Data Quality Assessment in the IPUMS: Processing, Formal Review, and the Research Community

Lara Cleveland*
IPUMS at the University of Minnesota, Minneapolis, USA – clevelan@umn.edu

Matthew Sobek
IPUMS at the University of Minnesota, Minneapolis, USA – sobek@umn.edu

Abstract

The Integrated Public Use Microdata Series International (IPUMS International) is a global project to inventory, preserve, harmonize, and disseminate census microdata. The project facilitates comparative international research by providing integrated sample microdata from 301 censes in 85 countries for research use. Countries provide data to IPUMS accompanied by varying degrees of documentation, which can vary depending upon the age of the data, expertise of the staff, and financial resources of the statistical office at the time of the census. IPUMS transforms data to a standard format, creates a harmonized set of variables in which variables are coded consistently across all samples, and provides extensive documentation. IPUMS has a three-fold interest in assessing data quality: 1) understand the incoming data provided by census agency partners; 2) ensure that IPUMS data harmonization efforts do not introduce undue error in the data; and 3) provide high-quality documentation describing the distributed data. In this paper, we describe two sets of data quality checking activities undertaken by the IPUMS team. The first are a set of built-in data investigation and quality assurance checks associated with various steps in the IPUMS data processing system. While designed primarily to catch errors made by IPUMS before releasing processed samples to the public, many of the tests also serve as indirect evaluations of the underlying quality of incoming data. The second are more formal research evaluations of intra-cohort coherence across samples for each country in the IPUMS International database. The results of most assessments are encouraging. We find structurally sound datasets with high degrees of coherence across samples. Researchers should take great care when interpreting the results of quality evaluations. Clean data may be heavily edited data. Well-documented but messy data may be treatable and extremely useful. In our experience, quality assessments are excellent tools for targeting areas that require further investigation. Documenting reasons for acceptable data anomalies can help ensure responsible data analysis.

Keywords: census data; microdata; IPUMS; coherence.

1. Introduction

The Integrated Public Use Microdata Series International (IPUMS International) is a global project to inventory, preserve, harmonize, and disseminate census microdata. It is a collaboration of the Minnesota Population Center, National Statistical Offices, international data archives and experts from participating countries. The database has grown dramatically over the 15 years of its existence and currently contains more than 600 million person records across 301 census samples from 85 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 facilitates access to this vast database. Users select only those records and variables necessary for their analysis and download the data file to their computers for analysis. Researchers must apply for access, demonstrating a reasonable scientific need for the data; but once approved, they have access to the entire database free of charge. The data access system, available at www.ipums.org/international, has been used by thousands of scholarly and policy researchers worldwide.

Sample data in IPUMS International are provided by the national statistical offices or census agencies (NSOs) with the intention that IPUMS will integrate and harmonize their data for scholarly and policy use. The availability of metadata, the amount of detail in codebooks, and overall documentation of the

incoming data varies widely depending upon the age of the data, the expertise of the staff, or the financial resources of the statistical office at the time of the census. Statistical offices differ widely in their approaches to and procedures for data editing, allocation, and confidentiality. Documentation of such practices for public use is scarce. IPUMS International assists researchers by requesting and providing as much information as possible about data collection techniques, post-enumeration processing, and sample characteristics.

For researchers and NSOs, questions of quality of the samples disseminated by IPUMS are of great concern. Baffour and Valente (2012) define census quality as 'fitness for use', characterized by six elements: relevance, accuracy, timeliness, accessibility, interpretability and coherence (p. 122). When users of the IPUMS database ask about the quality of the data, their questions are many. How good were the field operations? What is the coverage? How well were the records handled, coded, and processed by the national statistical office? Is the sample structure reasonable and representative? Did IPUMS define the data correctly? Did IPUMS conceptualize the data correctly in recoding to create harmonized variables? Did IPUMS program data transformations accurately? In other words, users want to know the ways in which IPUMS might introduce errors in the data as well as whether the census office did a good job in collecting, processing and preparing the data.

Investigating data quality is another way to ensure that documentation is thorough and accurate. IPUMS has a three-fold interest in assessing data quality: 1) understand the incoming data provided by census agency partners; 2) ensure that IPUMS data harmonization efforts do not introduce undue error in the data; and 3) provide high-quality documentation describing the distributed data. Day-to-day work at IPUMS includes regular and repeated investigations of data quality. These checks address some questions about how well the NSO handled data and how well IPUMS processed it. In addition to regular internal quality assurance activity, our research staff have also undertaken more formal investigations of data quality. Both types of assessments, internal data processing and the formal evaluations, are discussed in this paper.

