New Projects of the Minnesota Population Center

An Introduction

STEVEN RUGGLES
Minnesota Population Center
University of Minnesota

This is the third theme issue of *Historical Methods* devoted to historical data projects at the University of Minnesota. In winter 1995, the journal published a double-length issue entitled "The Minnesota Historical Census Projects" (Ruggles and Menard 1995). That issue introduced the Integrated Public Use Microdata Series (IPUMS), the series of U.S. census microdata we had released just a few months earlier. We also described our procedures for creating national samples of the 1850, 1880, and 1920 censuses. In summer 1999, a second theme issue, "IPUMS: Integrated Public Use Microdata Series," provided an update on the IPUMS project, describing the second release of the database (Ruggles and Hall 1999).

The IPUMS is a coherent series of representative samples drawn from the U.S. censuses of the period from 1850 to 2000. Altogether, the data series describes the characteristics of approximately 70 million persons. The data and documentation of the IPUMS are harmonized, making it easy to use multiple census years simultaneously. Since 1996, we have distributed the IPUMS through an online data access system that allows users to pool censuses, select subpopulations, and select variables for analysis.

The primary goal of the IPUMS project was to stimulate quantitative historical research, and the objective has been met. Since the publication of the 1999 issue, over 7,000 new users have registered to use our data access system, and they have created some 45,000 customized data sets.[1] This massive data distribution is bearing fruit. In the 1999 issue, we bragged that the IPUMS had "already served as the basis for three books, ten completed dissertations, and over forty-five articles" (Sobek and Ruggles 1999, 108). Three years later, the IPUMS bibliography lists 23 books, 80 dissertations, 261 articles, and hundreds of working papers, conference presentations, new articles, and courses (http://www.ipums.org/usa/research.html). IPUMS-based research has appeared in the leading journals of economics, sociology, and history, including *Demography* (14 articles), *Journal of Political Economy* (6 articles), *Social Forces* (5 articles), *American Economic Review* (5 articles), *Quarterly Journal of Economics* (6 articles), and *American Sociological Review* (5 articles).

The present issue—separated into Part One (Winter) and Part Two (Spring)—reports on new historical infrastructure projects launched in the past three years that build on our experience with the IPUMS. The new projects have the same basic goals as the IPUMS project: in each case, they seek to make new historical data available on the Web in harmonized form with extensive documentation of comparability issues. By making these historical resources freely available and easy to use, these efforts aim to democratize access to large data collections.

We are continuing to add new files to the existing IPUMS database. These include samples of slave schedules for 1850 and 1860 (described in Alexander et al., pp. 21–26 in Part One of this issue) and the U.S. Current Population Surveys for the period 1962–2002 (described in King and Tertilt, pp. 35–40 in Part One of this issue). In addition, we are increasing the size of some of the IPUMS samples. For the past several years, we have been augmenting the small samples of the 1900 and 1910 censuses originally created by Samuel Preston. The new samples will include 1 percent of the population, which is the same sample density as the other IPUMS census years between 1850 and 1960.

A major new infrastructure project is extending the IPUMS paradigm to aggregate statistics. Summary data are essential for the study of the spatial organization of population.[2] Except for the most recent census years, however, aggregate summary data for the United States are difficult to use. The data are stored in a bewildering array of different formats, some of which are obsolete. Moreover, the files are poorly documented and are scattered across many data archives. The National Historical Geographic Information System (NHGIS) will address these issues by gathering, reformatting, and documenting machine-readable aggregate census data for the period since 1790. The project will then make the statistics, maps, and documentation available through a Web-based data access tool (see Fitch and Ruggles, pp. 41–51 in Part One of this issue).

Another important initiative is the expansion of the IPUMS model to include international census microdata. Machine-readable microdata covering the period since 1960 exist for most countries of the world, but with few exceptions these data are not available to scholars. The IPUMS-International project (see the articles by Ruggles et al. and Esteve and Sobek in Part Two of this issue) seeks to preserve and disseminate samples of these censuses. A second international effort, the North Atlantic Population Project (NAPP), is a collaboration of researchers in Canada, Great Britain, Iceland, Norway, and the United States. The NAPP is exploiting a unique collection of individual-level census data to compile a harmonized database comprising the entire population of all five countries in the late nineteenth century (see both NAPP articles by Roberts et al. in Part Two of this issue).

We describe each project separately in the rest of this issue, but several common themes are worth highlighting. The following sections address three issues: expanding the scale of available historical microdata, rationalizing census geography, and improving the technical infrastructure for data access.

