IPUMS INTERNATIONAL: A DATA RESOURCE FOR STATISTICS EDUCATION

Patricia Kelly Hall, Lara Cleveland and Matthew Sobek
Minnesota Population Center, University of Minnesota, USA
pkelly@umn.edu

IPUMS-International is the world's largest collection of high-precision census data samples containing individual-level information on 544 million people in 74 countries spanning five decades. These data are available for download at no cost to educators, students and researchers for scholarly, educational, and policy-related analysis. The database, built in cooperation with national statistical offices, provides remarkable access to data for educators wishing to expose students to real-world governmental statistics. Variables with distinct census responses for each person are coded consistently across time and place; documentation is thorough, harmonized and easily accessible; and the web delivery system allows registered users to create and download customized data sets pooled across time and place. Individual level responses mean that data can be used in analyses that range from simple descriptive tables to advanced statistical modeling.

IPUMS-INTERNATIONAL: HARMONIZED MICRODATA AND METADATA

The Integrated Public Use Microdata Series-International (IPUMS International) is a data infrastructure project which disseminates high-precision census microdata samples to teachers and researchers world-wide, free of cost, and on-line at https://international.ipums.org. In partnership with more than 100 national statistical agencies, as well as data archives, research centers, and international organizations, we have assembled the most comprehensive collection of census microdata in the world (Figure 1). The microdata records describe a billion persons nested within families and households, with information about the inter-relationships of all members of each residential group. For every person, the data include detailed information about geographic location, demographic characteristics, and economic activities. The data also cover education and literacy, migration and place of former residence, marital status and consensual unions, disabilities, characteristics of the building, and a host of other characteristics. Researchers can analyze multiple census years and multiple countries as a single pooled dataset.

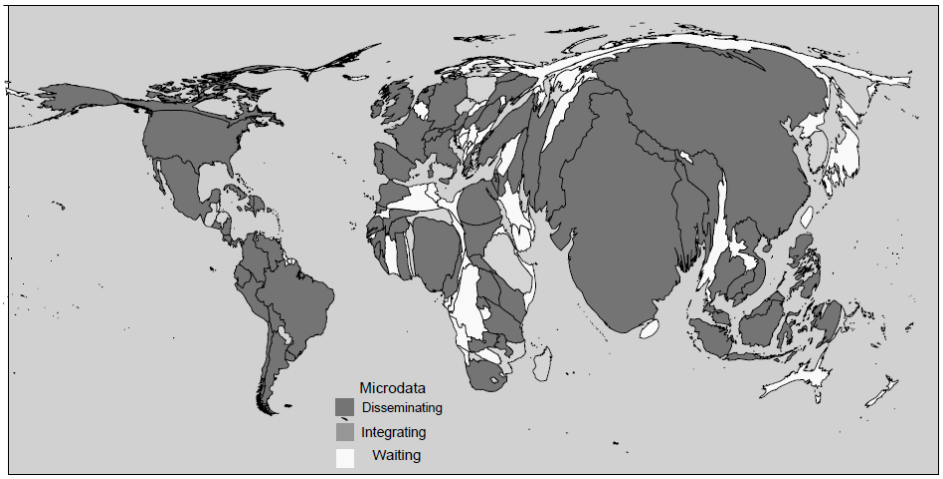

Figure 1. Census data distributed by IPUMS-International, with Country Area Weighted by that Country's Proportion of the World's Population in 2014. (*Data Sources:* Original data contributed by national statistical offices are harmonized and distributed by the IPUMS-International census project.)

Census microdata collected by countries around the globe contain a wealth of information useful to teachers and researchers (McCaa & Ruggles, 2002). Although large census microdata samples exist for many countries, access has been limited and the documentation has often been inadequate. Where microdata are available, comparisons across countries or time are difficult because of inconsistencies in data and documentation. IPUMS-International addresses these issues

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

by converting census data for multiple countries into a consistent format, supplying comprehensive documentation, and making the data available through a web-based data dissemination system.

Harmonized Census Microdata.

The harmonization process begins with verifying the universe of individuals whose responses are reported in each variable. Frequently, those not asked the question—such as males on a female-only fertility question—are nevertheless reported with a value which has substantive meaning. In the fertility example men, who are not in the universe, are sometimes given the value of 0 which is also used for women with zero births. Once the universe is verified and cleaned, general categories such as "no response" and "not available" are given standard IPUMS codes to make the data more easily and more quickly understood by users.

