The IPUMS International Census Microdata Project

Matthew Sobek, Patricia Kelly Hall, and Lara L. Cleveland
University of Minnesota, Minneapolis, Minnesota, USA

Corresponding author: Matthew Sobek, email: sobek@umn.edu

Abstract

The Integrated Public Use Microdata Series (IPUMS) International partnership is a project of the Minnesota Population Center and national statistical agencies, dedicated to collecting and distributing census data from around the world. The project has collected the world's largest archive of publicly available census microdata samples. Currently disseminating more than 200 census samples from 68 countries, the goals of the IPUMS International are to collect and preserve data and documentation, harmonize data, and disseminate data cost-free to researchers. The data series includes information on a broad range of population characteristics, including fertility, nuptiality, life-course transitions, migration, labor-force participation, occupational structure, education, ethnicity, and household composition. This paper describes sample characteristics and data structure; the data harmonization process including the creation of constructed family interrelationship variables; and the flexible dissemination system that enables researchers to build customized extracts of pooled census samples across time and place.

Keywords: census microdata, data harmonization, data dissemination, demographic analysis

1. Introduction

The Integrated Public Use Microdata Series (IPUMS) is the world's largest collection of individual-level census data available to researchers. Each IPUMS record describes a person, thus the term "microdata." The database currently contains data on 480 million persons from 68 countries. The data are coded consistently across censuses, enabling researchers to readily make comparisons between countries and across time periods. A web-based data access system allows users to easily browse this vast database, selecting only those records and variables necessary for their analysis. The requested data are downloaded to the researcher's computer where they can be analyzed. Researchers must apply for access, demonstrating a reasonable scientific need for the data; but once approved, they have access to the entire database. The data access system is available at www.ipums.org/international.

2. Goals

The IPUMS project began in 1999 with funding from the U.S. National Science Foundation and subsequently from the U.S. National Institutes of Health. The primary goals of the project have remained unchanged: preservation, confidentiality, harmonization, and cost-free dissemination to qualified researchers.

The most urgent goal is to preserve the census data. Much data has been lost over the years, and more is at risk of destruction. IPUMS is primarily focused on censuses, which were collected at great public expense but which often fall into neglect once the results are published and a national statistical office turns its attention to the next census. And the data are of little use without the metadata explaining how to interpret them; both must be preserved if future generations are to benefit from these rich data sources. IPUMS archives the data and metadata entrusted to it, ensuring their survival. In many cases,

IPUMS has funded the recovery of old data off of aging magnetic tapes that were otherwise unreadable by their statistical offices.

IPUMS does not just archive data, nor does it simply post datasets for researchers to download—the project harmonizes the data so the same codes mean the same things for all times and countries in the database. Data intended for public dissemination to researchers must also be confidentialized to remove the possibility of identifying individuals. This is critical for public trust and is stipulated by the dissemination agreements with the statistical offices that provide the data. Finally, IPUMS is committed to the goal of democratizing access to these invaluable data. All IPUMS data are available to researchers everywhere completely free of cost. Anyone with an Internet connection and a viable research project can access the data. No one is privileged with special access, and no one conducting legitimate research is turned away. Over 7,000 users representing more than 70 countries have accessed the database so far.

3. National and International Partners

The IPUMS-International partnership draws on the census expertise of more than 100 national statistical offices around the world and the data integration and harmonization experience of the Minnesota Population Center. The IPUMS project has also benefited from the cooperation and support of international organizations who work to facilitate research access to census microdata for qualified policy makers and social science analysts worldwide. A complete list of national and international partners is available at: https://international.ipums.org/international/international_partners.shtml.

4. Data

As noted, IPUMS is composed of microdata. Each record represents an individual and is composed of variables describing all of that person's characteristics as collected by the relevant census. Individuals are organized into households in nearly all IPUMS samples. This data structure provides substantially more power than would a simple sample of individuals. Thus, a researcher has access not only to the person's characteristics, but to all the characteristics of the people with whom they lived, family and non-family. This allows the construction of new variables drawing from information across individual person records, such as the number of wage earners in a family, or whether a mother has children under age five. Each set of persons also has a household record that contains information shared by household members, such as geography, and—in most censuses—the attributes of the dwelling in which they lived, such as presence of piped water or number of rooms.

