauthorized disclosure of microdata include removing direct identifiers such as names, addresses, and Social Security numbers. High-risk records (e.g., individuals with very large incomes or unusual jobs) are identified to ensure their visibility within the file is decreased. Other characteristics are considered for their

uniqueness and their contribution to any increase in reidentification (disclosure) risk.

DISCLOSURE AVOIDANCE METHODS FOR MICRODATA

For any given microdata file, the Census Bureau has used a combination of the techniques described below.

Removing Information to Protect Microdata

Remove Direct Identifiers

Beginning with the obvious, the Census Bureau removes direct identifiers such as name, address, and telephone number.

Topcoding and Bottom-Coding

Topcoding and bottom-coding are used to eliminate outliers in a file. They are used for continuous variables such as age and dollar amounts. When topcoding, the top 0.5 percent of all values or the top 3.0 percent of all nonzero values (whichever effects the least amount of records) are cut off. They can be replaced with the topcode (cut off) value, or the mean or interpolated median of all topcoded values. At least three values must be included in the topcode or it will be lowered to meet that threshold. Bottom-codes are the same except on the other end of the distribution. An example of a bottom-coded variable might be the year that a building was built or gross income. For variables that are part of a sum, the individual parts are topcoded (or bottom-coded) prior to their summation.

Recoding and Rounding

Recoding is done for categorical and continuous variables. Each category of a variable must contain nationwide at least 10,000 weighted people or households, depending on the universe of the table. Otherwise, the category must be combined with another until the threshold rule is met. For continuous data values that the Census Bureau knows are public information (such as property taxes which has its own recoding scheme) and for some dollar amounts, recoding is also applied.

Other dollar amounts may follow one of two rounding/recoding schemes.

Round to the nearest two significant digits, or use this recoding scheme:

- Zero rounds to zero.
- 1 to 7 rounds to 4.
- 8 to 999 rounds to the nearest multiple of 10.
- 1,000 to 49,999 rounds to the nearest multiple of 100.
- 50,000 and greater rounds to the nearest multiple of 1,000.

Any totals or other derivations are calculated using the rounded numbers.

Geographic Population Thresholds

All geographic areas identified on PUMS must have a weighted population of 100,000 or more. When figuring out the population of an identified area, all geography-related variables on the file must be cross-tabulated to obtain the final population count. For example, other geographic variables may be urban/rural, Metropolitan Statistical Area (MSA) status, and other geographic areas named such as Congressional District. All geographic pieces identified after crossing all geographic variables must meet the required threshold for that PUMS.

Altering Information to Protect Microdata

Data swapping, the generation of partially synthetic data, and noise infusion are current methods for the protection of frequency count data from the decennial census and ACS. While the three methods are used mainly to protect tables for very small geographic areas, they are performed on the underlying microdata before tabulation. The PUMS files are sampled from the altered data.

Data Swapping for Household Data

The purpose of any swapping methodology is to introduce uncertainty into the data so that the data user doesn't know whether real data values correspond to certain respondents. Household records with a high risk of disclosure are typically identified through software and called uniques because they have a unique combination of certain variables. The unique records are targeted for data swapping. In the swapping procedure, a small percentage of records are matched with other records in the same file on a set of predetermined variables used as swapping attributes. A set of other variables are then swapped between the two records without disturbing the responses for nonsensitive and nonidentifying fields.

The variables may be continuous or categorical. A household record is typically swapped with another household within a large area but in a different smaller area within the larger one, for example, across tracts but within the same county. Again, the swapping technique is targeted to protect frequency count tables from censuses and ACS, but the PUMS files are sampled from the swapped data, and this adds a small amount of confidentiality protection (Zayatz, 2002 and 2003) to the microdata.

Partially Synthetic Data

Applying data swapping to GQ data does not work well. Imagine swapping a nursing home (or someone who lives there) with a college dorm (or someone who lives there). The resulting data would make no sense, so the Census Bureau relies on the generation of partially synthetic data to protect GQ data from the decennial census and ACS.

