INTEGRATION OF THE PUBLIC USE SAMPLES OF THE U.S. CENSUS

Steven Ruggles
University of Minnesota, Minneapolis 55455

KEYWORDS: Census, microdata, public use sample

Public use microdata samples of the U.S. census of population covering nine census years between 1880 and 1990 are currently available or in preparation. Taken together, these microdata comprise our richest source of quantitative information on long-term changes in the American population. Because these samples were created at different times by different investigators, however, they have incompatible documentation and a wide variety of record layouts and coding schemes. These differences among the samples inhibit their use as a time series.

At the Social History Research Laboratory of the University of Minnesota, we are planning to convert the series of public use samples into a single coherent form. The success of this project will depend on the usefulness of the data series to a broad range of social scientists. This paper describes the planned Integrated Public Use Microdata Series in the hope of stimulating comments and suggestions before the design is finalized.

Background and Significance

Sociologists, economists and demographers have developed a variety of quantitative data sources to study social change, including retrospective surveys, repetitions of early social surveys, and longitudinal surveys. Although such data sources are essential, they are usually limited to the analysis of changes during the past thirty years. The study of longer term change -- over the past 100 or 150 years -- has been sharply constrained by the limited availability of consistent data series. Analysts of nineteenth-century society have often turned to institutional and bureaucratic records, such as those generated by churches and the military, but these sources are typically available only for the distant past and they are ordinarily limited to the study of specific population subgroups.

The decennial census is the most consistent general source of information about the American population over the past two centuries. Quantitative studies of long-term social change have always relied on the published tabulations of the census, but these data have substantial limitations. In each period, the topics addressed by census publications have focussed on contemporary concerns, and these concerns have shifted dramatically over the past century. Moreover, the high costs of tabulation before the introduction of modern data processing equipment meant that few cross-classifications of census data were possible, and much of the information collected by the census was never tabulated at all. Even for recent census years, the published census volumes have significant limitations for the study of social change. Despite the dramatic increase in the quantity of published census data in recent years, the Census Bureau cannot anticipate all the questions social scientists want to ask.

The Census Bureau has addressed these problems by producing individual-level public use samples of the census for each census year since 1960 (U.S. Bureau of the Census 1972a, 1973, 1982, 1989b). These samples are machine-readable transcriptions of enumeration forms with the names and other identifying characteristics of individuals removed to preserve confidentiality. From the beginning, the public use samples have proven to be a valuable resource, since they allow researchers to make tabulations tailored to their specific research questions. The samples are especially useful for studies of social change, because they allow researchers to avoid many of the problems of incompatibility in the published data for different census years. In addition, the public use samples have allowed researchers to move beyond simple tabular analysis and apply increasingly sophisticated multivariate techniques. The existence of these data has significantly expanded the power of quantitative social science research.

Since 1980, four historical public use samples have been created for the census years 1900, 1910, 1940, and 1950, and an additional sample for 1880 is currently in preparation at the Minnesota Social History Data Archive (Graham 1980; U.S. Bureau of the Census 1984a, 1984b; Strong et. al. 1989; Ruggles and Menard 1990). Although these files have only been available for a few years, they have already led to an outpouring of new research on the nature of long-term social change. As each new sample is created, the value of the previous public use files has been enhanced, since they become increasingly usable for the analysis of change.

To date, however, only a small proportion of the research based on public use samples has fully exploited the potential for the study of change over time. Many investigators are using the samples as isolated cross-sections. A preliminary bibliography of recent research using the public use samples compiled by the Social History Data Archives at the University of Minnesota reveals that over 80 percent of studies use only one of the eight public use samples currently available.

It is difficult to use more than one of the public use samples at a time because each sample has a different format, different coding schemes, and different documentation. Six separate research teams have been involved in the creation of the samples, and each of them has had their own ideas on how to organize the data. We are faced with eight different occupational classifications that have a total of 3200 different categories, and seven incompatible classifications for variables such as birthplace, household relationship, and institution type. Documentation for the eight existing samples is contained in eight separate volumes totaling about 3000 pages. These volumes are for the most part organized differently from one another, and their treatment of comparability issues is often cursory.