IPUMS includes 211 census microdata samples from 68 countries (Table 1). In all cases the samples are nationally representative. The modal sample density is 10% of the national population, but 5% is also common, and some samples are lower density. The median sample size is 828,000 records, and the database in total has 480 million persons. Roughly three quarters of the countries and two thirds of the samples are from developing countries. The temporal scope of the database is 1960 to the present. Because most countries have samples from multiple census years, it is usually possible to analyze change over time nationally and internationally. It is not possible to track individual persons across censuses.

The topical coverage of the samples is dictated by the censuses from which they are derived. Where long-form variants of a questionnaire were used, IPUMS samples are typically drawn from those richer records.

Table 1. Number of IPUMS Samples by Country (211 Total)

Argentina	4	Germany	4	Malaysia	4	Saint Lucia	2
Armenia	1	Ghana	1	Mali	2	Senegal	2
Austria	4	Greece	4	Mexico	7	Sierra Leone	1
Belarus	1	Guinea	2	Mongolia	2	Slovenia	1
Bolivia	3	Hungary	4	Morocco	3	South Africa	3
Brazil	5	India	5	Nepal	1	Spain	3
Cambodia	2	Indonesia	9	Netherlands	3	Sudan	1
Canada	4	Iran	1	Nicaragua	3	Switzerland	4
Chile	5	Iraq	1	Pakistan	3	Tanzania	2
China	2	Ireland	8	Palestine	2	Thailand	4
Colombia	4	Israel	3	Panama	5	Turkey	3
Costa Rica	4	Italy	1	Peru	2	Uganda	2
Cuba	1	Jamaica	3	Philippines	3	UK	2
Ecuador	5	Jordan	1	Portugal	3	USA	6
Egypt	2	Kenya	2	Puerto Rico	5	Uruguay	5
El Salvador	2	Kyrgyzstan	1	Romania	3	Venezuela	4
France	7	Malawi	3	Rwanda	2	Vietnam	3

Numerous countries not included in Table 1 have already sent data to the project that are currently awaiting processing. IPUMS processes on average 25 new samples every year while trying to maintain geographic dispersion as the database is expanded.

5. Processing

Census data and documentation primarily come to the IPUMS project directly from the national statistical offices that produced them. The data come in a wide range of formats, from software system files to multitudes of fixed-column files that need to be matched together. The metadata describing the data can be even more heterogeneous, especially for the older censuses. The documentation is usually in the native language of the various countries. The first step is to translate this material into English, the working language of the IPUMS project.

Protecting the confidentiality of census respondents is a primary principle of the project and of great importance to the national statistical offices that supply the data. The greatest protection is the lack of names in the data. An additional powerful protection is that the data are only samples of the population, so there is no guarantee that any particular individual is even in the data. To augment these built-in barriers, small areas are not identified, and variable categories are suppressed if they represent very small numbers of persons in the national population (they are combined with other categories). A small number of cases is also swapped across geographic units to further impede identifying individuals. If the country deems any variable too sensitive for distribution, IPUMS suppresses it.

Many countries supply IPUMS with 100% of their population microdata, entrusting the project with drawing the scientific-use sample. Usually, a 10% sample is created in these cases, and the full-count data are archived. The IPUMS sample design is simple but powerful: a systematic selection of every 10th dwelling from a geographically sorted sample. Most of the countries that draw their own samples for the project follow this basic design. The resulting samples are perfectly representative geographically. They are implicitly stratified geographically, producing better-than-random sampling of all the

personal and housing characteristics that are correlated with geography. The sample design is also easy to explain, which has additional benefits.