Reconciliation of variable codes across census samples is the essential task of data harmonization. We identify and recode every variable to the lowest common denominator of detail available—and comparable—for all or most samples. We also retain all detail in the original dataset, even if it is unique to that country and that year, through the use of a composite coding scheme: the most common level of detail is found in the first digit of a multiple-digit code, the second most common detail is captured in the second digit, and so on. Users can select general (first digit) or detailed codes. As part of our commitment to preserving all original detail, we have included the *source variable* with the original, unharmonized codes in the data extract system (Esteve & Sobek, 2003). For a more complete overview of IPUMS-International harmonization principles, see https://international.ipums.org/international/harmonization.shtml

Harmonized, Interactive Metadata Browser.

Data are useful only when researchers understand what they mean. Accordingly, we provide harmonized English-language documentation on each sample. This documentation covers enumeration procedures and instructions; definitions of households, dwellings, group quarters, and other enumeration units; and scanned images of original-language questionnaires. We provide detailed descriptions of the sources for each variable, including question wording and instructions (in the original and translated into English), universe definitions, frequency distributions, and variable codes. Comparability discussions describe any deviations of particular censuses from the standard variable definition and address differences over time and across countries.

The metadata browser limits the information displayed to only those elements relevant to a given research project, as defined by the user. This is possible because the metadata are encoded with tags for flexible display options, and the documentation pages are constructed dynamically. Suppose, for example, a user selects censuses only for Kenya. The question text, enumeration instructions, comparability discussions, and frequency tables would then cover only the Kenyan datasets. This filtering capacity creates individually-tailored documentation that highlights subtle problems of comparability without overwhelming users with unnecessary information.

Customized Data Extract System.

The IPUMS data access system pioneered web-based dissemination of large-scale microdata collections with software that allowed users to merge datasets, select variables, and define population subsets. Current dissemination software has enhanced data security and powerful new tools allow users to construct customized variables and draw subsamples tailored to their specific needs. The IPUMS disseminates pooled extracts in a single dataset, custom-tailored to the precise research needs of the user. Each user-generated extract contains only the requested microdata accompanied by the corresponding set of DDI (Document Data Initiative) compatible metadata and a codebook suitable for constructing a system file in SPSS, SAS or STATA.

CLASSROOM ACCOUNTS AND TOOLS FOR INSTRUCTION

IPUMS-International users can now set up classroom accounts. The accounts enable instructors to share extracts directly with students through the IPUMS web site. Student registrations are also easier, and instructors can see who is registered and therefore legally entitled to receive IPUMS data. For more information, please see "Teaching with IPUMS" at https://international.ipums.org/international/teaching.shtml

Additional features of IPUMS-International data add significant value to the raw data for classroom use: extensive documentation and guidance on variability in sample design, a data tabulator, household relationship pointer, GIS data, and a virtual data carrel for each user.

- *Private Virtual Data Carrel.* Each registered user of IPUMS-International has a private, password protected extract history page. Contained here are statistical package syntax files and data for download of recent extract requests, as well as the extract syntax file and description (if user provided) for each data order ever requested by that user. With a simple click an old extract can be re-requested or opened for modification and submission. This is particularly useful in the classroom where a complex classroom exercise or exam can be re-used for a new class by modifying the data request with a different country or year.
- *Customized Extract Size.* The IPUMS data extract system has a feature which allows users to customize the size of an extract by number or percentage of available cases. The system provides the syntax to adjust the weights in proportion to change in the extract size. Smaller extracts are especially useful in testing syntax or for use on lower-speed classroom equipment.
- *Sample Documentation and Guidance.* Census samples in IPUMS-International employ a variety of sample designs. All IPUMS samples contain individual level data, most are clustered by household, many are stratified, and some are differentially weighted. The IPUMS-International samples are either systematically drawn from full count data by IPUMS (or according to IPUMS specifications) or they are drawn by the statistical offices of the country of origin. Where possible, IPUMS-International provides systematic 10 percent samples of census data. Samples drawn by countries of origin may employ a variety of complex sampling techniques that include oversampling, clustering and stratification which are not always documented. Detailed sample descriptions are available on the project web site at https://international.ipums.org/international/samples.shtml; guidance on variance estimation is described in Cleveland, Davern and Ruggles (2011). For a discussion of potential statistical hazards with data perturbation techniques in census samples, see Cleveland, et al. (2012).
- *On-line Tabulator.* Quick tabulations can now be made with the IPUMS International Online Data Analysis System. The IPUMS online analysis system uses high-speed tabulation software developed at UC-Berkeley's Computer-assisted Survey Methods Program. Researchers registered with IPUMS International may specify samples and variables of interest and get quick calculations output to their computer screen or mobile device.
- *Pointer Variables.* IPUMS data include powerful constructed variables that aid users in utilizing information about household structure implicit in the census data samples. These variables, referred to as "pointers," include a consistent, versatile, and reliable set of constructed variables that describe a variety of family interrelationships among individuals within the same household. Researchers use them to link characteristics of one family member to another—spouses to spouses, children to either or both parents, and so on— thereby speeding up analyses of family structures and characteristics (Sobek & Kennedy, 2009).
- *GIS Boundary Files.* As geospatial measurement techniques have advanced, so has user demand for additional geographic information and tools for utilizing spatial data. We have recently enhanced documentation and harmonization of spatial information contained in the household geography of the census records. IPUMS-International has added spatiotemporally harmonized geographic variables and accompanying boundary files (shapefiles) to facilitate national and international data mapping. Users can create maps with IPUMS-International data using a statistical software program and ArcGIS (a GIS mapping software).
- *User Support.* If you or your students have a question when using IPUMS data, simply email ipums@umn.edu. The staff there will provide answers to common questions promptly. More complicated questions will be referred to one of the IPUMS-International senior staff.
- *Additional Datasets.* The Minnesota Population Center (MPC) has other datasets available for use in the classroom (Hall, 2011; Ruggles, 2014; see https://www.ipums.org). All are built on the same harmonization principles and use a common dissemination system. Finzer et al. (2007) provide a review of the MPC's collaborative efforts to expand the use of United States data in the classroom.

