# The Public Use Microdata Samples of the U.S. Census: Research Applications and Privacy Issues

**Minnesota Population Center**
**and**
**Inter-University Consortium for Political and Social Research**
**Census 2000 Advisory Committee**

Prepared for
Census 2000 Users' Conference on PUMS
Hilton Alexandria Mark Center
Alexandria, VA

May 22, 2000

This report was prepared by Steven Ruggles, Distinguished McKnight University Professor at the University of Minnesota and chair of the ICPSR Census 2000 Advisory Committee, with the assistance of Catherine Fitch, coordinator of the IPUMS project at the University of Minnesota and Matthew Sobek, managing editor of *Historical Statistics of the United States* at the University of California Riverside. The members of the committee include Ilona Einowski, Assistant Director of the University of California Data Archive; W. Reynolds Farley of the Institute for Social Research, University of Michigan; Halliman H. Winsborough, Interim Director, ICPSR and Bascom Professor of Sociology at the University of Wisconsin; Erik Austin, Director, Archival Development, ICPSR; and Peter Granda, Assistant Archival Director, ICPSR.

## Table of Contents



NOTE: Due to their length, the appendices have been removed from this version of the report.

For a full version of the report, including appendices, please download this file (1.01 Mb).

For the HTML version of the report, please visit our web site

## Summary

The Census Bureau is considering significant reductions in the level of subject detail for the Public Use Microdata Samples (PUMS) of the 2000 Census in order to reassure the public about respondent confidentiality. In particular, the Bureau is considering a plan that would:

- group ages into five-year categories for persons aged 65 or older and reduce the topcode for age from 90 to 85;
- reduce the number of ancestry categories from 560 to 105;
- reduce the Hispanic origin categories from 206 to 23;
- reduce the number of identified occupational groups from 505 to 67;
- reduce the number of identified industry groups from 244 to 70;
- reduce the number of language categories from 393 to 83;
- eliminate 298 foreign countries of birth and substitute 14 continents and U.S. possessions.

To provide feedback to the Census Bureau on the potential impact of these changes on the academic research community, the Minnesota Population Center and the ICPSR Census 2000 Advisory Committee carried out a survey of 1,006 users of the Integrated Public Use Microdata Series (IPUMS). This report summarizes the reaction of researchers to the proposed changes and recommends an alternate strategy to ensure respondent confidentiality.

The research community expressed alarm at the proposal to reduce the level of detail in the PUMS. Survey respondents described hundreds of ongoing research projects that will have to be abandoned if the changes go forward in their current form. With remarkable and perhaps unprecedented unanimity, the academic community urges the Census Bureau to maximize historical compatibility of the 2000 census files and to avoid making precipitous decisions that will do permanent damage to social science research in the United States.

The ICPSR Census 2000 Advisory Committee recognizes that some census categories that appeared in the 1990 PUMS are too small, and may pose at least a theoretical, if not a practical, risk of disclosure. These gaps in the confidentiality safety net should be closed. There are some categories in the 1990 PUMS files that represent a national population of fewer than 1,000 persons. But the changes under consideration by the Census Bureau are excessive. Under the proposed system, most detailed categories would represent upwards of a million individuals in the general population. This thousand-fold reduction in the finest available level of detail would severely compromise the scientific utility of the data for academic and policy research.

The goal of ensuring that specific individuals cannot be identified in the census would be equally well served by imposing a minimum population threshold of 10,000 or 25,000 persons for every sensitive census category. Such a 10 to 25-fold reduction in detail for the smallest census categories for 1990 would result in a substantial loss of information, and some researchers will oppose such a drastic reform. Nevertheless, this alternative is far preferable to the radical change under consideration by the Bureau. If a minimum population threshold is adopted and the collapsing of categories is carried out with sensitivity to historical compatibility, the PUMS will remain the single most important source in American social science. It would also provide unprecedented security for respondent confidentiality.

We also recommend that the Bureau close one remaining confidentiality gap. There are a small percentage of cells in the tables of the summary files that contain a single individual when the tables are aggregated over an entire PUMA. We recommend that the Bureau suppress the PUMA codes for such individuals when they appear in the PUMS. This would foreclose the possibility of identifying any particular individual in the PUMS, but would not significantly reduce the precision of the samples because the number of excluded cases would be extremely small.