The original data are modeled using a general linearized model. The process then continues with identifying unique records by cross-tabulating certain values and flagging records in the resulting cells with a count of one. Because these are GQ data, the uniques represent people rather than households. Those variable values that are causing the disclosure risk problem in a given unique record are then blanked and replaced with values generated from the model. Geography and type of GQ are never altered, and the numbers of people of less than age 18 and age 18 or more are never changed. Occasionally, a modeled (simulated) value may coincidentally be the same as the original value. Again, the partially synthetic data generation technique is targeted to protect frequency count tables from censuses and ACS, but the PUMS files are sampled from the partially synthetic data, and this adds a small amount of protection to the microdata.

Noise Infusion

At this time, noise infusion is not widely used for the protection of microdata. It is used to hide very unusual characteristics of a person or household at a given point in time that is not caught by the 10,000 threshold rule for individual categories described above. For example, consider a person who gave birth to seven children at one time, or a person who is a practicing physician at the age of 15 (both very unusual circumstances that would probably be in the news). Also very large households may present a disclosure risk. Editing procedures capture and alter many, but not all, of these unusual occurrences. The

Census Bureau does not publicly describe precisely how noise is added to protect this type of data.

1960

PUMS Data

Decennial censuses gather information from questions asked of the entire population, or from those same questions, as well as many others, asked of only a sample of the population. Those questions asked about every person and household are called 100 percent (or short-form) questions. The other questions are called sample (or long-form) questions. In the 1960 Census, 1 in 4 households received the long-form questionnaire.

The Census Bureau was the first statistical agency to publicly release microdata files (Ruggles, 2013). In 1962, the Census Bureau drew a sample of the long-form data records that would represent 1.0 percent of the population nationwide. The Census Bureau published two microdata files in the form of punch cards from those records, both using a geographic population threshold of 250,000 for each area identified. Areas that were identified on the PUMS files were called Public Use Microdata Areas (PUMAs). Areas could not cross state lines. The first file contained records from a 1-in-1000 sample of the population, and the second contained a 1-in-10,000 sample of the population. The second file was a subset of the first file. The smaller file was published for data users who may not have had the computer power or the need to process the larger file. In 1973, DualLabs published the records from the full 1 percent sample, which was recoded to match the record layout and categories of the 1970 public-use samples, <<http://users.hist.umn.edu/~ruggles/JSM2005-000189.pdf>>. The files contained person-level and household-level information, with persons linked to their households. Demographic data users were immensely pleased to have these files because they gave researchers the ability to retabulate and manipulate data without constraints imposed by a fixed set of predefined, printed tables, <www.icpsr.umich.edu/icpsrweb/icpsr/series/13>.

DA Techniques

The only DA techniques used for these files were the removal of direct identifiers and a geographic population threshold of 250,000.

1970

PUMS Data

In 1970, there were two long forms with some overlapping questions and some different questions. One long-form questionnaire was sent to 15 percent of U.S. households and GQ individuals, and the second was sent to 5 percent of households and GQ individuals. Six PUMS files were released from the 1970 Census. See Appendix A for an illustration. For both the 15 percent and the 5 percent long-form data, three PUMS files were released for different types of geographic areas: a file based on areas within a state, a file based on county groups mainly determined by MSAs, which can cross state lines, and a file based on neighborhood characteristics, <https://usa.ipums.org/usa/resources/codebooks/1970_pums_codebook.pdf>.

All six PUMS files were based on stratified samples of the two long-form datasets and were nonoverlapping in terms of households and people. They each contained data on 1 percent of the population nationwide. They were self-weighting. Each person or household had a weight of 100. For all six PUMS files, subsamples were drawn that represented 0.1 percent and 0.01 percent of the population to accommodate users who could only handle smaller files.

DA Techniques

All direct identifiers were removed from the files. A geographic population threshold of 250,000 per identified area was imposed and for the neighborhood characteristics files, the only geographic areas directly identified were census region and census division due to the fact that neighborhood characteristics can divide geographic areas into smaller pieces.

1980

PUMS Data

In 1980, there was just one long form that was sent to approximately 1 in 5 households. Individuals living in GQ and vacant units were also sampled. PUMS files included a 5 percent, state-based file (Sample A); a 1 percent, MSA-based file (Sample B); and a 1 percent, state-by-urban/rural-based file (Sample C). See Appendix B for a summarization of these samples and a comparison between the 1970 and 1980 PUMS. All three 1980 PUMS files were based on stratified samples of the long-form dataset, and were nonoverlapping in terms of households and people. They were self-weighting. Each person or household in the 1 percent files had a weight of 100, and each person in the 5 percent file had a weight of 20. The files had the same subject content. For users desiring smaller files, a subsample of 0.1 percent of

the population was also released for each of the three files, <<https://www2.census.gov/prod2/decennial/documents/D1-D80-PUMS-14-tech.pdf/>>.

