

Using Public-Use Microdata Samples With SAS®

Warren A. Brown
Cornell University

Lisa King
Cornell University

What are PUMS?

If you wanted to answer the following questions, where would you turn for a source of data?

In our city, how many low-income families are at risk of becoming homeless?

In which industries are immigrants to our labor market area finding jobs?

The 1990 Census collected information which could be used, but none of the tables in the summary tape files or the printed reports gets at these questions. If you could access the Census Bureau's confidential files with individual household and person information then you could prepare your own tabulations. Although the Census Bureau is not going to give you access to these confidential files, they have prepared a public use version of the individual household and person responses to the census which will allow you to pursue the previous questions along with virtually any other tabulation or statistical analysis of census data you may wish to pursue. The product is called the Public Use Microdata Sample, or PUMS for short.

The Public Use Microdata Samples are computer files 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 files permit users with special needs to prepare their own tabulations of census data.

In the 1990 version, PUMS is available in two standard configurations: a 5 percent county level file, and a 1 percent metropolitan area file of the population and housing of the United States.

In addition to these two, some other PUMS configurations likely to be prepared are:

- a 3% elderly file
- a 1% labor market area file; and

a contextual variables file.

Comparison of Summary Data and Microdata

Figure 1 illustrates the basic distinctions between summary data and microdata. Summary data are the type of data found in census printed reports and summary tape files. The basic unit of analysis is a specific geographic area (for example, a census tract, county or state) for which counts of persons or housing units in particular categories are provided. In microdata, the basic unit is an individual housing unit and the persons who live in it.

There are two types of microdata: confidential microdata and public-use microdata. Confidential microdata include the census basic record tapes, computerized versions of the questionnaires collected from households, as coded and edited during census processing. The Census Bureau tabulates these confidential microdata in order to produce the summary data that go into the various reports, summary tape files (STFs), and special tabulations. Public-use microdata samples are extracts from the confidential microdata taken in a manner that avoids disclosure of information about identifiable households or individuals.

Records on public-use microdata samples contain no names or addresses. Also, the Bureau limits the detail on place of residence, place of work, high incomes, and selected other items to further protect the confidentiality of the records. Microdata records identify no geographic area with fewer than 100,000 inhabitants. Microdata samples include only a small fraction of the population, drastically limiting the chance that the record of a given individual is even contained in a microdata file, much less identifiable.

Use of Microdata Files

Public-use microdata files essentially make possible "do-it-yourself" special tabulations. The 1990 files, for the most part, contain the same variables and recodes as Summary Tape File 3 (STF3). Subject to the limitations on sample size

and geographic identification, it is possible for the user to construct a seemingly infinite variety of tabulations interrelating any desired set of variables. Users have the same freedom to manipulate the data that they would have if they had collected the data in their own sample survey, yet these files offer the precision of census data collection techniques and sample sizes larger than would be feasible in most independent sample surveys.

Microdata samples will be useful to users who are doing research that does not require the identification of specific small geographic areas or detailed cross tabulations for small populations. Microdata users frequently study relationships among census variables not shown in existing census tabulations, or concentrate on the characteristics of certain specially defined populations, such as housing costs as a percent of income among low income households, or employment characteristics of immigrants.

Microdata are also necessary for answering some kinds of questions in a methodologically correct manner. For instance, if one wanted to establish a relationship between minority status and income level for a region of the country, and summary data were all that was available, you might be tempted to fall into the ecological fallacy by establishing that income level among a group of counties or census tracts is correlated with the percent of the population which is non-white. Rather, by using microdata one can look at the relationship using households or individuals as the unit of analysis and be methodologically correct.

Sample Design and Size

Each microdata file is a stratified sample of the population, actually a sub-sample of the full census sample that received census long-form questionnaires (roughly 1 in 6 households). Sampling is done household-by-household in order to allow study of family relationships and housing unit characteristics. Sampling of persons in institutions and other group quarters was done on a person-by-person basis. Vacant units were also sampled.

There are two independently drawn samples in the 1990 PUMS, designated the 5% and 1% samples, each featuring a different geographic scheme. The 5% sample will give the user records for over 12 1/2 million persons and almost 5 million housing units.