In the 36-year history of the census microdata samples, no respondent has ever been identified by anyone outside the Census Bureau. We consider such identification to be highly improbable even under the current standards. A minimum threshold for identification of sensitive population categories may be justified, however, on the grounds that it would reassure the public that the risk of disclosure is negligible.

The PUMS files are the crown jewels of American social science. They generate more population research than any other data source, and are the only available data source for a wide range of topics that bear on pressing policy issues. The ICPSR Census 2000 Advisory Committee strongly recommends that the Bureau proceed cautiously, and delay a final decision on the redesign of classifications until the long-form data are available for analysis. The current (as of 22 May 2000) classifications appear to be arbitrary, unsystematic, and hastily prepared. We further recommend that any reduction of detail maximize historical compatibility and be carried out in close consultation with the user community, even if that means a significant delay in the release of the data.

## A. Background and Significance

### A1. What is Census Microdata?

Most population data are available only in aggregated tabular form. The PUMS are microdata, which means that they provide information about individual persons and households. This makes it possible for researchers to create tabulations tailored to their particular questions. Since the PUMS files include nearly all the detail originally recorded by the census enumerations, users can construct a great variety of tabulations interrelating any desired set of variables. The flexibility offered by microdata is particularly important for historical research because the aggregate tabulations produced

by the Census Bureau are often not comparable across time, and until recently the subject coverage of census publications was limited.

Microdata do pose some limitations, however. Most important, for the period since 1940 census microdata are subject to strict confidentiality measures that limit their usefulness for some applications. The available samples for these years include no names, addresses or other potentially identifying information. To further ensure that no individuals can be identified, the Census Bureau limits the detail on place of residence, place of work, very high incomes, and several other variables. Most important, the microdata records for the period since 1940 identify no geographic areas with fewer than 100,000 inhabitants (250,000 in 1960 and 1970).

## A2. Historical Background

The founding fathers of the United States envisaged the census not simply as a tool for apportioning representatives, but rather as a means of gathering information about American society. James Madison, the primary author of the constitution, argued forcefully that the census should "embrace some objects besides the basic enumeration of the population." Madison pressed for the enumeration of occupations, since it would provide the "kind of information . . . all legislatures had wished for" and "it would give . . . an opportunity of marking the progress of society, and distinguishing the growth of every interest . . . this would furnish ground for many useful calculations" (Magnuson 1995: 12-13). After the first census, Thomas Jefferson and his fellow members of the American Philosophical Society lobbied for the expansion of detail on age, birthplace, and occupation "in order to ascertain more completely the causes which influence life and health" and "the conditions and vocations of our fellow citizens" (Magnuson 1995:15).

From 1790 to 1950, the census expanded dramatically, both in terms of the questions asked and the number of tabulations produced. But the basic mode of access to census data remained unchanged: throughout that period, census results consisted of counts of the number of persons in each geographic area who had a particular characteristic or combination of characteristics. The advent of electronic computers, both within the Census Bureau and on university campuses, allowed a ground-breaking shift in this paradigm for 1960. In an effort to meet the needs of scholars who required specialized tabulations, the Census Bureau created a 1 in 1000 extract of the basic data tapes they had used to create tabulations for the published census volumes (U.S. Bureau of the Census 1964). To preserve confidentiality, the Census Bureau removed names, addresses, and other potentially identifying information.

The 1960 public use sample revolutionized the analysis of the American population and led to an explosion of new census-based research. Not only did it allow researchers to make tabulations tailored to their specific research questions, but it also allowed them to apply new methods to the analysis of census data, especially multivariate techniques. But the sample did have two significant limitations. First, the sample size was relatively small. The 1 in 1000 sample density yielded about 180,000 person records. Given the modest capacity of computers in 1964, this was a lot of cases, but as

researchers began to use the sample for detailed analysis of small population subgroups, its limitations became apparent. Second, the 1960 public use sample provided highly limited geographic information. To ensure confidentiality, the Census Bureau stripped off all information on places below the state level. This meant, for example, that it was impossible to extract a subsample of the New York City population.

