Cross-Census Assessment of Age-Sex Ratios:

An application of newly updated UN assessment guildelines to microdata census samples from the IPUMS International

Lara Cleveland
University of Minnesota, Minneapolis, Minnesota, USA
clevelan@umn.edu

Data about social phenomena are essential to researching patterns of social life, and census data are among the most meticulously collected social data. However, even these data cannot perfectly represent the social characteristics of the populations they are meant to describe. It is, therefore, essential that researchers using such data be as familiar with their source data as possible. Although census data samples are becoming increasingly available through the IPUMS International project, information about data structure and quality can be quite limited. The more researchers understand patterns of error, omission and bias that stem from complex data collection, manipulation, dissemination and estimation processes, the more accurately the y can describe the populations they analyze. The IPUMS International data collection is comprised of 211 census microdata samples from 68 countries of the world and adds new samples each year. In addition to providing the data itself to researchers at no cost, the IPUMS provides a wealth of documentation about the enumeration procedures used to collect the data. Recently, IPUMS International has undertaken efforts to provide users with additional information about design an d structural characteristics of the microdata samples in the collection . This paper reports on data assessment results comparing age and sex ratios across years for a number of national census samples within countries in the IPUMS International collection. We base our assessment procedures on the newly updated United Nations recommendations detailed in the "Tools for demographic estimation," or Manual XI, a preliminary version of which is available at demographicestimation.iussp. Compared to age heaping assessments such as the Whipple Index or Myers Blended approach, the graphing techniques recommended in the guidelines provide superior means of isolating potential structural issues in the census samples.

Key words: Data quality, age-sex ratios, demographic estimation, census microdata

For more than ten years, the IPUMS International in partnership with statistical offices and census bureaus around the world and with funding from the U.S. National Science Foundation and U.S. National Institutes of Health, has been building a collection of census microdata that has become the world's largest. The primary goals of the project have remained unchanged: preservation, confidentiality, harmonization, and cost-free dissemination to qualified researchers. The database currently contains data from more than 200 censuses 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.

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 that may never have been prepared for usage beyond compiling the published census reports decades ago. All data and documentation are transformed into standard formats for processing. Data are reformatted into the household-person structure understood by the software undergirding the project. Data are reviewed by staff in a variety of ways to complete the tasks of harmonization. IPUMS International staff members run basic diagnostics on household composition and family structure, make comparisons to total population counts where necessary, examine frequency distributions and verify universes for all variables.

A key strength of the interactive web-based data delivery technology is the wealth of data documentation and the organization of that documentation. The work of the first decade has been focused primarily on preservation, harmonization and dissemination of data. These remain the primary goals, but as our user base expands and becomes more statistically sophisticated, we have received increased demand to provide additional information about structural aspects of the data. While we prefer to leave the task of data quality assessments to the data producers and users of the data, we are beginning to provide some basic assessments to aid researchers in targeting their efforts. This paper grows out of a preliminary effort to assess age-sex ratios across multiples years of census data within IPUMS International samples, beginning with countries in Asia.

IPUMS International samples are provided for public use by the statistical or census agencies of the countries of origin. The countries provide responses to basic questions about census data collection strategies and sample design. However, statistics offices differ widely in their approaches to and procedures for data editing, allocation and for preserving confidentiality. Documentation of such practices is scarce. Conducting and presenting this research raises philosophical issue in assessing data structural issues. Census agencies providing public use microdata samples are highly sensitive to issues of data quality and to potential misunderstandings about sampling error that could raise any questions about the validity of published official census statistics. Measures that label structural data anomalies "errors" by default, regardless of whether deviations from smooth age-sex trends reflect "true" population characteristics contribute to hesitation on the part of census and statistical agencies. We assess data structure rather than data quality in this paper.