The incompatibility of the public use samples in their present form means that multi-sample studies require a large initial investment to prepare the data for use. The number of investigators using multiple public use samples is growing rapidly. Most have proceeded by creating a set of special-purpose semi-compatible extracts containing a limited number of variables and minimal documentation. This ad hoc approach has already led to increasing duplication of effort. Moreover, given the complexity of the files and the often subtle differences among them, the potential for error is large.

The Social History Data Archives plans to convert the public use samples for 1880, 1900, 1910, 1940, 1950, 1960, 1970, 1980, and 1990 into a single consistent format and to prepare an integrated set of documentation oriented to the use of the samples as a series. In the long run, we anticipate adding data for all the remaining census years for which individual-level census enumerations survive; these years are 1850, 1860, 1870, 1920 and 1930. We have already applied for funding to create a new sample 1920, and plan future applications for the 1850, 1860, 1870 and 1930

census years. When complete, the Integrated Public Use Microdata Series will provide individual-level data for a large random sample of Americans in 14 census years spanning a century and a half. The series will constitute a resource of unprecedented power for the study of long-term social change.

Design of the Data Series

Detailed planning of the Integrated Public Use Microdata Series is a major undertaking. The following sections are intended to raise the most important design problems and to explain our general approach to them.

To set the context for this discussion, Table 1 shows the availability of the most important variables in each of the public use samples. Eleven basic questions were asked in every census year, and twenty-two inquiries are available for at least seven of the nine census years. There are also, however, a significant number of variables not shown in Table 1 that are available for only one or two census years. In addition to the differences in available information across census years, there are also multiple versions of the samples for recent years with slightly differing variables.

1. Record layout and general coding design. Following conventional practice for public use samples, the Integrated Public Use Microdata Series will consist of numeric codes arranged in a column-format hierarchical structure. Variables common to the household as a whole -- such as geographic indicators and housing questions -- will appear on a household record. Each household record will be followed by a series of person records describing individual-level characteristics.

The design of the record layouts will stress column-compatibility rather than compactness. In general, all variables available in multiple census years will appear in the same columns in every year. When a variable is not available for a given year, the columns will be filled with a missing data value. This means that the integrated versions of the public use samples will be substantially larger than the originals; we anticipate a record length in excess of 250 bytes, which is almost twice the average of the existing samples. The great advantage of column compatibility is that it simplifies the construction of multi-year data files and minimizes the potential for user errors. In view of the rapid decline in the cost of mass data storage, it makes sense to focus on efficiency of use rather than efficiency of computing resources.

In the 1940 and 1950 census years, individuals falling on a designated "sample line" of the census form were asked an extra set of questions. The public use samples were constructed so that each household contains one sample line individual, and the extra sample-line variables are provided on a separate sample-line record. The integrated public-use microdata series will eliminate the sample-line record by embedding the sample variables in the person records. Individuals who were not asked the extra questions will receive a missing data code for all sample-line variables.

The public use samples employ differing numeric classification systems in every census year, and reconciliation of these classifications is a major part of this project. For most variables, it is impossible to construct a single uniform classification without an unacceptable loss of information. Some census years provide more detail than others, and if we reduced all census years to their lowest common denominator we would sharply reduce the power of the data series. For example, the household relationship classification for the 1960-1970 census years consists of only fifteen categories. All the other census years are more detailed; in fact, the 1910 census distinguishes 124 categories of household relationship. If we were to adopt the 1960-1970 classification as a standard, we would lose the ability to distinguish such household relationships as nephews, aunts, farmhands, and domestic servants.

To avoid such problems and still maximize the compatibility of coding systems, we will design composite coding systems for most variables. The first one or two columns of each variable will be entirely compatible; in the case of household relationship, for example, the first two digits will be equivalent to the fifteen-category coding system of the 1960-1970 censuses. An additional one or two columns will provide added detail for particular census years or groups of years. This approach maximizes ease of use and minimizes information loss at modest cost in space. It will be applied to all the complex categorical classifications except for occupation, which is discussed at length below.