IN CONCLUSION—AN INVITATION TO EDUCATORS

The IPUMS-International partnership is eager to see this valuable, free data international resource put to greater use in classrooms around the world. Although IPUMS-International was originally conceived as infrastructure for social science research, analysis of recent user data revealed that more than 20 percent of the 9,230 extract requests were for class assignments—ranking above Ph.D. and other theses. The number of instructors registering to use IPUMS data for teaching has also risen dramatically. So we issue this invitation to all statistics educators:

- Visit the website, register, and explore the classroom possibilities for yourselves.
- Let us know what we might do to make IPUMS even easier for you and your students to use. Virtually all tools and enhancements to the first IPUMS were made at the suggestion of our research data users. The classroom accounts, themselves, were introduced at the urging of a data user. We are eager for your input on what tools and features we might add.
- Share data exercises using IPUMS that others might bring to their classrooms. A few of ours are on the Minnesota Population Center website at https://www.pop.umn.edu/data-user-resources/data-support. Send yours to ipums@umn.edu. If we use them, we will credit you and your institution with their development.

Finally, let us know if you are interested in helping train others in the use of these data in the classroom. Together with our national statistical office partners, we are developing training materials and classroom exercises to facilitate the use of national census data in the classroom.

ACKNOWLEDGEMENTS

The IPUMS-International data infrastructure project at the Minnesota Population Center, University of Minnesota, is supported by grants from the National Institutes of Health and the National Science Foundation: 5 R01 HD043392, 5 R01 HD047283, and SES-085141.

REFERENCES

Cleveland, L., Davern, M., & Ruggles, S. (2011). Drawing Statistical Inferences from International Census Data. *IPUMS International Working Paper*. Available on the web at https://international.ipums.org/international/variance_estimation.shtml

Cleveland, L. L.; McCaa, R., Ruggles, S., & Sobek, M. (2012). When Excessive Perturbation Goes Wrong and Why IPUMS-International Relies Instead on Sampling, Suppression, Swapping, and Other Minimally Harmful Methods to Protect Privacy of Census Microdata. in J. Domingo-Ferrer and I. Tinnirello (Eds.), *Privacy in Statistical Data 2012*, 7556 LNCS: 179-187.

Esteve, A., & Sobek, M. (2003). Challenges and Methods of International Census Harmonization. *Historical Methods 36*: 66-79.

Finzer, W., Erickson, T., Swenson, K., & Itwin, M. (2007). On Getting More and Better Data into the Classroom. *Technology Innovations in Statistics Education 1,* 1-12. Available on the web at https://escholarship.org/uc/item/09w7699f

Hall, P.K. (Ed.) (2011). Big Data: Large-Scale Historical Infrastructure from the Minnesota Population Center. *Historical Methods Special Issue, Part 1 (Vol. 44, No. 1),* and *Part 2 (Vol. 44, No. 2).*

McCaa, R. & Ruggles, S. (2002). The Census in Global Perspective and the Coming Microdata Revolution. *Scandinavian Population Studies, 13,* 7-30.

Meier, A., McCaa, R., & Lam, D. (2011). Creating Statistically Literate Global Citizens: The Use of IPUMS-International Integrated Census Microdata in Teaching. *Statistical Journal of the IAOS 27,* 145-156.

Ruggles, S. (2014). Big Microdata for Population Research. *Demography, 51,* 287-297.

Sobek, M., & Kennedy, S. (2009). The Development of Family Interrelationship Variables for International Census Data. *Minnesota Population Center Working Paper Series 2009-02.* (http://www.pop.umn.edu/sites/www.pop.umn.edu/files/Working%20Paper%202009-02.pdf)