For each sample type, the record contains an individual weight. There are housing unit weights on the housing records, and person weights on their person records. This is similar to the weighting procedures for the Current Population Survey, and a change from 1980 when the samples were self-weighting. This somewhat more complicated weighting procedure allows users to produce estimates which more closely approximate those provided by the tables in Summary Tape File 3.

Subject Content

With only minor exceptions, microdata files will contain the full range of population and housing information collected in the 1990 census. Figure 2 lists the subject content of the 1990 PUMS. An excellent way to familiarize yourself with the subject matter is to review the tables produced from STF3. Since the variables and the coding are essentially the same, any subject matter dealt with in STF3 can be examined in PUMS.

There are no "missing data" categories in files. Substitutions or allocations have been made for any missing data. Substitution and allocation flags appear on each record indicating whether the data were "substituted" for the entire housing unit, whether they were substituted for all the person data, or to identify each item which has been allocated. Thus, a user desiring to tabulate only actually observed values can eliminate those cases with allocated values.

Geographic Identification

The PUMS file includes identification of geographic areas, the smallest of which must exceed 100,000. These are called PUMAs. That's right, this time around you will need to keep your TIGERs and PUMAs straight. PUMA stands for Public Use Microdata Area. A PUMA can be a group of tracts, a place, a group of places, a minor civil division, a group of MCDs, a metropolitan area, a county, or group of counties.

Many of the uses of the data will involve tabulations or research of a local or regional nature. Because of this, the Bureau of the Census allowed organizations within their outreach network to participate in the formation of the PUMAs. The State Data Centers were responsible for coordinating input within their state and submitting a proposal for the PUMAs to the Bureau.

Structure of the File

The PUMS is structured as a hierarchical file. Each file consists of a series of logical records of two types—housing units and persons. Each housing unit record is followed by a variable number of person records, one for each member of the household or none if vacant. Each person in group quarters has two records—a dummy housing unit record as well as a person record.

Figure 3 is a representation of the hierarchical structure. Housing Unit 1 is occupied by five persons and Person a is the householder. Nothing in the structure tells you whether this is a married couple family with three dependent children, or an apartment unit with five college students sharing the housing unit. This type of detail is contained in the variables. Housing Unit 3 is an unoccupied housing unit, and thus there are no person records associated with it. Housing Unit 4 could represent a single person household, or it could represent a person living in a group quarters—a college dormitory, prison, or military barracks for example—for which the housing unit record is essentially a dummy record.

Geographic identifiers appear only on the housing unit record, so most tabulations of person characteristics require manipulation of both housing unit and person records. SAS can handle the hierarchical structure, but users may want to work with the data in rectangular form anyway, especially if you are planning to use the data with software which is limited to rectangular files only. In order to rectangularize the files it will require extracting person records with any desired household data repeated on each person record.

Strategies for Using PUMS with SAS

Below are two different examples of analyses that can be done with the PUMS files, one looks at household level data and the second at person level data. These examples demonstrate strategies for approaching some typical research questions using PUMS with SAS, and illustrate the power that PUMS files give you to carry out such analyses. Since the 1990 PUMS files are not available at this time, we use 1980 PUMS data for the Kansas City area in our examples.

EXAMPLE 1: A Household Level Analysis

In this example, we look at the issue of housing affordability among low-income households in Kansas City, Missouri. The indicator we are after is shelter costs as a percentage of household income for households whose incomes in 1979 fell below 125% of the federal poverty level. The file we use is the 1980 PUMS-A 5% sample.

Figure 4 shows the program we use to extract the appropriate data from the PUMS file and put it into a SAS data set for analysis. Due to the hierarchical layout of records in the PUMS files, it is necessary to read through the file and pick out households of interest, skipping over person records in between. To distinguish household from person records, check record type in column 1 of the file. Person records are arranged within household in order of relationship. The householder record immediately follows the housing unit record (for occupied units). This simplifies tabulation of households or families by race or ancestry of householder, and even by poverty status, since the desired indicators are always on the first person record. Note that the program reads the record type, county group code, and tenure (i.e. owner vs. renter status) for each household, and then uses those values for selection purposes as follows. We want to consider only household records, and only for owned or rented households in Kansas City. Once such a household has been identified, we check the poverty status of the household, which is to be found on the householder record (the person record immediately following the housing unit record). Notice that we must check tenure FIRST, in order to get rid of the vacant units for which there are no subsequent person records.