Developing Large Microdata Samples

The 1 percent samples that make up most of the current IPUMS are too small for many research projects. Since 1980, the Census Bureau has released 5 percent samples of each census. Although originally intended primarily as tools for state and local policy analysis, the 5 percent files have also become a principal source for national research as computing costs have declined. Since 1990, 80 percent of *Demography* articles based on 1980 or 1990 census microdata have used the 5 percent samples, and during the past three years these large samples have become the most widely used data source in that journal. Much of this research focuses on population subgroups too small to analyze effectively with the 1 percent files, such as American Indians, father-only families, same-sex couples, and the grandchildren of immigrants. Moreover, the large samples permit the use of innovative methods: to take just one example, they have allowed demographers to carry out national multilevel contextual analyses by making it feasible to assess the characteristics of small geographic areas.

The extensive use of the 5 percent files for the 1930 and 1990 censuses has shown the value of very large samples, and the success of the IPUMS has shown the importance of historical census microdata. Heeding these lessons, we are pursuing several strategies to develop new high-density samples for the period before 1930. Each of these strategies is based on collaboration with public or private partners to develop cost-effective approaches to the creation of large census microdata collections.

- As described in Goeken et al. (pp. 27–34 in Part One of this issue), we are collaborating with the Church of Jesus Christ of Latter-day Saints to create a complete-count database of the U.S. census of 1880, including information on all 50 million persons who were enumerated. Similar projects in Great Britain and Canada form the core of the North Atlantic Population Project mentioned above.
- In cooperation with the U.S. Census Bureau, we have embarked on an even larger scale project involving over a billion long-form and short-form census microdata records for the period since 1960. As detailed in Ruggles et al. (pp. 9–19 in Part One of this issue), our goal is to make these restricted data available to researchers in IPUMS-compatible format through the Census Bureau Research Data Centers. We hope that this project will also lead to new public census data products, including a 5 percent sample of the 1960 census.
- As noted in Ruggles et al. (in Part Two of this issue), we have agreements with the national statistical offices of 16 Latin American countries to draw new 10 percent samples of recent censuses as part of the IPUMS-International project.
- Most recently, we have embarked on an exciting new collaboration with ProQuest Information and Learning to develop new 5 percent samples for the 1900 and 1930 census years. ProQuest is keying a substantial portion of the 1930 census records to construct a genealogical index. We have contracted with the company to key every variable for a 5 percent sample of the 1930 census, including some 6 million cases. Because of efficiencies gained by creating the sample and the genealogical index simultaneously, the per-case cost of the sample will be reduced by a factor of four compared with traditional methods. Data transcription for 1930 is now in full-scale production and will be completed by early 2004; we plan to release a nationally representative sample of the data shortly thereafter. We have applied for funding to collaborate with ProQuest on a similar project to raise the sample size for the 1900 census sixfold. If funded, this sample will fill a half-century gap in large samples between 1880 and 1930.

Rationalizing Census Geography

Among the most frustrating issues facing researchers who use census data for the analysis of social and economic change is the incompatibility of geographic information over time. The Census Bureau has revised the geographic units identified in the census in every decade. The problem of geographic incompatibility over time hinders geographic analysis of both census microdata and aggregate summary files.

We are investing substantial resources to help address the problem of geographic incompatibility. As part of the NHGIS project, we are developing compatible electronic boundary files for census tracts from 1910 to 2000 and for counties from 1790 to 2000. The availability of electronic boundary files will allow us to interpolate comparable estimates of aggregate population characteristics of small areas over time.

Building on the NHGIS cartographic database, the IPUMS Redesign project (Ruggles et al., pp. 9–19 in Part One of this issue) is developing boundary files for every geographic unit identified in public census microdata covering the period since 1940, including state economic areas, county groups, and public use microdata areas. We will use these files to improve the compatibility of geographic codes across census microdata files, locating all places that can be consistently identified in multiple census years. Similarly, we will design a consistent small-area identifier for use with the large Census Bureau restricted-access census microdata files mentioned above.

Finally, we have undertaken an ambitious effort to revise geographic coding in the IPUMS samples for the period before 1940. The raw data for these census years identifies every census-defined place, but the current IPUMS samples provide codes for only a fraction of census places. Our goal is to provide full geographic detail for every census year by developing and applying a compatible classification system. Our system will be compatible with the Federal Information Processing Standards (FIPS) coding schemes for minor civil divisions and incorporated places, modified to accommodate historical change in places identified by the census.

Infrastructure for Data Access

By the standards of the Web, the IPUMS data access system is antique. When Todd Gardner designed the first IPUMS data extraction system in November 1995, most Web sites were read-only. This was only a year after the first meeting of the World Wide Web Consortium and just months after the appearance of such early commercial Web sites as yahoo.com and amazon.com. Gardner's extraction tool allowed users to combine data from multiple censuses, select a specified set of variables, and extract subsets of the IPUMS based on population characteristics (Ruggles, Sobek, and Gardner 1996). Over the next three years, IPUMS staff incorporated complete hypertext documentation into the extraction tool, making it an integrated system for access to both data and metadata.

