2003

Issued January 2004

PRODPR/03-3

Public Use Microdata Sample (PUMS) Files

a data products update from the U.S. Census Bureau

PUMS Files at a Glance

- Detailed demographic and housing data sets for the advanced user
- Extracts of responses from the Census 2000 long-form questionnaire
- Complete U.S. coverage through Public Use Microdata Area (PUMA) geographies, plus Puerto Rico
- Downloadable files on the Web through File Transfer Protocol or purchase on CD-ROM/DVD
- Customizable data

Need Data to Do Your Own Research?

If you are looking to create your own custom tabulations using decennial census data, then the Public Use Microdata Sample (PUMS) files are just what you need. PUMS files show the full range of responses from the long-form questionnaires, for example, how one household or one household member answered questions on occupation, place of work, and so forth. The files contain records for a sample of all housing units, with information on the characteristics of each unit and/or each person in it. The user actually sees the responses made by a household — with certain modifications.

All identifying information is removed to ensure confidentiality. The records selected are a sample of those households that received the questionnaire. The questionnaire included questions on age, sex, tenure, income, education, language spoken at home, journey to work, occupation, condominium status, shelter costs, vehicles available, and other subjects.

The full range of population and housing information collected in Census 2000 is available in the PUMS: over 450 occupation categories, age by single years up to 89, and so forth.

Why Use PUMS?

For many data users, the Census 2000 Summary Files and tabular profiles will suffice. Microdata are for those users who want to create do-it-yourself tabulations, to be able to further draw on the richness of detail recorded in the census.

Who Can Use PUMS?

Microdata users frequently want to look at relationships among variables not shown in the Summary File products offered by the Census Bureau. For example, what characteristics do families with four or more children have in common? What are the incomes at different educational attainment levels? The advantage of PUMS is that data users can tabulate data according to the characteristics they need to know.

PUMS files are perfect for people, such as students, who are looking for greater accessibility to inexpensive data for research projects. Data users in academic life — economists, psychologists, and sociologists — have found the PUMS useful for regression analysis and modeling applications.

Figure 1.
PUMS Products/File Types

Public Use Microdata File Types

	1-percent file	5-percent file
Data detail	Most detail in PUMS files	Some detail filtered out for confidentiality reasons
Geography	400,000 population threshold Super-Public Use Microdata Areas (Super-PUMAs)	100,000 population threshold Public Use Microdata Areas (PUMAs)

United States Census 2000

www.census.gov

USCENSUSBUREAU

Helping You Make Informed Decisions

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU

Confidentiality Issues

Because of the rapid advances in computer technology and the increased accessibility of census data to the user community, the Census Bureau has had to adopt more stringent measures to protect the confidentiality of public use microdata through disclosure-limitation techniques. At the same time, the Census Bureau recognizes the needs of data users for greater characteristic detail and greater geographic specificity. Hence, two sets of files are produced: one that provides a fuller range of detailed characteristics (the 1-percent files) and one that provides greater geographic detail, but less characteristic detail (the 5-percent files).

In the Public Use Microdata Samples, confidentiality is protected by the use of the following processes: data-swapping, top-coding of selected variables, geographic population thresholds, age perturbation for large households, and reduced detail on some categorical variables.

Data swapping is designed to protect confidentiality in tables of frequency data (the number or percent of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases. Swapping is applied to individual records and, therefore, also protects microdata.

Top-coding is a method of disclosure limitation in which all cases in or above a certain percentage of the distribution are placed into a single category.

Geographic population thresholds prohibit the disclosure of data for geographic units with population counts below a specified level. (See descriptions of Public Use Microdata Areas (PUMAs) and super-PUMAs).

Age perturbation, that is, modifying the age of household members, is required for large households (households containing ten people or more).

Detail for categorical variables is collapsed, if the categories do not meet a specified national minimum population threshold.

Geography of the 1-Percent PUMS File

The 1-percent PUMS files provide users the maximum amount of social, economic, and housing information available. There is no national minimum threshold for the identification of variable categories, with the exceptions of a national minimum population of 8,000 for race and Hispanic origin. The goal of these files is to provide a similar level of detail, as was available in the 1990 PUMS files.