Both of these problems were addressed by the 1970 public use samples. The 1 in 1000 density of the 1960 sample was increased dramatically; the Census Bureau provided six independent public use samples for 1970, each of which had a 1 in 100 density. Users who required an exceptionally large number of cases could combine the samples to obtain a six percent density, or about 12 million person records. In addition, the 1970 samples provided a variety of alternate geographic codes, although the Census Bureau still did not identify any places of less than 250,000 population.

In conjunction with the 1970 public use samples, the Census Bureau released a new version of the 1960 public use sample. The Bureau enlarged the sample density from 1 in 1000 to 1 in 100, and at the same time reorganized the coding schemes and record layouts to be compatible with the samples from 1970. This compatibility made it relatively easy for investigators to pool data from 1960 and 1970, and thus incorporate change into their analyses.

By the late 1970s, the public use samples for 1960 and 1970 had become one of the essential tools of American social scientists. It was in this climate that Halliman Winsborough and a group of others at the University of Wisconsin developed the idea of creating historical public use samples for earlier census years. They obtained funding from the National Science Foundation and contracted with the Census Bureau to create 1 in 100 samples for the censuses of 1940 and 1950 (U.S. Bureau of the Census 1984a, 1984b).

In addition to the census microdata samples covering the period 1940 through 1970, the Census Bureau has released samples for the 1980 and 1990 censuses, and plans to create a sample of the 2000 census (U.S. Bureau of the Census, 1982a). The 1980 and 1990 samples included significantly greater geographic and subject content detail than either the 1960 or 1970 public use samples. We now have a continuous series of Census Bureau microdata samples for six census years consisting of anonymized records spanning the period from 1940 through 1990, and the much-anticipated 2000 PUMS will give researchers a seventh sample in the series in 2002 or 2003.

The series of national census microdata files also extends backwards to the more distant past. The individual-level enumerations of the Census are released to the public after an interval of 72 years. Accordingly, historical demographers have created large national samples of these census enumerations for the period from 1850 to 1920.

The Integrated Public Use Microdata Series (IPUMS), created in 1995, harmonizes these early census samples with the seven Census Bureau samples covering the recent period (Ruggles and Sobek 1998). The combination of free and open access, a

user-friendly access system, and integrated comprehensive hypertext documentation has attracted many users to the IPUMS. Since 1995, the IPUMS project has distributed over two terabytes of IPUMS data to users around the world. The project is now distributing about 95 gigabytes of data per month, or an average of 130 megabytes per hour, 24 hours per day. The IPUMS automated data-extraction system has prepared approximately 20,000 custom extracts of IPUMS data since May 1996 for about 3,000 researchers around the world, and is now processing approximately 1,000 data extract requests per month. Even though the database has only been available for only five years, there is already a substantial body of IPUMS-based research. To date, the database has been used in 120 articles, 3 books, and 35 Ph.D. dissertations as well as hundreds of conference papers and research reports. Many of these articles appeared in leading journals such as the *American Economic Review*, the *American Sociological Review*, the *American Historical Review*, *Social Forces* and *Demography*. Most of these studies use the IPUMS to assess recent change in areas of current public policy concern, so compatibility of the 2000 PUMS with the earlier samples is critical.

## A3. Strengths of the PUMS Files for Social Science Research

The PUMS have become a mainstay of American social science. Among population scientists, the PUMS files are the single most important source of data. Among articles published in *Demography*—the leading journal in the field—PUMS data was used 36% more often than the next most important source during the period 1994-1998 (Ruggles 1998). The need for PUMS data is not limited to demographers. Economists represent the single largest group of IPUMS users, and according to Joshua Angrist, an economics professor at MIT, "In my view, the PUMS is the most important research dataset in economics." PUMS data are also a mainstay of academic research in the fields of sociology and history, and they are an indispensable resource for urban planners and policy analysts.

Why are the PUMS files so widely used? The national census files incorporated in the existing IPUMS database have three key strengths: broad chronological scope, large sample populations and fine detail. Social scientists have increasingly recognized that we cannot understand contemporary social behavior without investigating processes of change. It is the relative continuity of Census Bureau classifications that allows for consistent comparisons across many decades. The PUMS is the only source of microdata that allows researchers to assess the effects of policy changes on the American population across significant periods of time or between cohorts, and the IPUMS design makes such investigations comparatively simple.