All missing data values were allocated (imputed) and allocation flags for each variable were included in the PUMS. Washington, DC, and Puerto Rico were treated as states.

DA Techniques

All direct identifiers were removed from the files. Income was grouped into \$10 intervals and was topcoded at \$75,000, and age was topcoded at 90.

A geographic population threshold of 100,000 per identified area (PUMA) was imposed. PUMAs were not always contiguous. PUMAs in the state-based file (Sample A) could not cross state boundaries. Many PUMAs in the MSA-based file (Sample B) did cross state boundaries. PUMAs in the state by urban-rural file (Sample C) sometimes had to combine states. The PUMAs for this file consisted of the cross tabulation of state by urban/rural designation. If there was not enough population designated as urban or rural in a given state to meet the 100,000 population threshold for a PUMA, that state was combined with another. Region and division boundaries were not crossed.

1990

PUMS Data

In 1990, there was just one long form that was sent to approximately 16 percent of all U.S. households. Individuals living in GQ and vacant units were also sampled. People sampled from within the same GQ were not identifiable as such. PUMS files included a 5 percent, state-based file (Sample A); a 1 percent, MSA-based file (Sample B); and a 3 percent, elderly file for households with at least one person of age 60 or more (Sample C). All three PUMS files were based on stratified samples of the long-form dataset and were nonoverlapping in terms of households and people. Each household and person record was assigned its own weight. The files had the same subject content. For users desiring smaller files, a subsample of 0.1 percent of the population was also released for each of the three files, <https://www2.census.gov/prod2/decennial/documents/D1-D90-PUMS-14-techm.pdf>. Washington, DC, and Puerto Rico were treated as states.

In 1990, three different household sampling rates were used for the long form: 1 in 8, 1 in 6, and 1 in 2 for an overall average of about 1 in 6. For GQ, there was only one sampling rate of 1 in 6, <www.census.gov

</history/pdf/1990proceduralhistory.pdf>>. The variable sampling rates were used to arrive at high-quality estimates for tables published for small geographic areas and to decrease respondent burden for densely populated areas. The rates were based on precensus population estimates of incorporated places, census tracts, and block-numbering areas.

All missing data values were allocated (imputed), and allocation flags for each value were included in the PUMS. "In rare instances during the implementation of the sample weighting process, the sample size was considered inadequate to make estimates of sample data. In collection block groups with a designated 1-in-2 sampling rate, augmentation was employed if the 100 percent housing unit count was at least six and the observed sampling rate was less than 1 in 12. In collection block groups with a designated 1-in-6 or 1-in-8 sampling rate, augmentation was employed if the 100 percent, housing unit count was at least 12 and the observed sampling rate was less than 1-in-30. Augmentation was performed separately for GQ persons using the same criteria as for the 1-in-6 or 1-in-8 designated sampling rates. Augmentation was achieved by selecting a sample of housing units (or GQ persons) to increase the observed sampling rates to at least 1 in 12 or 1 in 30. Using the 100 percent characteristics, the selected households (or GQ persons) were matched by a hot deck procedure to similar housing units (or GQ persons) with sample data. The sample data were then copied to the augmented housing units (or GQ persons). The augmentation rate was very small. Most augmentation occurred for GQ persons," <https://www2.census.gov/prod2/decennial/documents/D1-D90-PUMS-14-techm.pdf>. Augmentation is sometimes referred to as whole household imputation or, for GQ, whole person imputation.

DA Techniques

All direct identifiers were removed from the files.

The Census Bureau limited the detail on files by using recodes and topcodes for place of residence, place of work, type of GQ, income values, age, and other selected items to further protect the confidentiality of the data. Most economic items were topcoded on a national basis. The topcode was set at either 0.5 percent of all values or 3 percent of all nonzero values, whichever was the larger of the two cutoff values. If a state had at least 30 cases above a given topcode, the state median of all topcoded values was released.

A geographic population threshold of 100,000 per identified area (PUMA) was imposed for all three samples. PUMAs were not always contiguous. PUMAs in the state-based file (Sample A) did not cross state boundaries. PUMAs in the MSA-based file (Sample B) often did cross state boundaries. The elderly file (Sample C) was produced for states only. Region and division boundaries were not crossed by any PUMAs.