In order to provide the level of characteristic detail for the 1-percent files described above, the minimum population threshold needed to be raised above 100,000 (the Public Use Microdata Area (PUMA) minimum). A new geographic entity was created — the super-PUMA. Super-PUMAs have a minimum population of 400,000 and are composed of a PUMA or PUMAs. (PUMAs are shown only in the 5-percent files. Super-PUMAs are identified in both the 1-percent and 5-percent files.) Each state is identified, and any state with a population of 800,000 or greater can be subdivided into two or more super-PUMAs. Super-PUMAs and PUMAs do not cross state lines.

Geography of the 5-Percent PUMS Files

The 5-percent PUMS files provide data for both PUMAs — geographic entities with a minimum population of 100,000 — as well as super-PUMAs. The minimum PUMA threshold was held at 100,000 by increasing the degree of variable collapsing. To maintain confidentiality, while retaining as much characteristic detail as possible, a minimum threshold of 10,000 in the national population is set for the identification of variable categories within categorical variables in the 5-percent files.

The 1-percent files contain geographic equivalency files showing the relationship between super-PUMAs and standard Census 2000 geographic concepts (e.g., counties, etc.). The 5-percent files contain geographic equivalency files showing the relationships among super-PUMAs, PUMAs, and standard Census 2000 geographic concepts. In addition, maps of super-PUMAs and PUMAs are available at: *http://www.census.gov/geo/www/maps/puma5pct.htm.*

Figure 3 below illustrates a map, accessible online in PDF format, of the 1-percent super-PUMA areas available for the state of Connecticut.

Figure 3.
1-Percent Super-PUMAs for Connecticut

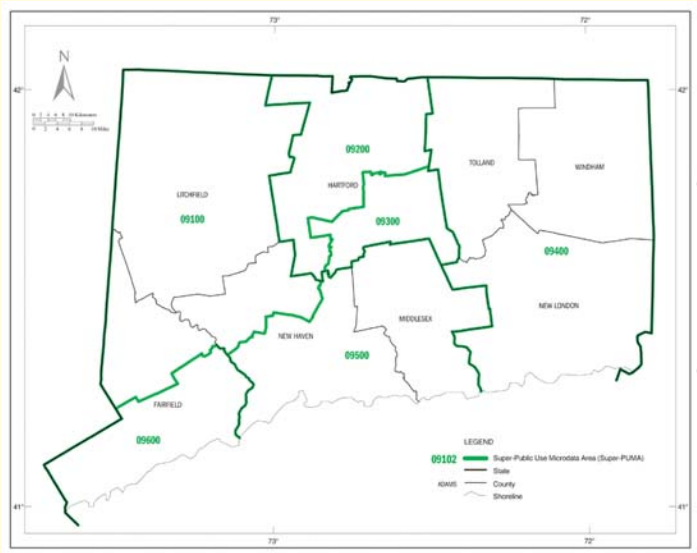


Figure 2.
PUMS Records

Items on the housing record include:		Items on the person record include:	
▪ Bedrooms ▪ Condominium fee ▪ Contract rent (monthly rent) ▪ Cost of utilities and fuels ▪ Family income ▪ Family, subfamily, and household relationships ▪ Fire, hazard, and flood insurance ▪ Gross rent ▪ House heating fuel ▪ Household income ▪ Household type ▪ Kitchen facilities ▪ Linguistic isolation ▪ Meals included in rent	▪ Mortgage status and selected monthly owner costs ▪ Plumbing facilities ▪ Presence and age of own children ▪ Presence of subfamilies in household ▪ Property value ▪ Real estate taxes ▪ Residence state ▪ Rooms ▪ Telephone service ▪ Tenure ▪ Units in structure ▪ Vacancy status ▪ Vehicles available ▪ Year householder moved into unit ▪ Year structure built	▪ Ability to speak English ▪ Age ▪ Ancestry ▪ Citizenship ▪ Class of worker ▪ Disability status ▪ Educational attainment ▪ Hispanic origin ▪ Hours worked ▪ Income by type ▪ Industry ▪ Language spoken at home ▪ Work status ▪ Marital status ▪ Means of transportation to work ▪ Migration ▪ Years of military service, veteran period of service	▪ Mobility status ▪ Occupation ▪ Personal care limitation ▪ Place of birth ▪ Place of work ▪ Poverty status ▪ Race ▪ Relationship ▪ School enrollment and type of school ▪ Sex ▪ Time of departure for work ▪ Travel time to work ▪ Vehicle occupancy ▪ Weeks worked ▪ Work limitation status ▪ Year of entry

Summary Data and Microdata — What's the Difference?