The second strength of the public use census files is their large size. The number of cases available for each census year ranges from the hundreds of thousands to the tens of millions. This allows the study of small and geographically dispersed population subgroups. Even the largest surveys are too small to allow analysis of small population subgroups such as Native Americans or particular occupational groups. Moreover, there are presently no national surveys large enough to be used for policy research at the municipal level. Finally, the large size of the PUMS files together with their national

coverage permits multi-level analyses of the effects of local conditions on individual and family behavior.

The third strength, fine detail, is an essential complement to both the chronological scope and large scale of the database. Without detailed categories, it would be impossible to make the datasets compatible over the long run. Moreover, it is the detail of the samples that allows us to identify small population subgroups, and to capitalize on the large scale of the PUMS. Therefore, without detailed population categories, the PUMS files would have limited application for either the analysis of change or the study of population subgroups.

## B. Survey Results

### B1. Design and Execution

To clarify the specific needs of researchers, the Minnesota Population Center conducted a survey of PUMS users during the week of May 9-16, 2000. We emailed a request to fill out the survey to approximately 1,400 persons who recently registered to use the IPUMS data extraction system. The email message and survey form are reproduced in Appendix A. We received 440 responses within 48 hours and 1,006 responses overall. It is difficult to calculate a precise response rate to our request, because we received many responses from persons other than the 1,400 we asked, but we estimate, based on name matching, that approximately 60% of the responses came from registered IPUMS users, yielding a response rate of approximately 43%.

149 survey responses, or 14.8 percent of the total, came from persons who indicated that they had used the PUMS for one or no studies. We eliminated these responses so that we could focus on experienced users of the data. Appendix B Tables 1-3 give descriptive statistics on the remaining respondents. Over two thirds of respondents were academic users, and most of the rest were policy researchers. The largest group of respondents is faculty members, followed by graduate students and nonacademic researchers. The most important fields of research are economics, demography, and sociology. The geographic distribution of respondents was very broad; about two-thirds of respondents came from institutions with five or fewer respondents. Table 4 lists the 28 institutions with six or more respondents. The list includes most of the leading institutions in demography, empirical economics, and public policy.

### B2. Statistical Results

Detailed survey results are presented in Appendix B Tables 5 through 17. The results suggest a remarkable consensus on the importance of the PUMS and the need to preserve as much detail as possible. At the time we prepared the survey, the particulars of the Census Bureau proposal were not yet available, but the survey results are nonetheless clear. Two-thirds of the faculty respondents indicated the PUMS were indispensable for research in their field, and 96 percent of these experienced researchers indicated that historical comparability was very useful or indispensable for work in their field.

The results indicate a clear preference for maximum detail, especially among academic researchers. The respondents were especially concerned with geography, income, and age variables, but approximately 90% of faculty researchers also said that a reduction in the detail available on occupation and race would have "catastrophic" or "very harmful" consequences for research in their field. Moreover, there is no consensus at all on which aspects of occupation should be preserved if the detail of occupation were to be substantially reduced; a plurality think the occupational classification should maximize comparability with earlier classification systems, but many others think the system should focus on socioeconomic status or type of work. It is therefore clear that no single broad classification could meet the needs of most researchers.

At the time the survey was designed, we did not know that the redesign of the PUMS called for eliminating all specific foreign countries of birth. If we had included a query on this proposal, we expect that it might have generated more concern than any other aspect of the proposed plan.

Perhaps the most revealing quantitative indicator was our question that posed a trade-off between promptness of data release and detail of categories. The question was worded as follows:

> "The Bureau might make the decisions on the reduction of detail before the data are processed or they might wait until afterwards to allow analysis of the data to determine the need for confidentiality. The latter strategy might allow greater detail, but could result in a significant delay in release of data. I recognize that one would need to know more to make an informed decision, but in general, would you prefer a less detailed dataset released more promptly, or would you be prepared to wait if it might mean a more useful dataset?"

Over 90 percent of all respondents and 94% of faculty respondents indicated that they would rather wait for the release of the PUMS, if this delay offered some prospect of greater detail in the data. The consensus on this point is unequivocal.

