Chapter 3.
How To Use This File

INTRODUCTION

This chapter serves as a guide for data users to both the data files and the technical documentation. Novice users trying to understand how to use the documentation and the file should read this chapter first.

DATA FORMAT AND ACCESS TOOLS

The 2000 Public Use Microdata Sample (PUMS) data files are available in flat ASCII format. Users of the DVD/CD-ROM may access the PUMS data in two ways: with software and without software.

- The DVD/CD-ROM with software is designed to perform basic cross tabulations of any desired set of variables on the PUMS file.
- For the DVD/CD-ROM without software, users can utilize off-the-shelf standard statistical software packages to manipulate the data. (Also, files are available for downloading via FTP from the Census Bureau Web site.)

The 2000 PUMS are accompanied by electronic data dictionaries in a format that will allow the user to read in ASCII characters and prepare statements transforming the variables and their corresponding descriptions and values to the proper statements required by the software package of choice.

TECHNICAL DESCRIPTION

The 2000 PUMS file structure is hierarchical and contains two basic record types of 314 characters each: the housing unit record and the person record. The PUMS files are released in this format because of the tremendous amount of data contained in one record.

Each record has a unique identifier (serial number) that links the people in the housing unit to the proper housing unit record. The inclusion of the serial number on both record types affords the option of processing the data either sequentially or hierarchically. The file is sorted to maintain the relationship between both record types, so that a user does not have to be concerned about keeping the record sequence as the file was delivered. Each housing unit record is followed by a variable number of person records, one for each occupant. Vacant housing units will have no person record, and selected people in group quarters will have a pseudo housing record and a person record. The only types of group quarters that are identified are institutional and noninstitutional.

A housing unit weight appears on the housing unit record and a person weight appears on the person record. Weights allow users to produce estimates that closely approximate published data in other products.

Geographic identifiers and subsample identifiers appear only on the housing unit record. Thus, most tabulations of person characteristics require manipulation of both housing unit and person records. The item “PERSONS” on the housing unit record indicates the exact number of person records following before the next housing unit record. This feature allows a program to anticipate what type of record will appear next, if necessary. Most statistical software packages are capable of handling the data either hierarchically or sequentially. Many users may still want to create extract files with household data repeated with each person’s record. All fields are numeric with the following exceptions. (1) Record Type is either “H” or “P.” (2) The Standard Occupational Classification (SOC)-based code for occupation and the North American Industry Classification System (NAICS)-based code for industry may have an “X” or “Y.”

RECORD SEQUENCE

The files are released on a state-by-state basis. Records on these files are sorted by geographic area within state. On the 5-percent sample, all households sampled within a particular Public Use Microdata Area (PUMA) appear together. Super-PUMA is a new geographical entity that comprise areas of at least 400,000 people. On the 1-percent sample, all households sampled within a particular super-PUMA appear together. On the 5-percent sample, PUMAs are sequenced in ascending order within super-PUMA within state. Super-PUMAS are sequenced in ascending order within state. In order to provide an extra measure of protection from disclosure of individual households within each geographic area, we scramble the records to avoid any implication of geographic information beyond that which meets Census Bureau disclosure rules for the 2000 PUMS.

The householder record always immediately follows the housing unit record for an occupied unit. This feature simplifies tabulation of households or families by race of householder, ancestry of householder, and even poverty status—since the desired indicators are always on the first person record. The next person record following the householder record is the spouse (if there is a spouse) followed by all family member records, in no particular order. Nonfamily members come last in the household, in no particular order. People sampled from within the same group quarters are not identifiable as such, since each person has an independent pseudo-housing unit record.

METROPOLITAN AREAS

The following items on the housing unit record refer to metropolitan areas. Substitutions should be made as shown.

- AREATYP1, AREATYP5 (substitute “PUMA” wherever super-PUMA is mentioned),
- MIGAREA1 (substitute “super-PUMA of migration” wherever super-PUMA is mentioned),
- MIGAREA5 (substitute “PUMA of migration” wherever super-PUMA is mentioned)
- POWAREA1 substitute “super-PUMA of place of work” wherever super-PUMA is mentioned)
- POWAREA5 (substitute “PUMA of place of work” wherever super-PUMA is mentioned)

Metropolitan Area (MA) codes are based upon June 30, 1999 Office of Management and Budget definitions. A “fully-identified” MA indicates that the entire MA—and no other territory—is shown in one or more super-PUMAs. A “partially-identified” MA indicates that at least one portion of the MA is contained within a super-PUMA (or super-PUMAs) that also contains territory outside of the particular MA.

Example 1. Two-county MSA (containing county A and county B) with the only central city (as well as other noncentral city part) in county A. Super-PUMA 1 only contains county A and Super-PUMA 2 only contains county B. Super-PUMA 1 receives the code “13” indicating that it “contains only metropolitan territory both inside and outside central city (MSA part of fully-identified MSA).” Super-PUMA 2 receives the code “12 ” indicating that it “contains only metropolitan territory outside central city (MSA part of fully-identified MSA).”