Summary data are predefined cross tabulations of characteristics. The basic unit of analysis is a specific geographic entity — state, county, etc. — for which estimates of persons, families, households, or housing units are provided. The Census Bureau has created the tables that users say they need — but users occasionally need tabulations not covered in the standard data products. It is you, the user, who determines the structure of the tabulation and the characteristics to be tabulated in the PUMS files.

In microdata, the basic unit is an individual housing unit and the people who live in it. The record shows all the information associated with a specific housing unit, except for names, addresses, or other identifying information.

Only large geographic areas are identified on microdata records. These geographic entities are delineated specifically for this product. The Census Bureau uses a minimum population threshold to help avoid disclosure of information about any household or individual.

To further protect confidentiality, there is limited detail on items such as place of residence, place of work, high incomes, and other items.

PUMS Records

There are two basic record types: the housing unit record and the person record. Each has a unique identifier. Each of the records contains a serial number that links the persons in the housing unit to the proper housing unit record. The file is sorted to maintain the relationship between both record types.

The Census Bureau releases the PUMS in this format because of the tremendous amount of data contained in one record. Although these records are extremely large, they can be handled by most statistical or report-writing software. Each record has an individual weight that allows users to produce population estimates close to those in other products showing sample data.

How Are the Microdata Provided?

The U.S. Census Bureau provides two sets of Public Use Microdata Sample (PUMS) files: 1-percent state files and 5-percent state files. These files provide the greatest possible detail, while protecting the confidential nature of the data. For Puerto Rico, 1-percent and 5-percent files also are created. For Guam and the U.S. Virgin Islands, a 10-percent file was created for each island area.

These maps are available at *http://www.census.gov/geo/www/maps/sup_puma.htm.*

Additionally, maps for PUMAs and super-PUMAs can be found in the reference maps area of the Census Bureau's main online data dissemination tool, the American FactFinder®(AFF). While most decennial census data are in AFF, PUMS data are available only by file transfer protocol download or from a CD-ROM or DVD.

Figure 4 illustrates the PUMAs and super-PUMAs for Connecticut using the reference mapping function of American FactFinder®.

Population Thresholds for PUMAs

Each geographic unit in the 5-percent files — PUMAs — must meet a minimum population threshold of 100,000. The minimum PUMA threshold is held at 100,000 people by increasing the degree of variable collapsing to an appropriate level to maintain confidentiality. The 100,000 minimum population threshold allows for a wide variety of local-level geographic analyses, such as studies of nonmetropolitan, metropolitan, and intrametropolitan areas, conducted by public agencies, academic researchers, and others in the private sector. The 100,000 minimum population threshold — the threshold set for both the 1980 and 1990 PUMS files — permits historical comparability.

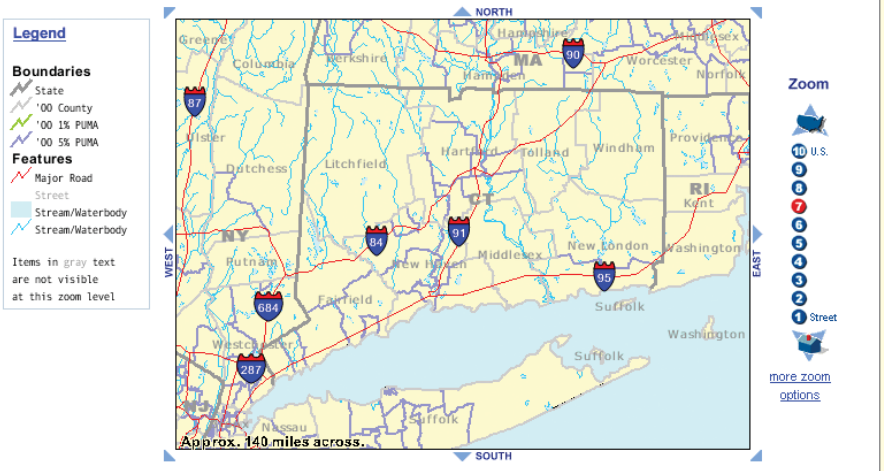
Minimum Population Threshold for Categorical Variables

To maintain confidentiality, while retaining as much characteristic detail as possible, a minimum threshold of 10,000 in the national population is set for the identification of groups within categorical variables in the 5-percent PUMS files.