A confidentiality edit was performed on the underlying 1990 sample data prior to publication of all data products. It was mainly used to protect data in tables that were published for very small geographic areas, but it also affected the PUMS files. An imputation methodology was used to provide DA for sample data in small block groups. This methodology involved the blanking of a sample of the data values (population and housing items) for one of the sample housing units in each small block group and imputing those values using the 1990 Census imputation methodology. This technique was known as Blank and Impute. Once sample data imputation was completed, the resulting sample data file (for which DA had been applied) was used to prepare all subsequent census sample data products. This data imputation methodology for providing DA for sample data added very little to the level of error of the estimates. A major reason for this is that the relative increase in imputation rates was very small (Griffin et al., 1989).

2000

PUMS Data

In 2000, there was just one long form that was sent to approximately 16 percent of all U.S. households. Individuals living in GQ and vacant units were also sampled. People sampled from within the same GQ were not identifiable as such. Households and GQ people in outlying areas, such as Guam and the U.S. Virgin Islands, all received the long form. The PUMS files included a 5 percent and a 1 percent file, both state-based. Both PUMS files were based on stratified samples of the long-form dataset, and were nonoverlapping in terms of households and people. Each household and person record was assigned its own weight, <www.census.gov/prod/cen2000/doc/pums.pdf>. Washington, DC, and Puerto Rico were treated as states. The 5 percent file identified PUMAs with a geographic population threshold of 100,000. The 1 percent file identified Super-PUMAs with a geographic population threshold of 400,000. The 1 percent file had much more variable detail (less recoding) than the 5 percent file, hence the larger areas. PUMAs and Super-PUMAs were not

always contiguous, and they did not cross state boundaries. There were also PUMS files for Guam and the U.S. Virgin Islands that contained records from 10 percent of the population and had the same level of detail as the 5 percent file.

As in 1990, three different household sampling rates were used for the long form: 1-in-8, 1-in-6, and 1-in-2 for an overall average of about 1-in-6. For GQ, there was only one sampling rate of 1 in 6. The variable sampling rates were used to arrive at high-quality estimates for tables published for small geographic areas and to decrease respondent burden for densely populated areas. The rates were based on precensus population estimates of incorporated places, census tracts, and block-numbering areas.

All missing data values were allocated (imputed), and allocation flags for each value were included in the PUMS. Also as in 1990, "In rare instances during the implementation of the sample weighting process, the sample size was considered inadequate to make estimates of sample data. In collection block groups with a designated 1-in-2 sampling rate, augmentation was employed if the 100 percent, housing unit count was at least six and the observed sampling rate was less than 1 in 12. In collection block groups with a designated 1-in-6 or 1-in-8 sampling rate, augmentation was employed if the 100 percent, housing unit count was at least 12 and the observed sampling rate was less than 1-in-30. Augmentation was performed separately for GQ persons using the same criteria as for the 1-in-6 or 1-in-8 designated sampling rates. Augmentation was achieved by selecting a sample of housing units (or GQ persons) to increase the observed sampling rates to at least 1 in 12 or 1 in 30. Using the 100 percent characteristics, the selected households (or GQ persons) were matched by a hot deck procedure to similar housing units (or GQ persons) with sample data. The sample data

were then copied to the augmented housing units (or GQ persons). The augmentation rate was very small. Most augmentation occurred for GQ persons," <<https://www2.census.gov/prod2/decennial/documents/D1-D90-PUMS-14-techm.pdf>>. Augmentation is sometimes referred to as whole household imputation or, for GQ, whole person imputation.

DA Techniques

All direct identifiers were removed from the files.

The Census Bureau limited the detail on files by using recodes, topcodes, and bottom codes for place of

residence, place of work, type of GQ, income values, age, and other selected items to further protect the confidentiality of the data. Most economic items were topcoded on a national basis. The topcode was set at either 0.5 percent of all values or 3.0 percent of all nonzero values, whichever was the larger of the two cutoff values. The topcode had to include at least three values in each state. If not, the topcode for a given state was lowered until the threshold was met. The mean of the topcoded values for each state was shown on the files.