2. Reconciliation and integration of data dictionaries, 1880-1910. Although the public use samples consist entirely of numeric codes, information on topics such as occupation, household relationship, and birthplace was originally entered in alphabetic form. These alphabetic strings were converted into numeric categories by means of data dictionaries which assign a code to each unique alphabetic string. For the census years 1880-1910, we have access to the alphabetic transcriptions of the original census enumeration forms. We plan to maximize coding consistency across these census years by merging the data dictionaries and eliminating any incompatibilities. This will ensure that identical responses are coded identically in each year. The procedure cannot be carried out for the 1940-1990 census years, since the uncoded data are still protected by privacy rules; we must therefore make do with numeric classifications constructed by the Census Bureau.

3. Design of occupational and industrial classifications. Occupation is the most complex variable collected by the census. It is also among the most important census variables for analysis of long term social change, because the early census years provide few alternative indicators of socioeconomic status or labor-force participation. The census bureau has modified its classification systems every decade, so all comparisons of occupation and industry require extensive reconciliation of codes. There are nine different occupational classification systems consisting of between 285 and 550 categories each.

We plan to construct three different consistent coding systems for occupation. Our efforts to date have focussed on conversion of all census years into the 1950 Census Bureau classification system, because the 1950 system poses the fewest technical difficulties (Jelatis 1990). Some historians have argued that the dramatic changes in occupational classification systems between 1900 and 1950 are a reflection of fundamental changes in the labor force, and that the earlier classifications may be especially appropriate for study of the late nineteenth and early twentieth centuries (Conk 1980). We therefore plan to provide an older classification in addition to the 1950 system, and have tentatively decided to use the 1880 classification system. Finally, we will recode all census years into a simplified version of the 1990 classification system in order to facilitate comparisons with recent data. We will also include the original contemporary classifications for each census year.

For the period since 1940, the recoded classifications will be imperfect. In each year, the Census Bureau split some categories and merged others, so the various systems are not fully compatible. The Bureau has, however, provided sufficient technical information to determine the number of misclassified persons in each category, and we will include this information as part of our documentation (U.S. Bureau of the Census 1968, 1976b, 1989a). For the census years prior to 1940 we have access to the original

archive versions of the data files, and these include the original alphabetic transcriptions of the occupation fields. These data are sufficiently detailed to allow reasonably precise classification according to the 1880, 1950, and 1990 systems.

The census classifications of industry have been much more consistent than those of occupation. In fact, it is possible to construct a single four-digit classification of industry with virtually no loss of information. We plan to adopt the three-digit 1950 system as the standard, with an additional digit giving more detailed information for later census years.

The census did not ask a separate question on industry prior to 1910. Instead, enumerators were expected to provide information on industry as part of the occupational question. Because we have access to the detailed alphabetic transcriptions of occupation in the early census years, we expect that we will be able to classify industries accurately in most cases.

4. Design of constructed occupational status variables. None of the Census Bureau occupational classifications are sequentially ordered according to socioeconomic status or occupational prestige, so we plan to construct several additional variables to capture these dimensions of occupational structure. There is considerable evidence that the hierarchy of occupational prestige has changed only modestly since the mid-nineteenth century (Tyree and Smith 1978; Hauser 1982; Treiman 1976; Hodge, Seigel and Rossi 1964; Sharlin 1980). Therefore, modern occupational prestige scores -- such as those described in Treiman (1977) and Reiss (1961) -- should provide a reasonable approximation of prestige through the entire period.

The prestige scores will be supplemented by objective occupational measures of socioeconomic status. We will create an income score and an education score based on occupational titles. These scores will reflect the median income and education of persons with each title in the 1950 census. The scores will be accompanied by precision indices based on the variability of income and education within each title. The 1950 census is a reasonable standard for these measures, since it is the earliest census year to provide full information on income, and it offers the most broadly compatible occupational classification system. We will also explore the construction of a variable reflecting the nineteenth-century economic rank of occupations, based on contemporary wage and salary data.

5. Design of constructed variables on household composition. Each of the public use samples includes constructed household composition variables on the household record, but these variables differ from year to year. The most useful classifications will be incorporated in the integrated microdata series (see Table 2). We will also construct the 19-category Laslett/Hammel household classification for all census years; despite its limitations, the Laslett/Hammel scheme remains the most widely used household classification for historical analysis.