### B3. Qualitative Results

The qualitative results were even more revealing than the quantitative ones. The first qualitative question asked respondents "Please comment below on ways in which reduction of PUMS detail might affect your research. Be specific as possible. For example, if you specialize in aging research, comment on the sorts of analyses that would be precluded by grouped age data." We had hundreds of responses to this inquiry, and they are reproduced in full in Appendix C. Many respondents focus their research on particular population subgroups, which they might be unable to identify in a less-detailed sample. Some examples include studies of immigrant groups within metropolitan areas, the changing demographics of hotel and restaurants workers in San Francisco and Los Angeles, the legal profession, and specific Asian-Pacific populations. Policy makers in

particular expressed concern about the loss of the detail needed to study issues such as school voucher programs, welfare reform, residential segregation, urban poverty, and measures of social and economic inequality at the local level.

A few representative examples make the cost of the possible changes palpable. Robert Hauser, a demographer at Wisconsin, wrote that:

> As far as I am concerned, elimination of the detail of age, race, ancestry, income, occupation, and geography would essentially eliminate the value of data from the long form. This is a shameful, cowardly, and ludicrous proposal. I hope it will disappear promptly and not be raised again.

The respondents detailed problems that would arise in several research areas. One of the most frequently mentioned topics was aging. Tony Deitz, a social gerontologist from Central Florida, focuses on the damage that would be done by grouping the ages of the aged:

> I am a social gerontologist who focuses primarily upon socioeconomic and racial/ethnic minority status in my research. I am interested in and conduct research on a number of areas within gerontology. Recently, for example, I have written manuscripts that deal with mortality among older adults by ethnic group status. It is imperative that I be able to differentiate between people of different ages (i.e. 65 to 66, even). This is particularly important when we begin to look at age effects by ethnic group membership because such things as the mortality crossover effect is believed to occur at different ages for different ethnic groups (even, for example, between different Hispanic populations). So, given that the census represents our nation's most comprehensive survey of the population, it is very important that as much detail as possible be retained so that reliable, valid estimates and reports can be made.

Caroline Hoxby, a Harvard economist, stresses the use of the PUMS for the analysis of education:

> For analyzing the rate of return to education, it is absolutely necessary to know a person's exact age, so that we can associate him with the regime (laws, funding) under which he attended school. We also need the exact age for estimating consistent wage equations.

Many respondents expressed concern about occupational categories and income information. Charles Nam, the developer of the Nam-Powers socioeconomic scale of occupations, writes:

> The scale depends on a detailed occupation classification (although comparability with the occupational classification in earlier censuses is not important), as well as a fair amount of detail on education and income (determinants of the score for each detailed occupation). Without the detail for those three variables, it would not be possible to create the scale

> (nor would it be possible for competing scales, e.g., Duncan SEI, to be constructed).

Donald Bogue, a Chicago demographer, writes about detail across several categories:

> The need to introduce age cohorts into demographic analysis requires that age groups be as refined as possible. The occupational categories used in 1990 were about as minimal as acceptable. Income distributions need detail in order to meet the needs of a variety of studies, but also to adjust income distributions between censuses for comparability. Race and ethnicity are powerful variables in differentiating the population (expecially for change over the 1990-2000 decade), and the detail sould be as great as for 1990, and as comparable as possible.

We also asked respondents for their thoughts about the confidentiality issue. The voluminous results are reproduced in Appendix D. A few common themes run through the responses. Many researchers echo the comment of Halliman Winsborough of the University of Wisconsin: "The first PUMS file came out about 35 years ago. Has anyone been identified? I think not." Bogue comments

> I have never heard of a case of a breach of confidentiality using PUMS data. Is there a documented case on record? If so, what privacy was breached and how serious was it? Abuse of privacy of credit, bank, and other records is rampant, with little effort to control or regulate it. Persons seeking information about particular others would find PUMS about the least productive entry-point for gaining personal information.

Many also pointed out the impossibility of identifying individuals in the PUMS with certainty. Russell Davis, a graduate student at Louisiana State University, writes "I have it open and running right now, and I would challenge anyone to find someone they know." It was also commonly noted that it does not make sense that anyone would attempt to identify someone in order to learn their income, since financial information on specific individuals is readily available on the Internet at modest cost.