Example 2. Two-county MSA (containing county A and county B) with the only central city (as well as other noncentral city part) in county A. Super-PUMA 1 only contains county A and Super-PUMA 2 contains county B, plus a non-MA county. Super-PUMA 1 receives the code “23 ” indicating that it “contains only metropolitan territory both inside and outside central city (MSA part of partially-identified MSA).” Super-PUMA 2 receives the code “70” indicating that it “contains both metropolitan and nonmetropolitan territory.”

MACHINE-READABLE DOCUMENTATION

Every file includes a machine readable “data dictionary ” or record layout. The record layout is the same for the 1-percent and 5-percent files. A user can produce hard copy documentation for extract files or labels for tabulations created; or with minor modifications, can use the data dictionary file with software packages or user programs to automatically specify the layout of the microdata files.

The PUMS Equivalency Files also are available in machine-readable form. These files lists the geographic components (counties or MCDs, places, tracts where available) and their assigned PUMA and super-PUMA codes for the 5-percent and 1-percent samples, respectively. See Appendix J. Equivalency Files.

PREPARING AND VERIFYING TABULATIONS

Estimation. Estimates of totals may be made from tabulations of public use microdata samples by using a simple inflation estimate, that is, summing the weights associated with that variable (e.g. for housing characteristics, use the housing unit weight; for person characteristics, use the person weight.) Those users using subsample numbers to vary the sample size must apply an appropriate factor, or, otherwise adjust the weights to derive an appropriate estimation of totals. We further explain the use of weights and subsample numbers in Chapter 5. Sample Design and Estimation.

Estimation of percentages. A user can estimate percentages by simply dividing the weighted estimate of people or housing units with a given characteristic by the weighted sample estimate for the base. Normally, this yields the same as would be obtained if one made the computation using sample tallies rather than weighted estimates. For example, the percentage of housing units with air conditioning in a 1-percent sample can be obtained by simply dividing the tally of sample housing units with air conditioning by the total number of sample housing units.

Verifying tabulations. Producing desired estimates from the PUMS is relatively easy. File structure and coding of items is straightforward. There are no missing data (see the section "Use of Allocation Flags" in Chapter 4). Records not applicable for each item are assigned to specific NA categories, and it is frequently not necessary to determine in a separate operation whether a record is in the universe or not. PUMS "universe"and variable" definitions may differ from other products produced from sample data primarily because of concerns about disclosure risks (e.g. PUMS files may have different topcodes from SF 3, or the recodes may vary because the components were topcoded). Thus, user tabulations should be verified against other available tallies. Two ways for the user to verify estimates follow:

1. Using control counts from the samples. Total unweighted and weighted population and housing counts are provided for each state. See Appendix I.
2. Using published data from Census 2000. Tabulations from the Census 2000 data base are available in the printed census publications and on the summary data files. Users may check the reasonableness of statistics derived from PUMS against these sources. A familiarity with summary data already available may also facilitate planning of tabulations to be made from microdata. Those publications series likely to be of greatest use for this purpose are listed in PHC-2, Summary Social, Economic, and Housing Characteristics and Summary File 3 (SF 3). In comparing sample tabulations with published data, one must carefully note the universe of the published tabulation. For instance, on PUMS person records, Industry (character position 211-213) is reported for the civilian labor force and for people not in the labor force who reported having worked in 1995 or later. Industry tabulations in Census 2000 publications are presented only for the employed population.

Thus, a tally of industry for all people from whom industry is reported in PUMS records would not correspond directly to any published tabulation. A user should always pay particular attention to concept definitions, as presented in Appendix B. Definitions of Subject Characteristics. One cannot, of course, expect exact agreement between census publications that are based on the complete census count, full sample estimates, or a subsample of the census sample and user estimates based on tallies of a 5-percent or smaller sample. They will inevitably differ to some extent due to chance in selection of actual cases for PUMS.

Chapter 5. Sample Design and Estimate discusses sampling variability and its measurement. User experience has indicated that careful verification of sample tabulations is essential—so important that it may frequently be advisable to include additional cells in a tabulation for no other reason than to provide counts or to yield marginal totals, not otherwise available, which may be verified against available tabulations.

1990-2000 SUBJECT COMPARABILITY

Most of the items for 2000 are comparable to 1990. A few items found in the 1990 PUMS are not in the 2000 PUMS file, primarily because the questions were not asked. Full descriptions of item comparability are given in Appendix B. Definitions of Subject Characteristics.

2000 items not on 1990 files

Grandparents as care givers

1990 items not on 2000 files

Children ever born
Source of water
Sewage disposal
Condominium status

Concepts substantially changed

Race. Users were allowed to identify multiple races.
Geography. The concept of Super-PUMA is new.