Variations in sample design among the public use samples can introduce incompatibilities in household and family classifications. All the samples incorporate provisions for individual-level sampling of large units such as institutions and boarding houses, but the criteria for individual-level sampling have varied from year to year. As a result, some family relationships cannot be identified in all census years. For the census years 1940 through 1970, for example, no information on household composition or family relationships is available for units containing five or more persons unrelated to the household head. Moreover, since 1970 the Census Bureau has ceased to gather data on secondary family relationships.

The 1970 sampling rules constitute a lowest common denominator for household composition and family relationships; virtually all information on these topics available in 1970 is also available in all other census years. Moreover, sufficient information is available in all census years to determine the sampling unit of each case under the 1970 rules. Thus, it is possible in all census years to suppress information on household composition and family relationships if it would not have been available under the 1970 rules. Unfortunately, the 1970 rules are inappropriate for the early twentieth century, when many households included five or more boarders or servants and a high percentage of the population resided in secondary families. Our experience has shown that many users require the greater precision available in the early census years. We will therefore retain all available household and family information, but we will construct an individual-level variable indicating the sampling unit under 1970 rules (SU1970). The documentation will provide full instructions on the use of this variable to convert each family and household variable into fully consistent form.

6. Design of constructed variables on family interrelationships. Individual-level variables describing interrelationships among family members are even more important for most users than household level classifications. Such variables make it possible for users to create specialized measures of living arrangements tailored to their specific research topics, such as living arrangements of the elderly or of single parents. These measures also facilitate the construction of specialized own-child fertility measures and measures of marriage characteristics.

The 1940 and 1950 census years provide the most comprehensive set of variables on family interrelationships. These eight variables indicate the membership of each individual in a specific primary family, secondary family, and/or subfamily (FAMUNIT and SUBFUNIT); the size of each of these units (FAMSIZE and SUBFSIZE); the type of unit, defined by the presence of married couples and own children (FAMTYPE and SUBFTYPE); and the relationship of each individual in the unit to a reference person, ordinarily the first listed person in the unit (FAMREL and SUBFREL). We will construct these variables for all census years. In addition, we will create three pointer variables that give the location within the household of each individual's spouse, mother, and father (SPLOC, MOMLOC, and POPLOC). The pointer variables allow users to easily attach characteristics of these kin, and sophisticated users find them to be convenient tools for the construction of measures of fertility and coresidence. Finally, we will include several of the most commonly requested variables on own-children: number of own children, number of own children under five years old, age of eldest own child, and age of youngest own child.

In every census year, there is a small percentage of ambiguous relationships among family members. The information available for sorting out such ambiguous family relationships varies from sample to sample. For the sake of consistency, many investigators will want to use family interrelationship variables based entirely on information available in all census years. There are certain applications, however, for which the greater precision available in some years is required. The Integrated Public Use Microdata Series will accommodate both needs through the use of flags. The variables MOMLOC, POPLOC, SPLOC, FAMUNIT, and SUBFUNIT will be accompanied by flags indicating (1) if the link or unit membership would be the same even if minimal information were used; (2) if the link was only made because of extra information available in the particular census year; or (3) if the link is contradicted by extra information available in that census year.

7. Design of geographic codes. The geographic codes

are the most frustrating ones. Precise information on locality was gathered in every census year, but because of privacy regulations this information has been suppressed in the public use samples of the period 1940-1990. In 1960 and 1970, places with fewer than 250,000 inhabitants were not identified; for 1940-1950 and 1970-1990, the threshold for identification is 100,000. Within these constraints, the classification systems for identifying places within states have varied considerably. The 1940 and 1950 samples provide State Economic Area, which is a system for coding county groups, and Standard Metropolitan Area (SMA). No information on urban/rural residence is given for those years. In the 1960 public use sample, no geographic locations below the state level are available, but there is a variable on urban/rural residence. In both 1970 and 1980 the Census Bureau released three versions of the public use samples containing alternate geographic variables, and there will be two versions for the 1990 census. In spite of this, there remain significant problems of compatibility in the geographic codes for these three years. Both 1970 and 1980, however, do identify Standard Metropolitan Statistical Area (SMSA), and the 1990 census will identify the closely comparable system of Metropolitan Statistical Area (MSA). The definition of SMSA differs slightly from the earlier SMA, but they can be viewed as compatible for many applications. In addition, some county groups can be consistently identified from 1970 to 1990, and all three years identify urban/rural residence. The 1940-1950 and 1970-1990 census years also include sufficient information to identify consistently the residents of 61 of the largest cities.