A number of respondents discussed the politics of privacy and the census. For example, Maseo Suzuki, an economist at the Public Policy Institute of California, wrote

> I am concerned that the Census Bureau may be under political pressure, since I have read newspaper accounts of Congressmen and/or Senators questioning the need for census data. I think that there is a lot of public concern about privacy, but this needs to be directed at its true source, private-sector internet businesses, and not redirected at academic researchers who would have no interest in trying to identify individuals.

Marcus Stanley, a graduate student in economics at Harvard, had this to add:

> It seems to me this is a political stampede without much reason behind it. I know there have been a few people trying to whip up concern about the long form, but I don't know of any case where Census data has been used to identify individuals. . . it is just not a reasonable thing to try to do.

### B4. The Research Data Center Alternative

One possible solution to the problem of confidentiality would be to restrict access to the PUMS to the five Census Bureau research data centers. Although we neglected to ask our respondents if this would be an acceptable solution, several commented on it anyway. For example, an anonymous faculty member wrote:

> I am increasingly concerned about efforts to limit availability of high quality data such as PUMS to a broad spectrum of social scientists and other users, and to eliminate important information out of concern about privacy violations. The chances of such violations appear trivial. Insisting on secure-access sites is not a very appealing solution because doing so greatly disadvantages persons not located near such facilities. Restricting access or eliminating important information does not serve the broader purposes collecting such data - better understanding of social conditions and problems, information about the effect of public policies on social conditions, and insights into how to improve matters.

We believe that restricting detailed PUMS to the Research Data Centers would drastically reduce scholarly and policy use of the data. The cost of using a research data center is very high. For example, the two California centers charge $3,125 per seat per month. Travel, per diem expenses, and housing costs could easily double this figure for anyone not located in a research data center city. Thus, a large research project, involving several person-years of data analysis—as is typical for dissertation projects—could easily cost $200,000 simply to obtain access to the data. Major projects involving substantial manipulation, such as merging of confidential and public data sources, could cost many times more. We estimate that if we had carried out the IPUMS project at a Research Data Center, it would have added upwards of a million dollars to the cost. In fact, because of the logistics involved, we almost certainly would not have attempted it.

Because of the cost barrier, use of the centers will be restricted to a small number of well-funded researchers, and those researchers will be discouraged from use of the data by the inconvenience of travelling to another city. Even if the cost and travel issues could be overcome, the RDC's would have to be expanded at least 100-fold to accommodate the current number of PUMS users. Accordingly, we do not regard this as a viable strategy.

## C. Conclusions and Recommendations

The revision of the PUMS under consideration by the Census Bureau represents the most radical modification of the data series since it was introduced in 1964. It would undermine studies of such key topics as occupational structure and ethnicity, even as they are changing more rapidly than at any time in our history. If the proposed revision were to be carried out, it would represent the most damaging withdrawal of public data in the history of the Census Bureau, and would sharply reverse the 200-year tradition of

continuous improvement in Census Bureau data products which began at the urging of Madison and Jefferson.

We do not believe that there is an adequate scientific basis to justify such a drastic change. There is no evidence that the confidentiality protections of the past 36 years have been in any way inadequate. We know of not a single instance in which privacy has been breached. We find it implausible that anyone would turn to PUMS data files to attempt to uncover sensitive information about particular individuals. Even in the unlikely event that one could locate a unique exact match for a target individual, one could never be certain that the case actually represented that individual. The PUMS files are samples, and there will always be the possibility that another person exists, not included in the sample, who is also an exact match. The only exception to this is the comparatively rare cases in which a unique individual can be identified in a summary file, and that individual also appears in the PUMS. As proposed in the introduction, we suggest that the Bureau suppress the PUMA codes for such individuals, which would close this potential loophole.

Privacy in America is indeed under assault. There are now on the Internet some 500 web sites promising full investigative reports on any individual—including credit ratings, property records, marital status, and other information—for fees ranging from $35 to $150. Given this wealth of precise information readily available from a wide variety of private sources, it is simply not plausible that anyone would attempt to crack the PUMS security measures in order to obtain uncertain and outdated data. Thus, we think that fears of census disclosure are greatly exaggerated.