2. Data Processing Quality Assessments

Many IPUMS data quality assessments are built-in components of day-to-day processing activity. Building the integrated, harmonized, public-use IPUMS files from NSO data involves a series of transformative steps. Quality checks are essential parts of each step: reformatting; data dictionary review and documentation; universe verification; data integration and harmonization; documentation of sample-level characteristics, and web output review.

Reformatting. Data and documentation provided by NSOs typically come in a range of formats, from software system binary formatted files to fixed-width ascii files. The data can be in one large rectangular file with full information for each person on each record line, or data can be stored in separate files for each record type (dwelling, household, and person) or geographic area (e.g., province or region). Metadata describing the data can be even more heterogeneous than the data structures, especially for older censuses. When historic censuses can be located, they often retain only the format required to produce final output for the published census report. They often lack metadata for particular components. IPUMS research staff review NSO codebooks and any available metadata to define record structures and variable locations.

Before processing, various files and file types must be merged and reshaped to a standard IPUMS format. We never assume anything about the data structure, nor do we assume that metadata are 100 percent accurate, until we have conducted a number of verification checks. We have learned never to assume that identifiers meant to link records across files or file types are clean. During processing, we reformat each census data file into a standard fixed-width household-person format and create an associated IPUMS-style data dictionary. Both data and dictionary are suited for ingest in the custom-built data conversion software program. Each sample presents a distinct challenge.

Data review during reformatting includes a series of automated and manual evaluation steps to ensure that the structural integrity of the data is preserved. We check that the number of persons and households from the input data match the restructured version. Our reformatting tools automatically generate information about household characteristics which we append to the records as additional variables for internal use. These variables include household size; presence of multiple heads of households, multiple spouses, or duplicate records; and record position of the head within the household. Staff review and investigate any suspicious or outlying results. We frequently are able to fill gaps or correct structural issues based on these few straightforward checks; this is especially useful when we lack detailed metadata about household or dwelling identifiers.

Data documentation. Once the standardized data files are ready, automated software generates variable frequencies within the data dictionary aiding staff in reviewing frequency distributions. Staff members add, check, and correct labels for every variable based on available documentation and specify machine-readable associations between variables in the data dictionary and associated text in the census enumeration materials (questionnaires and instructions). Staff research and document correspondences between geographic units identified in the census data and contemporaneous maps, verifying population totals against published results. Finally, relying heavily on available documentation, but also on their social science training and data experience, research staff consider whether frequency distributions seem reasonable.

Universe verification. The most time consuming checks on reformatted input data involve empirically verifying and documenting the universe of each variable. Staff assemble information about expected universe parameters based on the census questionnaires and instructions, create a universe-defining variable, and cross-tabulate the universe in a statistical package. Staff document discrepancies between expected universe specifications and those found in the data, and implement minimal programming to separate “not in universe” from missing values. The objective of the universe checking procedure is to create verified documentation about the variable universe to share with users on the website. The equally important indirect benefit is that staff must thoroughly review the data. Variable descriptions, value labels, and overall documentation about the sample are improved during universe checking. The work described thus far is done for each sample individually. Only after many hours of reformatting, documenting, and verifying are the input samples ready to be integrated and harmonized.

Integrating data and harmonizing variables. Integration is the sample-level activity of standardizing data formats and documenting sample designs; harmonization is the variable-level process of making consistently coded variables. Samples are brought together into a single database through an open source metadata-driven software system developed at IPUMS. From cleaned, standardized, and formatted data, research staff members create a new set of harmonized variables, each of which is coded consistently across all samples in the database for which appropriate inputs are available. Harmonization is done on a variable by variable basis. The harmonization process provides yet another opportunity to identify inconsistencies in the data, this time by explicitly comparing distributions across census years within a country and across samples from other countries. Wherever possible, international standard classification systems drive the conceptual and interpretive decisions about how to categorize data for the integrated variables.

Comparisons of frequency distributions provide important information about the extent to which the classification systems can be applied effectively across time and place. These comparisons are made indirectly during harmonization work, then more formally upon reviewing the output. Data processing infrastructure tools produce tables and visualizations across samples for each harmonized variable. Staff members check the tables for coherent distributions across samples, particularly for samples across time within each country. In reviewing the data, researchers apply what they know about expected distributions, and about how changes in population distributions ought to change gradually over time, to identify anomalies or inconsistencies. For example, sex ratios are expected to be stable and nearly at parity, *unless* we know that gender preference and sex selection practices are prevalent in

a country. On the other hand, we expect characteristics indicative of economic development to trend upward in countries undergoing such transformations. For example, in most cases, we expect to see increasing rates in the availability of electricity, access to water, or modernization of sewage systems from one decennial census samples to the next. We might also expect to see associated social trends, such as decreases in family size over the same period. The review is extremely useful, but we have not developed statistical tolerance tests for acceptable levels of change across samples. Staff check data again when they are loaded to the website, which displays variable frequencies, and yet again when they download and review data extracts before releasing data formally.