There isn't much we can do about the geographic incompatibilities of the 1940-1980 census years, other than imposing consistent numeric codes for SMA and SMSA. What we can do is construct variables for the early census years that are compatible with the later public use samples. The samples for 1880-1910 provide full information on county, city, and enumeration district. It is therefore possible to construct the State Economic Area and County Group variables used in recent census years, although the correspondence will not be precise because of boundary changes and creation of new counties. We will also create the closest possible analogs of SMA and urban/rural residence. However, the definitions of these variables depend in part on commuting ties, telephone calls, and other measures of integration and metropolitan character that are not available for the earlier census years, so they will be only approximately comparable.

8. Design of all other coding systems and constructed variables. The classification issues discussed above are the most problematic ones, but there are a wide variety of other variables that will require significant work. For example, the development of a consistent scheme for classification of countries of birth and parental birth is complicated by dramatic boundary changes in many parts of the world since the mid-nineteenth century. Other complex classifications include mother tongue, institution type, and ancestry. Even such apparently straightforward classifications as income, race, employment status, and education require some manipulation to make them compatible across census years.

The data series will include two types of constructed variables not previously mentioned. First, we will provide tools to help users manipulate the data; these variables will include census year and sample number, record type, number of person records in sample unit, household sequence number, and person sequence number within households. Second, we will construct several household-level variables summarizing characteristics of the household as a whole, including economic status, racial composition, and life-cycle classification.

Documentation

The creation of integrated documentation is the highest priority of this project. By comparison with the usual standards of social science research, the existing documentation of the public use samples is quite good, but it still has significant limitations. Among the seven census years that have been available for a year or more, only two were ever documented as fully as had been originally planned. Indeed, in some cases it is impossible to use the samples correctly if one relies entirely on the documentation supplied with the data.

The documentation is particularly awkward when using the public use samples as a time series. With few exceptions, the documentation for each census year is organized differently. The combined documentation adds up to some 3000 pages, and there are no indexes. Simply learning how to look things up in each census year requires a substantial investment of time. Moreover, with the exception of the most recent census years, the discussions of comparability issues range from inadequate to nonexistent.

We plan a three-volume set of documentation, consisting of a general user's guide, a volume on comparability issues and procedural histories, and a volume on technical characteristics and error estimation. Each volume will be about 500 pages long.

The user's guide will contain the essential information for routine use of the data series. It will include a general description of the public use samples, guidelines for use of the data series, record contents descriptions, a glossary of terms, and a brief summary of sample designs and error estimates.

The volume on procedural histories and comparability issues will provide a comprehensive treatment of changes in the census that affect comparability of the public use samples. We will include capsule procedural histories for all census years and complete enumerator instructions organized by variable. We will focus especially on problems of comparability that stem from differences in enumeration procedures and on changes in post-enumeration editing and processing. For the period since 1950, our principal source will be the official procedural histories prepared by the census bureau (U.S. Bureau of the Census 1955, 1966, 1977, 1986-1989). In the case of the 1940 census, we will rely heavily on a new procedural history created in conjunction with the 1940 public use microdata sample (Jenkins 1987). For the early census years, where no official procedural histories exist, we will supplement published material (e.g. U.S. Census Office 1882; Walker 1888; American Economic Association 1899; Wright and Hunt 1900; Holt 1929; Eckler 1972; Anderson 1988) with manuscript sources located at the National Archives.

The third volume of documentation will contain additional detail on many of the topics covered briefly in the user's guide, including data on verification results, approximate standard errors, and allocation statistics. We will also provide full documentation of the conversion of the original public use samples into their integrated format, with particular attention to sources of imprecision in the occupational and geographic codes. Finally, we will include a set of frequency distributions for key variables.

Release of Data

The data series will be released through the Inter-University Consortium for Social and Political Research at the end of three years. We intend to release the data in several different formats. The public use samples for the period since 1960 are divided into files according to geo-

graphic area, whereas those for the earlier census years are broken into nationally representative subsamples. Geographic organization is most convenient for state and local policy analysts and others with an interest in particular regions, but for most academic investigators the nationally representative subsamples are more useful. We will therefore make the Integrated Public Use Microdata Series available in both forms.