Although we have continuously updated the system and have added a variety of new features, the structure remains essentially the same as it was in 1995. The data access system is a combination of Perl scripts, C programs, and HTML pages. We designed the system specifically for the original IPUMS files, and it is not easy to add new data sets. Even modifying a single variable is difficult, because it typically requires changes in at least eight different files, including statistical package data-definition files, tables used to build extract pages, and static HTML documentation pages.

Our new infrastructure projects pose a range of technical challenges for data access. For example, we expect that IPUMS-International will eventually incorporate over 100 censuses, and it is not feasible to display the documentation for all data sets simultaneously. Therefore, we must develop a system for dynamically filtering the documentation pages to display only the information needed for a particular analysis. The NAPP and the restricted Census Bureau data for the period since 1960 include much larger files than the existing IPUMS database, so we must develop methods to increase the efficiency of data extraction. The NHGIS project will require an entirely new approach to data access, because aggregate data and boundary files require different strategies for searching for information and delivering results.

The needs require new data access tools. We plan to develop systems that are as generalized and flexible as possible so that we can add new data sets and variables with minimal changes to software. We must also make the software more portable and transparent, since we plan to install it in several remote locations.

To accomplish these goals, we are building our new generation of data access software around a framework of machine-understandable metadata. As described by Block and Thomas (in Part Two of this issue), we have adopted the Data Documentation Initiative (DDI) standard for documentation encoding. The DDI format will allow us to simplify maintenance of both software and documentation. Machine-understandable metadata will allow us to write software to automate the incorporation of new data sets into the system. Moreover, when we modify a variable, we will be able to make the change in a single location and it will propagate throughout the system.

By moving beyond system-specific software built for a single application, we will reduce costs and improve the long-run sustainability of our data access systems. We also hope to improve functionality. Although many of the software changes will be invisible to existing IPUMS-USA users, we plan several new features to improve the flexibility of the system, including version control, improved navigability, user preferences, and customizable constructed variables.

Why the IPUMS Model Works

The key to our success in obtaining large-scale funding for historical data infrastructure is the widespread use of our data across many social science disciplines. Among the academic researchers who registered to use the IPUMS in 2002, 38 percent describe themselves as economists, 28 percent are sociologists or demographers, and 23 percent are scattered across a range of disciplines, including geography and political science. When we began the project, however, we hoped that it would also contribute to a revival of quantitative research by historians, and our success in that goal has been more limited: just 11 percent of the registered users describe themselves as historians. In absolute terms, however, the number of historians using the IPUMS is growing rapidly, so we are optimistic for the future.

Several factors have contributed to the popularity of the IPUMS. We have provided broad unfettered access to data, and we invested early in electronic dissemination tools. From the outset, we focused on prompt release of data with no privileged access for insiders. Our documentation is carefully written and coherently designed. We have devoted considerable effort to dissemination activities such as conference exhibits and workshops, and we provide good user support.

The most important reason for the popularity of the IPUMS, however, is the extraordinary significance and versatility of the data source itself; no other provides such rich information on long-run social and economic change. Moreover, unlike most historical data sources, the IPUMS connects the past to the present. Descriptions of the past provide the basic empirical foundation for theories of social change and projections into the future, so the series of IPUMS samples provides a unique laboratory for the study of economic and demographic processes. Thus, old census data are of interest not only to historians but are also of interest to social scientists, who are becoming increasingly aware that historical data are essential to basic social research and policy analysis.

NOTES

[1] Thousands of others have obtained IPUMS data through other distribution channels. Users can bypass the IPUMS extraction system and directly download entire files from our anonymous FTP site. Moreover, the data are redistributed through other social science and population center data archives and by our three private-sector dissemination partners: Public Data Queries (http://www.pdq.com), Querylogic (http://www.querylogic.com), and Key Curriculum Press (http://www.keypress.com/fathom).

[2] To protect the confidentiality of respondents, the individual-level files for the period since 1940 do not identify places with populations under 100,000. In the earlier period, the existing samples are generally too small to allow close analysis of small areas.

REFERENCES

Ruggles, S., and P. Kelly Hall, eds. 1999. IPUMS: Integrated Public Use Microdata Series (theme issue). *Historical Methods* 32: 102–58.

Ruggles S., and R. R. Menard, eds. 1995. The Minnesota Historical Census Projects (theme issue). *Historical Methods* 28: 6–78.

Ruggles, S., M. Sobek, and T. Gardner. 1996. Disseminating historical census data on the World Wide Web. *IASSIST Quarterly* 20: 4–18. (Accessed at http://www.iassistdata.org on 12/11/2002.)

Sobek, M., and S. Ruggles. 1999. The IPUMS project: An update. *Historical Methods* 32: 102–10.