Sample documentation. In the early stages of data reformatting, we collect as much formal documentation about census enumeration, census office processing, and sample structure as possible. We use relevant information to ensure that we have appropriately identified dwellings and households. At the end of data processing, integration, and harmonization, we revisit materials to write sample level metadata and construct or integrate household and person weights. Data are checked systematically to ensure that weighted totals for a subset of characteristics are in line with published totals. We then investigate and document any important discrepancies.

As the discussion above demonstrates, data are reviewed on multiple fronts during IPUMS processing. Although the number of times staff review frequencies may seem redundant and inefficient, checks at each step are essential for troubleshooting errors and preventing IPUMS from introducing new errors into the data. Insights into the structure of the data and logical inferences where metadata are scant also help improve overall documentation of the data.

3. Formal quality measures and coherence evaluation

Data quality review during normal processing comprises the bulk of IPUMS data quality efforts. However, in recent years, we have conducted more formal research and evaluation of data quality following recommendations from UN agencies. In particular, we have undertaken a number of intra-cohort coherence assessments. We also reap the benefits of quality reviews undertaken by users of the IPUMS database.

Standard tests. Several years ago, IPUMS assembled several data quality indicators for internal quality monitoring purposes. We constructed three standard measures of age-heaping, compiled summaries of missing data (variable and record level missing), compiled summaries of universe inconsistency rates, and reviewed processing notes about issues encountered during the reformatting process. The question we faced was how to interpret the results of the exercise for purposes of informing data users. Some input data files have been heavily edited before leaving the statistical office while other files remain virtually unedited. Age heaping, for example, is more a function of numeracy in the population than an indicator of the skills of the census analyst. The statistical office can certainly improve the quality of data about age by refining enumeration methods to elicit more precise age responses, but they cannot easily improve the numeracy of their population. Some missing values and universe errors are expected in data collection since humans are imperfect data collectors and imperfect respondents. Clean-looking data is actually more likely to have been edited than data with age heaping, missing values, and universe inconsistencies. In IPUMS, a few data sets that performed well in the tests are known to have come from censuses with field operation difficulties or from censuses processed in questionable political environments. Data editing experts treated the data to make it fit for public use. Data from a few censuses known to have very good field operations and from offices with highly skilled processing operations sometimes has a bit of messiness. Which data files are, then, of better quality? We have yet to determine whether or how to release a systematized report based on these measures of quality because doing so might actually lead to erroneous impressions about the quality of the individual samples.

Intra-cohort coherence: age-sex ratios. A better approach to assessing data quality of the type that interests IPUMS data users involves comparing data across data sources. Integration output review, described above, in which staff look for distributional consistency or reasonable distributional trends

in the data, is one such assessment. As a formal extension of the output review work, we undertook a more precise evaluation of cohort coherence across census samples. For characteristics that are remain relatively consistent from one census to the next (e.g., a person's sex or educational attainment after a certain age), we can expect rates among each birth cohort within the country (or geographic area) to remain consistent. Such expectations assume that differences attributable to migration or mortality are relatively small.

In the first of these assessments, we followed a four-step process recommended in the UNFPA guidelines (Moultrie 2012) for evaluating age-sex ratios:

1) Graph age by single year to assess age heaping;
2) Graph smoothed ages and age ratios by sex, noting differences between males and females;
3) Calculate sex ratios, graph by age, and flag distributions that are not sex-balanced; and
4) Produce and compare distributions across multiple years for the same birth cohort.

The assessment strategy relies on assumptions of evenness between sexes and smooth, even patterns across age. Smooth transitions across birth year are considered acceptable, but peaks or valleys in rates or ratios by age or differences in age ratios by sex are suspect. Samples in the IPUMS performed quite well. Peaks and valleys were found in several samples, but they were usually associated with real shock to the population corresponding to periods of conflict or famine. That is, assumptions about minimal effects of migration or mortality were violated in many of the "suspect" cases. Although generally reassuring of overall quality, the assessments are useful tools for locating data anomalies and in targeting areas for further investigation.