Post-Processing

To provide more precise means of protecting confidentiality so that as much detail as possible can be included, the 5-percent file variable collapsing requirements are determined after the microdata samples have been drawn. Each variable is analyzed, and only those values that do not meet the 10,000 minimum national population threshold are collapsed into more general categories.

Figure 4.
Reference Map in Amercian FactFinder® Illustrating the 1-Percent and 5-Percent PUMA and Super-PUMA Boundaries



Buy PUMS on CD-ROM or DVD

If you'd like to get all the data for your state, including software to access and manipulate it, then a CD-ROM or DVD is the tool for you. The CD-ROM and DVDs are also beneficial in alleviating long download times for PUMS data from the Internet. You may purchase discs from the Census Bureau's Customer Services Center at 301-763-INFO (4636).

The files on the disc come with proprietary software that operates on computers with Microsoft Windows 95/98/2000/NT/ME/XP operating systems. The software uses a step-by-step approach to creating the desired extract with a tabular interface. You begin by selecting the data element of the extract you need to use, and then, you build your variables from there. Drag-and-drop functionality is used when it comes to moving data elements or the various 'dimensions'. Among many functions, users can create percentage distributions, perform calculations, and apply different weighting factors to the data.

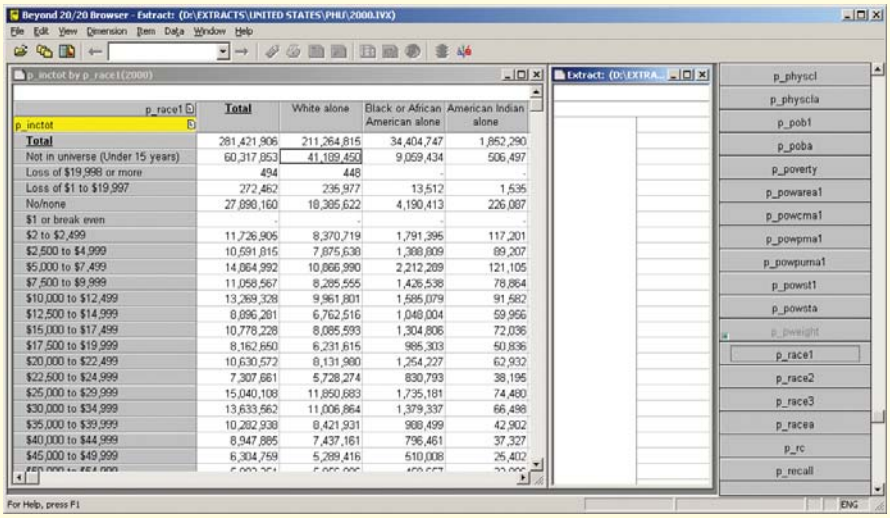
Figure 5 shows the basic table layout functionality of the proprietary software.

Download PUMS on the Internet

The Census Bureau has a File Transfer Protocol (FTP) Web site that allows users to download data sets for various products. Many of the standard decennial data products are available for download from the site. To ease the download burden, the data are typically broken down by state and then the files are segmented and compressed. Users download the zipped files to their own computers where they must then be decompressed or unzipped and reconstructed into a master file. Several of the downloadable products can then be manipulated using common spreadsheet or database packages.

The Public Use Microdata Files are more complex and require a higher level of sophistication for downloading. The PUMS FTP files are broken down by state and are offered in compressed and noncompressed formats. The files are presented as a compressed .zip file or a noncompressed .dat file. These files are designed to work with more advanced statistical software packages and may be too cumbersome for standard spreadsheet and database applications.