One confidentiality measure can have a particularly large impact on the kind of research that can be conducted: the suppression of low-level geographic detail. No locality smaller than 20,000 population is uniquely identified in the database. Smaller places are combined as necessary to meet the threshold. The 20,000 population threshold normally applies to the samples constructed by IPUMS, but in cases where the national statistical office created the samples, they sometimes provide even less geographic detail. Highly restrictive geographic information is most common among developed countries, where perception of disclosure risk may be greater.

All of the data contributed to IPUMS must be reformatted into the household-person structure understood by the software undergirding the project. The data formatting can be involved, but it is generally less onerous than processing the metadata. There are fewer standards governing metadata and there are language and terminology issuesAs mentioned, step one is to translate the documentation into English, as necessary. That injects some degree of interpretation that must be considered during subsequent work. Three key types of documents are required for processing: a data dictionary that describes the variables and their codes; the census questionnaire and instructions; and information on how the census was conducted and how the sample was drawn, if not by IPUMS.

The data and metadata are processed in tandem. Each variable must be tested for soundness and documented, including writing custom text that describes the variable and any issues that are not self-evident from the labels. One of the most innovative aspects of IPUMS lies in the handling of the census questionnaire text and instructions. Machine-readable tags are inserted into the text during processing. These tags mark the beginning and ending points for each block of text that applies to a specific variable in the sample data. In essence, they link the text to the variables in the data. The tags are not visible to users, but the IPUMS software uses them to pull this information together so users can easily view it online when examining a variable.

After all the original sample-specific variables are fully processed, IPUMS harmonizes the data across countries and over time. The project has made several hundred of these harmonized variables, where the same codes apply to all samples in the database. Thus, there is one marital status variable for all of IPUMS. Harmonization is made feasible by using a composite coding system. The first digit or two are usually fully comparable across all samples, while trailing digits retain details that are present in only some samples. Marital status, for example has four categories at the first digit—single, married/in-union, divorced/separated, and widowed—while the last two digits include variations for religious marriage, polygamy, etc. Finally, documentation is written to point out comparability issues that remain after harmonization.

Although the composite coding is intended to lose little or no information by retaining sample-specific details in trailing digits, this is not always practical. To keep variables from getting unwieldy or to make the concept from one census fit another, it is sometimes necessary to drop some detail in the harmonized variable. But another core principle of IPUMS is to lose no information. All of the sample-specific variables that are the source variables for harmonization are also available to researchers for downloading. They are not harmonized with other samples, but they accurately reflect the full information in the original data. This allows researchers to make their own informed judgments about how

to make comparable measures between samples, or to study one time and place efficiently.

6. Dissemination

The IPUMS data are accessed via a custom-designed web dissemination system. This is the only way the data are distributed—no one has early access or can obtain data not available to every other researcher. The web system allows users to browse for variables while filtering for the countries and time periods of interest to them. As they select variables, they are added to the data cart. The system defaults to showing the internationally harmonized variables, but the unharmonized sample-specific variables are available for browsing with a click. When the user is satisfied with their selections they instruct the system to create their data extract. The data are created off-line and the user receives an email when the data are ready to be downloaded. Researchers download and analyze the data on their desktop using the software package with which they are most familiar. The IPUMS system supports SAS, SPSS and STATA, and there are plans to add R in the near future.

As the variables are browsed, users have access to all the metadata via a series of tabs. They can see the universe of persons who were asked the question in each census. The codes page shows the availability of categories across samples, and it displays unweighted frequencies so the researcher can determine at a glance if the relevant samples have sufficient cases for a particular analysis. The comparability section explains all of the issues that the IPUMS team deemed necessary to highlight for researchers. But users do not have to depend on this judgment. The enumeration text tab shows all of the questionnaire text related to the variable for each census, including the specific wording of the question and the categories that were printed on the census form. Images of the original-language form are also available. The enumeration text feature lets researchers quickly get to the most fundamental information about how the data should be interpreted. If a user feels that harmonization may have lost useful information, there is also a tab that lists the input source variables for each harmonized variable. The source variables can be selected there to include in a data extra with or without the associated harmonized variable.