We also plan a compact edition of the data series. This version will maximize comparability at the expense of information. It will include only the common-format component of all composite variables, and will eliminate all variables not available in multiple census years. In addition, we will suppress information on household composition and family relationships if it would not have been available in all census years. The compact edition will be considerably simpler and smaller than the main version of the data series, and thus will be more efficient for users who do not require fine detail. Finally, we plan to release merged data files containing data from all nine census years. These files will be designed primarily for teaching purposes and for exploratory data analysis. They will contain a small representative sample of records drawn from the compact edition of each census year.

Conclusion

The decennial enumerations of the American population include a great deal of information on demography and social structure that can only be taken advantage of through the public use samples. We presently understand only the broad outlines of the social transformation that has taken place since the late nineteenth century; published sources provide only limited information on topics such as fertility behavior, urbanization, immigration, household composition, and occupational structure. The Integrated Public Use Microdata Series will allow the construction of cross-tabulations on a wide range of topics that were not covered by census publications or were incompletely tabulated. Perhaps even more important is the potential for multivariate analyses opened up by the availability of microdata. Used in combination, the nine data sets spanning a century of cataclysmic social and economic change will comprise our most important resource for the study of changing social structure.

This essay has summarized our general approach to the integration of the nine existing public use samples. We are still in the formative stages of design, so our plans should be viewed as tentative. It is our hope that potential users of the data series will provide us with as much feedback as possible before the design of the data series is cast in stone.

REFERENCES

American Economic Association (1899) The Federal Census. New York: Macmillan.

Anderson, Margo J. (1988) The American Census: A Social History. New Haven: Yale University Press.

Conk, Margo A. (1980) The United States Census and Labor Force Change. Ann Arbor: UMI Research Press.

Eckler, A. Ross (1972) The Bureau of the Census. New York: Praeger.

Graham, Stephen N. (1980) 1900 Public Use Sample: User's Handbook. Seattle: Center for Demography and Ecology, University of Washington.

Hauser, Robert M. (1982) "Occupational Status in the Nineteenth and Twentieth Centuries." Historical Methods 15: 111-126.

Hodge, Robert W., Paul M. Siegel, and Peter H. Rossi (1964) "Occupational Prestige in the United States, 1925-63." American Journal of Sociology 70: 286-302.

Holt, W. Stull (1929) The Bureau of the Census: Its History, Activities, and Organization. Washington, D.C.: U.S. Government Printing Office.

Jelatis, Virginia (1990) "Reconciliation of Occupational Codes in the U.S. Public Use Samples, 1880-1980." Paper presented at the annual meetings of the Social Science History Association, Minneapolis.

Jelatis, Virginia and Matthew Sobek (forthcoming) "Occupational Class and Socioeconomic Status in U.S. Census Data." Paper presented at the annual meetings of the Social Science History Association, New Orleans.

Jenkins, Robert (1987) Procedural History of the 1940 Census of Population and Housing. Madison: University of Wisconsin Press.

Laslett, Peter (1972) "Introduction," in Peter Laslett and Richard Wall (eds.) Household and Family in Past Time. Cambridge, England: Cambridge University Press.

Reiss, Albert J. (1961) Occupations and Social Status. Glencoe, Ill.: The Free Press.

Ruggles, Steven, (1991) "Comparability of the Public Use Files of the U.S. Census of Population. Social Science History. 15: 123-158.

Ruggles, Steven and Russell R. Menard (1990) "A public use sample of the 1880 Census of Population." Historical Methods 23: 104-115.

Sharlin, Allen (1980) "On the Universality of Occupational Prestige." Journal of Interdisciplinary History 21:115-125.

Sobek, Matthew (forthcoming) "Class Analysis and the U.S. Census Public Use Samples." Historical Methods.

Strong, Michael A., Samuel H. Preston, Ann R. Miller, Mark Hereward, Harold R. Lentzner, Jeffrey R. Seaman, Henry C. Williams (1989) User's Guide: Public Use Sample, 1910 Census of Population. Philadelphia: Population Studies Center, University of Pennsylvania.