We believe that the risk to privacy posed by the PUMS must be weighed against the social cost of restricting access to information. That cost is very great. The proposed changes would cripple the efforts of social scientists to understand the radical social and economic changes taking place in American society. For example, the study of immigration will be devastated, as not a single foreign country of birth is identified under the proposed design of the 2000 PUMS. Similarly, we will be unable to gauge the transformation taking place in American occupational structure, since if the Census plan is adopted we will no longer be able to identify such key groups as carpenters, plumbers, waiters, bus drivers, licensed practical nurses, family child care providers, electrical engineers, and bartenders. Even lawyers could not be uniquely identified. And without single-year age information on the elderly, efforts to understand the implications of rapid population aging will be seriously handicapped, since there is simply no other data source large enough to examine the issue. The recodes proposed by the Bureau to date appear arbatrary and unsystematic. Shoshone Native Americans and persons of Albanian descent are to be identified, but not persons of Mexican or Chinese birth. There is no explanation of such anomalies, and no description of the criteria the Bureau is employing to define categories.

We recognize that perceptions are important; we have as much interest in ensuring a high census response rate as does the Bureau itself. But we see no historical evidence that confidentiality measures taken in the past have had any impact on census

response rates. When the Census Office first made responses confidential in the 1890 Census, for example, the undercount did not change. And when the threshold for identifying geographic areas in census microdata was lowered from 250,000 to 100,000 in 1980, the change did not even register with the public. It seems implausible that the technical changes proposed by the Bureau would improve response rates in the future. Nor would they satisfy opponents of the Census. We expect that the only real effect of the proposed changes would be to prevent social scientists and policy analysts from doing their research.

Still, we think it is reasonable to collapse small subject categories. We therefore propose that sensitive variable categories that represent small populations—perhaps those that represent fewer than 10,000 or 25,000 persons—be collapsed into larger categories. Such a criterion would significantly reduce the detail of information available in the PUMS. If the 25,000 population criterion were imposed in 1990, for example, it would have eliminated 110 occupational categories, 180 industry categories, 194 language categories, 148 birthplace categories, and 140 ancestry categories. It would nevertheless allow the great majority of ongoing investigations to proceed.

The ICPSR Census 2000 Advisory Committee is also very concerned about historical comparability. One option under consideration by the Bureau is to create a one-percent "national" sample with greater subject detail but sharply reduced geographic detail. We would endorse this proposal if the national sample were to replicate the 1990 subject area classifications precisely and identified ten-PUMA geographic areas incorporating at least one million persons. This file could then serve as a bridge between the PUMS files of the past and the reduced-detail subject classifications in the Census 2000 five-percent sample.

The Bureau has built a laudable reputation over the past 150 years for scientific integrity. It has repeatedly resisted political pressures that would compromise its mission. That reputation is one of the Bureau's greatest assets, and we believe it would be a mistake to sacrifice it merely for the sake of appearances. The Bureau should not allow unreasonable fears to undermine the social science and policy infrastructure of the nation. The risk to confidentiality in the current PUMS samples is minimal; with measured steps to collapse very small categories, the Bureau can eliminate any practical risk of disclosure without foreclosing essential scientific and policy research applications of the data.

## Bibliography

Magnuson, Diana (1995). "The Making of a Modern Census: The United States Census of Population, 1790-1940." Ph.D. Dissertation, University of Minnesota.

Ruggles, Steven (1998). "Demographic Data and Data Dissemination in the New Millennium." Paper presented at the annual meeting of the Association of Population Centers, Albany, November 1998.

Ruggles, Steven, Matthew Sobek, et al. (1998) "IPUMS-98: Integrated Public Use Microdata Series" Minneapolis: Historical Census Project.

U.S. Bureau of the Census. (1964). Census of population and housing, 1960 public use sample: one-in-one-thousand sample. Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census. (1972). *Public Use Microdata Samples of Basic Records from the 1970 Census: Description and Technical Documentation*. Washington, D.C.: U.S. Government Printing Office.

U.S. Bureau of the Census. (1973). *Technical Documentation for the 1960 Public Use Microdata Sample*. Washington, D.C.: U.S. Government Printing Office.

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

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

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

U.S. Bureau of the Census. (1992). *Census of Population and Housing, 1990: Public Use Microdata Sample U.S. Documentation.* Washington, D.C.: U.S. Government Printing Office.