IPUMS is designed to facilitate cross-national research. The data extract system lets users define pooled datasets that include any variables they desire from as many times and places as they wish. Thus country and year can be variables in the analysis. Using the extract system it is feasible to build a single dataset containing selected variables for all 480 million persons in the database. If such a dataset would be too large, the system is capable of drawing a systematic subsample of cases. Of course, most analyses are more localized in time and place, but IPUMS offers the unique potential for truly globe-spanning research. This is a practical possibility not only because of the data extract system, but also due to the harmonization of the variable codes and to the documentation system that collates information at the variable level across samples. Thus, the primary logistical barriers to cross-national studies are removed, and researchers can focus on the substantive matters of interest to them.

7. Added Value

The IPUMS project provides value to the worldwide research community by preserving, documenting and disseminating census data. Harmonization of the data is also a significant contribution, as is the unique ability to pool data across samples into custom-designed extracts. But IPUMS adds value in other ways that go beyond what was specifically in the data it received.

As noted above, individuals are organized into households in the IPUMS. This hierarchical structure gives the data much of its power, making it possible to interrelate the characteristics of co-resident persons in creative ways. To fully exploit this feature of the data, IPUMS constructs "pointer" variables that identify the location within the household of each person's mother, father, and spouse, if they were present. This makes it simple to compare the characteristics of spouses, to attach parents' characteristics to children or vice versa, and to construct unique household or family-level measures. For example, one can make a variable for spouse's education, mother's birthplace, or father's migration status. On the last screen of the data extraction process, users are given the opportunity to construct such variables and append them to the person records.

IPUMS has prepared GIS boundary files that enable users to map the data for all countries at the first administrative level (states, provinces, etc). The first level has also been harmonized across time within countries, so users can be certain that the same geographic codes describe the same space in all periods. The project is currently engaged in a longer-term goal of providing boundaries and harmonized geography for the second level within as many countries as possible. The geographic work will be leveraged by a major related project at the Minnesota Population Center: Terra Populus. The new project will allow users to add environmental and land cover data from satellites to the person records in IPUMS. Thus, a researcher can make a variable for the percent of forest cover or annual rainfall in a geographic unit, and have that appear as a variable on a person's record. Of course, users will have to be approved to use IPUMS to be able to access the enriched microdata through Terra Populus. The new data system should become available to researchers in late 2013.

In an attempt to facilitate regional research, IPUMS has developed mirror sites with partners at the Autonomous University of Barcelona, the African Centre for Statistics, and the Economic Research Forum in Cairo. These sites provide access to IPUMS data to approved users while allowing the partner organizations to add content of specific value to their research communities.

The biggest initiative on the horizon for IPUMS is the possibility of developing a restricted access system that would give researchers access to high-density and even full-count data with full geographic detail. Access sites would be strictly controlled, and the data would be analyzed only remotely on the IPUMS servers. No data would be transmitted—only the results—and those would be subject to review by IPUMS staff before their release to ensure confidentiality protection. The restricted access system would allow new kinds of analyses that use small places to study such things as human-environment interactions, health outcomes related to location, segregation, or access to services. Such a system is only speculative at this point, but the IPUMS team is hopeful it can be accomplished in the coming years. The system will, however, be dependent on the willingness of the national statistical offices to allow this kind of research.

The IPUMS project is possible only because of the remarkable contributions of data by its NSO partners. The project must always adhere to the letter and spirit of its agreements with the NSOs and, indeed, must continue to find ways to return value to the statistical offices in whatever ways it can. One recent step in this direction was the addition of an online tabulator to IPUMS in 2013. The tabulator allows users to analyze the sample data without a statistical package. The system is most practical for basic two- and three-way tabulations, but it can perform more sophisticated analyses as well. The system can be handy for any user, but may be particularly helpful in developing countries where bandwidth is small, because the data never need to be downloaded. The system is also very fast, running through most samples in a few seconds.