Data swapping was performed on the underlying 2000 sample data prior to publication of all data products. It was mainly used to protect data in tables that were published for very small geographic areas, but it also affected the PUMS files. Once data swapping was completed, the resulting sample data file was used to prepare all subsequent census sample data products.

All categories of categorical variables on the file had to represent a nationwide universe of 10,000 weighted people. Another technique used to protect the data was noise infusion for large households. These techniques are described in detail in the section on Disclosure Avoidance Methods for Microdata.

2010

PUMS Data

In 2010, the long form had been replaced with the ACS. A PUMS file was still released from the 100 percent (short-form) data. The file was state-based and contained records for a systematic sample of 10 percent of the population in each state, Washington, DC, and Puerto Rico. Individuals living in GQ and vacant units were included. Persons sampled from within the same GQ were not identifiable as such. All missing data values were allocated (imputed), and allocation flags for each value were included in the PUMS. Each housing unit and person record was assigned a weight, <https://www2.census.gov/census_2010/12-stateside_pums/0tech_doc/2010%20pums%20technical%20documentation.pdf>. Whole household imputation and whole person imputation (for GQ) were performed.

DA Techniques

All direct identifiers were removed from the PUMS. All PUMAs had a geographic population threshold of 100,000. All categories of categorical variables on the file had to represent a nationwide universe of 10,000 weighted people. Other techniques used to protect the data included data swapping for

household data, partially synthetic data generation for GQ data, topcoding, bottom-coding, recoding, and noise infusion for large households. These techniques are described in detail in the section on Disclosure Avoidance Methods for Microdata.

CONCLUSION

The content of PUMS and DA techniques have evolved over the last six censuses. All of the PUMS files contained data on individuals living in both households and GQ. Direct identifiers were removed from all of the files, and they all had geographic population thresholds. Sample (long-form) data were used to create the 1960 through 2000 files, and 100 percent (short-form) data were used to create the 2010 file. Other DA techniques for the PUMS are summarized in the table on page 10.

Recently, the Census Bureau has embarked on an aggressive effort to replace its legacy DA methods with modern DA techniques based on formal privacy methods, <<https://privacytools.seas.harvard.edu/formal-privacy-models-and-title-13>>. Current methods will gradually change with the introduction of formal privacy (Nissim et al., 2018). Most of the current Census Bureau's DA research is focused on formal privacy for all types of data (Nissim et al., 2007). An algorithm operating on a private database of records satisfies formal privacy if its outputs are insensitive to the presence or absence of any single record in the input (Dwork, 2006). Census Bureau staff members are quickly learning about formal privacy and how it protects Census Bureau data products.

REFERENCES

- C. Dwork, "Differential Privacy," International Colloquium on Automata, Languages, and Programming (ICALP), 2006, pp. 1-12.
- R. Griffin, F. Navarro, and L. Flores-Baez, "Disclosure Avoidance for the 1990 Census," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1989, pp. 516-521.
- K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth Sensitivity and Sampling in Private Data Analysis," Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing, 2007, pp. 75-84.
- K. Nissim, T. Steinke, A. Wood, M. Altman, A. Bembenek, M. Bun, M. Gaboardi, D. O'Brien, and S. Vadhan, "Differential Privacy: A Primer for a Non-technical Audience (Preliminary Version), Harvard University Privacy Tools for Sharing Research Data, 2018, <<http://privacytools.seas.harvard.edu>>.

| Decennial censuses | Topcodes and Recodes | Blank and impute | Swapping | Category size thresholds | Noise infusion | Partially synthetic data |
|----------------------|----------------------|------------------|----------|--------------------------|----------------|--------------------------|
| 1960 | | | | | | |
| 1970 | | | | | | |
| 1980 | X | | | | | |
| 1990 | X | X | | | | |
| 2000 | X | | X | X | X | |
| 2010 | | | | | | |
| Households | X | | X | X | X | |
| Group quarters | X | | | X | | X |

S. Ruggles, "Big Microdata for Population Research," Minnesota Population Center, University of Minnesota, Working Paper No. 2013-04, 2013.

L. Zayatz, "SDC in the 2000 U.S. Decennial Census," In: Domingo-Ferrer, J. (eds) Inference Control in Statistical Databases, From Theory to Practice (Lecture Notes in Computer Science), Springer, Berlin, Heidelberg, 2002, vol. 2316.