If a household is low-income by our definition (i.e. below 125% of the poverty level), the program goes on to read the rest of the variables of interest from the household record. These are:

- Selected Monthly Owner Costs, which includes mortgage, real estate taxes, insurance premiums, utilities and fuels,
- Gross Rent, which includes the monthly rent payment plus the average monthly cost of utilities and fuels (to the extent those are paid by the renter), and
- Household Income.

The 1980 PUMS samples are self-weighting, i.e. they are selected in such a way as to be

representative of the entire population within each state. To estimate a complete-count census frequency, multiply the sample frequencies by the reciprocal of the sampling rate. In this case, we use a weighting variable called SELFWT, which is equal to 20 in all cases (the reciprocal of the 5% sampling rate).

Once the appropriate measures have been input, the program calculates shelter costs as a proportion of household income. Figure 5 contains a summary of our findings, weighted by the appropriate household weights. Note that a household is sometimes said to be vulnerable to homelessness if shelter costs exceed 30% of household income. Had this been an actual research project, we might also have looked at age, race, and family status profiles of the households that were spending more than 30% of their incomes on housing.

EXAMPLE 2: A Person Level Analysis

Person level analyses usually involve extracting some data on the housing unit as a whole from the household record, along with the measures available for each person in the household. In this example, we look at the industry of employment and labor force status of in-migrants to the Kansas City labor market area between 1975 and 1980. Due to budget restrictions, data on migration, place of work, and travel time to work were coded for only half of the households on the 1980 PUMS files. In the 1990 PUMS this information will be coded for all households in the file.

Among the members of the civilian labor force who moved to the region between 1975 and 1980, what were the primary industries of employment and how did the unemployment rates for these individuals differ by industry?

For this labor market analysis, we used a special 1980 PUMS tabulation known as PUMS-D, which was produced cooperatively by the US Department of Agriculture and various regional rural development centers. PUMS-D, which is a 1% sample, employs a unique geographic scheme in which counties are grouped into labor market areas based on patterns of residence and employment. Clusters of counties were formed based on county-to-county commuting patterns, and these clusters were then used as county groupings, or "labor market

areas." The Kansas City labor market area in 1980 consisted of 18 counties in Kansas and Missouri which formed a commuting region. Labor market areas seemed particularly appropriate for this example, since we are looking at employment opportunities.

Although PUMS-D is a 1% sample, recall that migration data was coded for only half the sampled households. Since we are including only individuals for whom residence in 1975 has been coded, we have only half of the 1% sample. Therefore, each individual person in our subsample is assigned a weight of 200 (two times the reciprocal of the 0.5% sampling rate).

Figure 6 shows the SAS program we use to extract the necessary person records. This program differs from that used in Example 1 in that it reads data from the housing unit record, and then attaches that data to each person record within that household (through the use of the RETAIN statement). More specifically, we read the current labor market area code and household type from the housing unit record, then go on to read labor market area of residence in 1975 and labor force status from the subsequent person records. Only those individuals who are members of the civilian labor force, live in the Kansas City labor market area in 1980, and resided elsewhere in 1975 are output to the final file.

From Figure 7 one can see the breakdown of industries in which this population of in-migrants were employed, and how the unemployment rates varied across these industries. This was a fairly simplistic example. If we had examined data for the entire country, we could also have looked at some characteristics of those identified as out-migrants from Kansas City, subject to the limitations of sample size. It would also be possible to determine where our population of in-migrants lived before moving to Kansas City.

As a final note, the hierarchical structure of the PUMS files lends itself well to manipulation by true relational database software such as Structured Query Language (SQL). Users of the PUMS might want to consider creating separate tables of household and person records via PROC SQL in order to take advantage of the powerful data handling capabilities of this new procedure.

SAS® is a registered trademark of SAS Institute, Inc., Cary, N.C., USA.

Suggested Reading

Census of Population and Housing, 1980: Public-Use Microdata Samples Technical Documentation / prepared by the Data User Services Division, Bureau of the Census. --Washington: The Bureau, 1983.

Myers, Dowell. *Analysis with Local Census Data: Portraits of Change*. Boston: Academic Press, 1992.

Poston, Dudley L. Jr. "The Use of Census Data by Social Demographers." In *Census '80: Continuing the Factfinder Tradition*. By Charles P. Kaplan, Thomas L. Van Valey and Associates. Washington, D.C.: GPO, 1980, pp.413-432.