Intra-cohort coherence: educational attainment. In another series of tests, we compared cohort educational attainment rates across census samples within countries. Baffour and Valente (2012:126) identify two types of coherence: internal (results for a single census are coherent within themselves) and external (comparisons between two or more censuses or other official sources). To achieve statistical coherence, definitions, concepts, frameworks and classifications must be clear and consistent both nationally and internationally. When these change, explanations are essential to describe similarities and differences between the old and the new. The Sixteenth Meeting of the United Nations Economic Commission for Europe Group of Experts on Population and Housing Censuses defines statistical coherence as follows (see UNECE 2014, p. 4, Section B.4.f):

> Coherence reflects the degree to which census information can be successfully brought together with other statistical information within a broad analytical framework and over time. The use of standard concepts, definitions, and classifications—possibly agreed at the international level—promotes coherence.

By harmonizing variables around like concepts, variables in the IPUMS are virtually ready-for-use in coherence assessments.

For the 2010-round of censuses, the United Nations Statistics Division recommended educational attainment as a core topic and, in post-enumeration processing, recommended the use of categories of the 1997 revision of the International Standard Classification of Education (ISCED) to facilitate international comparisons (UNSD 2008:149-150). ISCED1 constitutes primary education, typically 4-7 years completed with six years the most common (UNESCO 2012:17). For the education assessment, we ask a simple question: For each birth year, is the proportion reported completing primary school in the most recently available sample similar to that for the one from 10 years earlier?

There are at least three considerations in assessing coherence across census samples as proposed here: census agency practices; IPUMS harmonization; and bias. First, the questions, definitions and categories posed in the series of censuses and the training of the field enumerators must be considered, as well as how the data were processed and edited by the census authority. Second, the way IPUMS harmonized the microdata and potentially introduced error, are also relevant. Third, the method assumes that differences in mortality, migration or reporting by level of educational attainment are minimal. The method also assumes that no adult education campaigns took place between censuses on

a scale that would contribute to increases in the percentage graduating primary school after the normal age. At high ages, educational level may be associated with differential mortality rates. To the extent that migration rates or response rates correlate with educational attainment, additional adjustments in method may be required. Despite the potential complications, coherence across samples is remarkably high across countries in IPUMS.

The vast majority of samples in IPUMS held up remarkably well, showing high degrees of statistical coherence. A handful of countries require further investigation. Migration and mortality incidence must be investigated for samples where patterns seem to deviate. Interestingly, we learned that age heaping has little effect on evaluations of coherence where heaping intervals coincide with census intervals. However, in one country, when two censuses were conducted 11 years apart, rather than 10, the imbalance in cohort rates of educational completion differed significantly due to digit preference in reporting age for that census. Moving the trendline by one year to match the census interval yielded much more consistent results. While the example provides a reminder to consider the effects of age heaping in analyses that are heavily dependent upon the precision of single-year age reporting, it is not necessarily indicative of problematic census operations, NSO processing problems, or IPUMS data coding errors.

4. Discussion and Conclusion

Given variations in data editing practices among national statistical offices and the limited documentation of such practices, investigations of coherence constitute research about structural data characteristics rather than assessments of producer error or poor data quality. Cases in which data appear to match expectations perfectly may indicate high “quality” in terms of response rates and accuracy of data collection and data entry activity. Alternately, clean-looking data may simply reflect high degrees of editing of otherwise messy data. We should be careful about interpreting quality assessment results and be extremely cautious about applying low-quality stamps to data files. Assessment techniques are very useful for revealing data structure and provide important information for flagging data anomalies requiring further investigation. Using the knowledge gained from quality explorations to write good metadata improves the overall quality of the data and contributes to better data analysis.

References

Baffour, B. and P. Valente. 2012. An evaluation of census quality. *Statistical Journal of the IAOS* 28:121-135. DOI 10.3233/SJI-2012-0752.

Minnesota Population Center. 2014. *Integrated Public Use Microdata Series, International: Version 6.3 [Machine-readable database]*. Minneapolis: University of Minnesota.

Moultrie, T. 2013. "General assessment of age and sex data." Chapter 1 in *Tools for Demographic Estimation.* Moultrie, T. et al. Paris: International Union for the Scientific Study of Population.

United Nations Department of Economic and Social Affairs, Statistics Division (UNSD). 2008. *Principles and Recommendations for Population and Housing Censuses, Revision 2*. Statistical papers Series M. No. 67/Revision 2, New York.

UNESCO Institute for Statistics. 2012. *International Standard Classification for Education ISCED 2011*. Montreal.