Figure 5.
Creating Extracts Using the CD-ROM/DVD Software



For FTP purposes, the url is *http://ftp2.census.gov/census_2000/datasets/PUMS/*

For standard Web downloading, the url is *http://www2.census.gov/census_2000/datasets/PUMS/*

Other Public Use Microdata Files

Several of the Census Bureau's other data collection programs provide Public Use Microdata Files. These programs include:

- American Community Survey
- Survey of Income and Program Participation
- American Housing Survey
- Vehicle Inventory and Use Survey
- Current Population Survey

Many of these microdata files are available online at *www.census.gov* through the data dissemination utility known as the Federal Electronic Research and Review Extraction Tool (FERRET). FERRET is an advanced software tool used to make extracts from the various data sets.

For More Information About Public Use Microdata Samples and Other Census 2000 Data Products

- Visit the Census Bureau's Internet site at *www.census.gov* or call our Customer Services Center at 301-763-INFO (4636).
- Visit your local library. Many major university and public libraries participate in the Federal Depository Library Program and receive copies of Census Bureau reports and discs.
- Call or visit one of 1,800 state and local planning groups, libraries, chambers of commerce, and others that participate in a Census Bureau data center program. For a complete list see *www.census.gov/sdc/www/.*
- Call or visit a Census Bureau Regional Office. For the address and phone number of the regional office near you, visit: *www.census.gov/field/www/.*

Stay Current

To keep up with the release of the microdata discs and maps, as well as other Census Bureau products, users can subscribe to the *Census Product Update*, a biweekly newsletter available free online or in abbreviated form through email. To view the newsletter or subscribe go to *http://www.census.gov/mp/www/cpu.html.*

Other Census 2000 Data Products: Summary Files at a Glance

File	Source	Data	Tables	Smallest geography
Summary File 1 (SF 1)	Decennial census short form	Population, age, sex, race, Hispanic-origin, household relationship, tenure (owner/ renter), group quarters	Total: 286 Population tables: 230 Housing tables: 56	Census block
Summary File 2 (SF 2)	Decennial census short form	Same as SF 1; tables are repeated for 249 race and Hispanic-origin groups	Total: 47 Population tables: 36 Housing tables: 11	Census tract
Summary File 3 (SF 3)	Decennial census long form	Population, age, sex, race, Hispanic-origin, household relationship, marital status, education, employment status, occupation, industry, income, poverty status, citizenship, foreign-born, year of entry, ancestry, disability, migration, place of work, language spoken at home, ability to speak English, journey to work, veteran status. housing tenure, value, mortgage status, rent, plumbing, kitchen facilities, occupants per room, real estate taxes, units in structure, bedrooms, heating fuel, monthly owner costs, year structure built, year household moved in, vehicle availability	Total: 813 Population tables: 484 Housing tables: 329	Block group
Summary File 4 (SF 4)	Decennial census long form	Same as SF 3, but fewer tables; tables are repeated for 336 race, Hispanic-origin, and *ancestry* groups. First detailed data on occupation and industry.	Total: 323 Population tables: 213 Housing tables: 110	Census tract

Appendix I.

Additional Specifications for the PUMS Files

Additional PUMS file specifications are included for the following variables in the 1-percent and 5-percent files.

Dollar Amounts

Dollar amounts are rounded before all summations, ratio calculations, or presentations of amounts. The dollar amounts are represented, including negative amounts, in Figure 6.

This rule is applied to income types, utility costs, mortgage costs, rent, condominium fees, hazard insurance costs, and mobile home fees.

Implementing income top-coding. An individual's income is rounded on a graduated scale and independently top-coded by variable type. The value inserted for observations at and above the top-code is the state mean of all cases at and above the top-code minimum value.

Housing-related dollar amount variables. Property taxes are categorized in a similar way to 1990, with the exception of the higher tax categories. The categories for the 5-percent file are collapsed in order to protect confidentiality.

All other housing-related dollar amounts are treated similarly to income (see above). That is, the variables use the same rounding scale as for income, and each case receives the state mean of top-coded cases for each respective variable.