L. Zayatz, "Disclosure Limitation for Census 2000 Tabular Data," Working Paper #15, Joint ECE/Eurostat work session on statistical data confidentiality, 2003, <www.unece.org/stats/documents/2003/04/confidentiality/wp.15.e.pdf>.

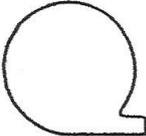
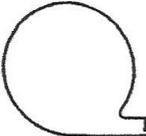
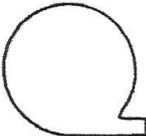
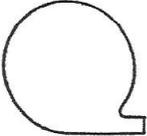
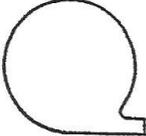
Appendix A

Public Use Samples of
Basic Records from the
1970 Census

Description and Technical
Documentation, p. 3

Prepared by the U.S. Census
Bureau, 1972

Figure 1. PUBLIC USE SAMPLES FROM 1960 AND 1970 CENSUSES

| | DATA CONTENTS | | |
|--|---|--|---|
| | 1970 15% Questionnaires (incl. 20% & 100% items) | 1970 5% Questionnaires (incl. 20% & 100% items) | 1960 25% Questionnaire (incl. 100% items) |
| <u>STATE Public Use Samples</u> Identify each State In larger States indicate Urban/Rural Metropolitan/Nonmetro Central City/Non CC Are available first |  |  |  |
| <u>COUNTY GROUP</u> <u>Public Use Samples</u> Identify all SMSA's over 250,000 pop. Identify related groups of counties elsewhere About 400 areas in all Do not identify urban or rural areas |  |  | |
| <u>NEIGHBORHOOD CHARACTERISTICS</u> <u>Public Use Samples</u> Identify only sections of the country: divisions Indicate rural, urban, & in- side urbanized areas (UA) the size of UA categories Households are associated by neighborhood Neighborhood character- istics include: Percent Negro Average household size |  |  | |

 Each tape symbol represents a separate one-in-a-hundred public use sample (30-33 tapes) from which one-in-a-thousand and one-in-ten-thousand subsamples (3 tapes and one tape respectively) will also be available.

Appendix B

1980 Census of Population
and Housing

Public Use Microdata Sample

Technical Documentation,
pp. 1 and 6

Prepared by the U.S. Census
Bureau, 1983

CHAPTER 1. INTRODUCTION

Overview

Public-use microdata samples are computer tapes which contain records for a sample of housing units, with information on the characteristics of each unit and the people in it. In order to protect the confidentiality of respondents, the Bureau excludes identifying information from the records. Within the limits of the sample size and geographic detail provided, these tapes permit users with special needs to prepare virtually any tabulations of the data they may desire.

Three separate public-use microdata samples are available, each representing five percent or one percent of the population and housing of the United States:

- o A Sample, 5%, identifying all States and various subdivisions within them, including most counties with 100,000 or more inhabitants;
- o B Sample, 1%, identifying all metropolitan territory and most SMSAs individually, and groups of counties elsewhere;
- o C Sample, 1%, identifying regions, divisions, and most States by type of area (urban/rural).

Three 1-in-1,000 samples are also prepared, one each extracted from the A, B, and C Samples.

Figure 2. Comparison of Features on 1980 and 1970 Public-Use Microdata Samples

| | -----1980 Samples----- | | | -----1970 Samples----- | | |
|------------------------------|------------------------|------------|------------|------------------------|--------------|--------------|
| | A | B | C | State | County Group | Neigh Chars |
| Sample Size | 5% 0.1% | 1% 0.1% | 1% 0.1% | 1-2% 0.1% | 1-2% 0.1% | 1-2% 0.1% |
| Areas Identified | | | | | | |
| Divisions | X | - | X | X | - | X |
| States | 51 | 20 | 28 | 51 | 4 | - |
| SMSAs of 100,000+ | 180 | 282 | - | - | 125 | - |
| Counties of 100,000+ | 350 | 236 | - | - | 104 | - |
| Places of 100,000+ | 123 | 135 | 58 | - | 12 | 5 |
| County Groups | 1154 | 1258 | - | - | 409 | - |
| Urbanized Areas | - | - | 73 | - | - | 6 |
| Metro/Nonmetro | - | X | - | 23 States | - | - |
| Urban/Rural | - | - | X | 42 States | - | X |
| Neighborhood Characteristics | - | - | - | - | - | X |