Treiman, Donald J. (1976) "A Standard Occupational Prestige Scale for Use with Historical Data." Journal of Interdisciplinary History 7:283-304.

_____ (1977) Occupational Prestige in Comparative Perspective. New York: Academic Press.

Tyree, Andrea and Billy G. Smith (1978) "Occupational Hierarchy in the United States: 1789-1969." Social Forces 56:881-899.

U.S. Bureau of the Census (1955) The 1950 Censuses: How they Were Taken. Washington, D.C.: U.S. Government Printing Office.

_____ (1966) The 1960 Censuses of Population and Housing: Procedural History. Washington, D.C.: U.S. Government Printing Office.

_____ (1968) Changes between the 1950 and 1960 Occupation and Industry Classifications, by John A. Priebe. Technical Paper 18. Washington, D.C.: U.S. Government Printing Office.

_____ (1972a) Public Use Samples of Basic Records From the 1970 Census: Description and Technical Documentation. Washington, D.C.: U.S. Government Printing Office.

_____ (1972b) 1970 Occupation and Industry Classifications in Terms of their 1960 Occupation and Industry Elements, by John A. Priebe, Joan Heinkel, and Stanley Greene. Technical Paper 26. Washington, D.C.: U.S. Government Printing Office.

_____ (1973) Technical Documentation for the 1960 Public Use Sample. Washington, D.C.: U.S. Government Printing Office.

_____ (1976) U.S. Census of Population and Housing: 1970 Procedural History. Washington, D.C.: U.S. Government Printing Office.

_____ (1982) Public Use Samples of Basic Records From the 1980 Census: Description and Technical Documentation. Washington, D.C.: U.S. Government Printing Office.

_____ (1984a) Census of Population, 1940: Public Use Sample Technical Documentation. Washington, D.C.: U.S. Government Printing Office.

_____ (1984b) Census of Population, 1950: Public Use Sample Technical Documentation. Washington, D.C.: U.S. Government Printing Office.

_____ (1986-1989) Census of Population and Housing (1980): History, 1980 Census of Population and Housing. Washington, D.C.: U.S. Government Printing Office.

_____ (1989a) The Relationship between the 1970 and 1980 Industry and Occupation Classification Systems. Technical Paper 59. Washington, D.C.: U.S. Government Printing Office.

_____ (1989b) 1990 Census of Population and Housing. Tabulation and Publication program. Washington, D.C.: U.S. Government Printing Office.

U.S. Census Office (1882) Report of the Superintendent of the Census (November 1, 1881) Washington, D.C.: U.S. Government Printing Office.

Walker, Francis A. (1888) "The Eleventh Census of the United States." Quarterly Journal of Economics. 2: 136-61.

Wright, Carroll and William C. Hunt (1900) The History and Growth of the United States Census. Washington, D.C.: U.S. Government Printing Office.

Table 1. Availability of Variables: Public Use Samples, 1880-1990

Blank = variable not available
Y = variable available
C = can be constructed
S = sample-line, 1940 and 1950
5 = five-percent sample, 1970 pus
15 = fifteen-percent sample, 1970 PUS
N = Neighborhood samples, 1970
ST = State samples, 1970 PUS
SM = SMSA samples, 1970 PUS
a = "A" sample, 1980 PUMS
b = "B" sample, 1980 PUMS
c = "C" sample, 1980 PUMS