Race and Hispanic-Origin Data

For the first time in Census 2000, respondents were allowed to mark more than one race. Whether someone is Hispanic or Latino was asked as a yes/no question separate from the race question. Data on race include "yes/no" variables for the five Office of Management and Budget (OMB) races [White, Black or African American, American Indian and Alaska Native, Asian, and Native Hawaiian and Other Pacific Islander] and Some other race on both the 1-percent and the 5-percent files. Presenting the microdata in this manner will allow data users to construct the 63 possible race combinations.

In addition, both the 1-percent and the 5-percent files will show all combinations of the 15 race categories shown on the census questionnaire, specific American Indian and Alaska Native tribes alone, and detailed Asian and Native Hawaiian and Other Pacific Islander groups alone that meet the relevant thresholds. In the 1-percent file, there is a national minimum population threshold of 8,000 for the identification of categories in the race and Hispanic-origin variables; in the 5-percent files there is a national minimum population threshold of 10,000 for the identification of categories in these variables. For example, the racial category "Black or African American and Filipino" is shown on both files, because there are more than 10,000 people in the United States who reported this combination on Census 2000.

Age Detail

For both the 1-percent and 5-percent files, single-year age categories are provided through age 89. There is one nationwide top-code (age 90) and each state receives the mean age of individuals in the state 90 years and over.

Ancestry Variables

The Census Bureau codes up to two responses for the ancestry question. For the 5-percent file, if the combined total national population from both of these responses for an ancestry group is 10,000 or greater, that group is identified by itself in both the first response and second response variables, even if the total for the category in either or both of the individual ancestry variables does not meet the 10,000 threshold.

Industry and Occupation

Two sets of codes for each occupation and industry are provided: (1) the census code and the Standard Occupational Classification (SOC)-based code for occupation and (2) the census code and the North American Industry Classification System (NAICS)-based code for industry.

Continuous Variables

Continuous variables are treated the same on both files. Additional specifications for departure time (when a person usually left for work in the week before their census form was filled out) and year of entry into the U.S. are described below.

Departure time is categorized as follows:

12 midnight - 2:59 a.m., in 30-minute increments

3 a.m. - 4:59 a.m., in 10-minute increments

5 a.m. - 10:59 a.m., in 5-minute increments

11 a.m. - 11:59 p.m., in 10-minute increments

Year of entry into the country will have a bottom-code of 1910.

Figure 6.
Represented Dollar Amount

Dollar amounts	Rounded dollar amount
No income	$0
$1-$7	$4
$8-$999	round to the nearest $10
$1,000-$49,000	round to the nearest $100
$50,000 or more	round to the nearest $1,000

Order Form

Yes! I want the Census 2000 Public Use Microdata Sample Files (PUMS). Please send me the CD-ROMs and/or DVDs I have chosen below. I have enclosed $ _____ (*check or money order only*) for my selection or provided credit card information below.

Name: _____ Date: _____

Company: _____

Address: _____
(City, state, ZIP Code) (No P.O. boxes)

Phone: ( ) _____ Fax: ( ) _____

E-mail address: _____

Product	Software provided	Price	Quantity	Total
PUMS 1%: U.S. on CD-ROM	None—ASCII Format	$50		
PUMS 1% and 5%: U.S. on DVD-ROM	Yes	$70		
PUMS 1%: U.S. on CD-ROM	Yes	$50		
PUMS 10%: Guam on CD-ROM	None—ASCII Format	$50		
PUMS 10%: Virgin Islands on CD-ROM	None—ASCII Format	$50		
			SUBTOTAL $	

TOTAL $ _____

METHOD OF PAYMENT (*please check one*)

☐ **Overseas Service** — *add $25.00*

☐ Check payable to Commerce-Census

☐ 9☐☐☐ Census deposit account

VISA, MasterCard, AMEX, or Discover account number:

_____

Expiration date: Month/Year: _____ /_____

Name on card: _____

Signature: _____

To place your **order by phone** *(have your credit card ready)* or for further information about PUMS data products, please call our Customer Services Center at **301-763-INFO (4636)**, fax your order toll-free to **888-249-7295**, or access our Web site at *http://www.census.gov* and select "Catalog."

For **mail orders only**, complete this order form and send it with your payment to: **U.S. Department of Commerce**
U.S. Census Bureau (MS 0801)
P.O. Box 277943
Atlanta, GA 30384-7943