	1880	1900	1910	1940	1950	1960	1970	1980	1990
Geographic Information									
State	Y	Y	Y	Y	Y	Y	N,ST	Y	Y(1)
Urban/Rural residence	Y	Y	Y			Y	N,ST	c	Y
Farm identifier(2)	C	Y	Y	Y	Y	Y	Y	Y	Y
Large cities	Y	Y	Y	Y	Y		SM	a,b	Y
Modified SMA	C	C	C						
SMA				Y	Y				
SMSA							SM	a,b	Y
County or county group	Y	Y	Y	Y	Y		SM	a,b	Y
Personal Characteristics									
Age	Y	Y	Y	Y	Y	Y	Y	Y	Y
Sex	Y	Y	Y	Y	Y	Y	Y	Y	Y
Race	Y	Y	Y	Y	Y	Y	Y	Y	Y
Marital Status(3)	Y	Y	Y	Y	Y	Y	Y	Y	Y
Household Relationship	Y	Y	Y	Y	Y	Y	Y	Y	Y
Duration of Current Marriage		Y	Y		S(4)				
Age at First Marriage				S		Y	5	Y	
Number of Marriages			Y	S(5)	S(5)	Y	5	Y	
Married in past year?	Y	Y	Y	C,S	C,S	C	C	C	
Children ever born		Y	Y	S	S	Y	Y	Y	Y
Children surviving		Y	Y						
Surname code	Y		Y	Y	Y				
Subfamily relationships	Y	C	C	Y	Y	Y	Y	Y	Y
Secondary fam. relationships	Y	C	C	Y	Y	Y			
Ethnicity and Migration									
Birthplace (country, state)	Y	Y	Y	Y	Y	Y	Y	Y	Y
Citizenship/Naturalization		Y	Y	Y	Y		5	Y	Y
Parental birthplace (country)	Y	Y	Y	S	S	Y	15		
Parental birthplace (state)	Y	Y	Y	S	S				
Residence five years ago				Y		Y	15	Y	Y
Year of immigration		Y	Y				5	Y	Y
Mother tongue			Y	S	S	Y	15	Y	Y
Speaks English?		Y	Y					Y	Y
Spanish surname	Y			Y	Y	Y	Y	Y	Y
Economic Status and Employment									
Wage and salary income				Y	Y	Y	Y	Y	Y
Total income					Y	Y	Y	Y	Y
Occupation	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry	C	C	Y	Y	Y	Y	Y	Y	Y
Home ownership		Y	Y	Y		Y	Y	Y	Y
Mortgaged?		Y	Y					Y	Y
Rent/Home value				Y		Y	Y	Y	Y
Class of worker				Y	Y	Y	Y	Y	Y
Period worked in census year				Y	S	Y	Y	Y	Y
Hours worked last week				Y	Y	Y	Y	Y	Y
Period unemployed	Y	Y	Y	Y	S			Y	
Year last worked						Y	Y	Y	Y
Currently unemployed			Y	Y	Y	Y	Y	Y	Y
Education and Veteran Status									
School enrollment	Y	Y	Y	Y	S	Y	Y	Y	Y
Can read	Y	Y	Y						
Can write	Y	Y	Y						
Years of schooling				Y	S	Y	Y	Y	Y
Veteran Status			Y(6)	S	S	Y	15	Y	Y

1. Not all geographic information indicated will be available for all versions of the 1990 sample.
2. Definition of farm varies.
3. The "separated" category of marital status is not available before 1950; however, the similar category of married, spouse absent can be constructed for all census years.
4. Duration of current marital status.
5. The 1940 and 1950 censuses indicated whether married more than once.
6. Civil war veterans only.

Table 2. Constructed Household Composition and Family Interrelationship Variables

I. Household Record

Name	Description	Notes
HHCOMP90	Household composition	1990 Census basis
HHCOMP40	Household composition	1940 PUMS system
EXTFAM	Extended family classification	1900 PUS system
EXTYPE	Type of extension	1900 PUS system
UNREL	Classification of secondaries	1880 PUS system
CAMFAM	Cambridge Group classification	Laslett/Hammel

II. Person record

Name	Description	Notes
SU1970	Sampling unit of case	1970 PUS rules
FAMUNIT	Family unit membership	1940 PUMS system
FAMSIZE	Size of unit indicated in FAMUNIT	1940 PUMS system
FAMTYPE	Family type and own children	1940 PUMS system
FAMREL	Family relationship summary	1940 PUMS system
SUBFUNIT	Subfamily unit membership	1940 PUMS system
SUBFSIZE	Size of unit indicated in SUBFUNIT	1940 PUMS system
SUBFTYPE	Subfamily type and own children	1940 PUMS system
SUBFREL	Subfamily relationship summary	1940 PUMS system
SPLOC	Spouse pointer (if present)	
POPLOC	Father pointer (if present)	
MOMLOC	Mother pointer (if present)	
NCHILD	Number of own children	
NCHLT5	Number of own children under 5	
ELDCH	Age of eldest own child	
YNGCH	Age of youngest own child	