Given the variation in data editing practices and the limited documentation of such practices, the following investigation should not be treated as an assessment of error or data quality. Rather, it is a documentation of data structural characteristics and could be compared to other official publications created internally at census agencies. Cases in which data appear to perfectly match expectations may indicate high "quality" in terms of response rates and accuracy of data collection and data entry activity. Alternatively, it may simply reflect high degrees of editing. Here we follow UNFPS guidelines for age-sex ratio assessment and find that the recommendations for visualizing data facilitate better understanding of the underlying patterns in the data. These assessment techniques are more revealing of data structure than calculated indices and provide more targeted diagnostics for investigating data anomalies.

According to the updated United Nations recommendations detailed in the "Tools for demographic estimation," or Manual XI[1], the following assessments "should be carried out as a

[1] A preliminary version of Manual XI is available at demographicestimation.iussp.

matter of course before embarking on a process of demographic analysis." Researchers are encouraged to seek as much relevant information as possible from the agency conducting the survey and to bring to bear as much relevant social, economic, historical, political and demographic information as possible.

Typical approaches to assessing age distributions include the Whipple index and Myers blended method, both of which assess age heaping. The UN has proposed an age-sex accuracy index which also yields a score associated with greater and lesser degrees of accuracy in age and sex reporting. The UN Index evaluates the given data against an ideal of smooth and even trendlines over the range of data points. Dips and spikes resulting from actual demographic events will yield scores associated with data "inaccuracy" but which may or may not signal actual errors in the data. Manual XI guidelines suggest eyeballing the age and sex distributions by graphing them.

The Manual recommends three steps in assessing age and sex distributions as diagnostics for targeting "suspect" patterns in the data. At all stages, the Manual recommends graphing data, as graphs present a clearer picture of underlying data structure.

Step 1) Graph age by single year, noting the extent to which data are heaped on particular ages in a patterned way and noticing any other unusual trend anomalies.

Step 2) Group ages to smooth the distribution.

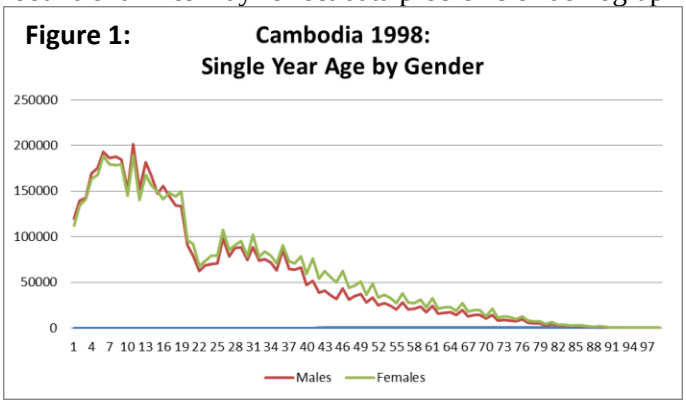
a) Graph grouped ages by sex, noting areas that dip or peak.

b) Calculate age ratios across adjacent age categories to further smooth data, graphing age ratios by sex and noting difference among males and females.

Step 3) Calculate sex ratios and graph by age, flagging distributions that deviate from balanced sex ratios at each age.

In all cases, if more than one census is available, comparisons across years can help identify whether anomalous patterns reflect underlying demographic structures or potential data errors.

Figure 1 illustrates the age distribution for males and females in the IPUMS International sample of the 1998 Cambodia census. Evidence of age heaping is seen in the spikes and valleys, gender gaps at various ages are seen in the gaps between the male and female trends and large divergences from smooth trend lines may reflect data problems or demographic events.

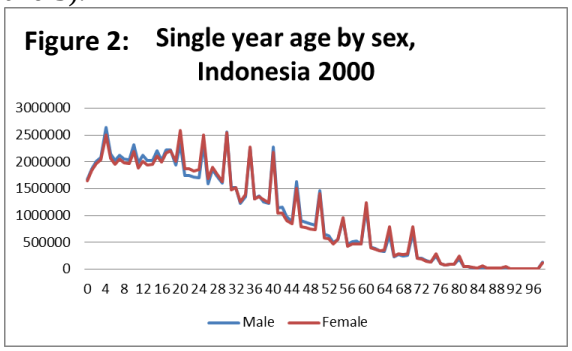


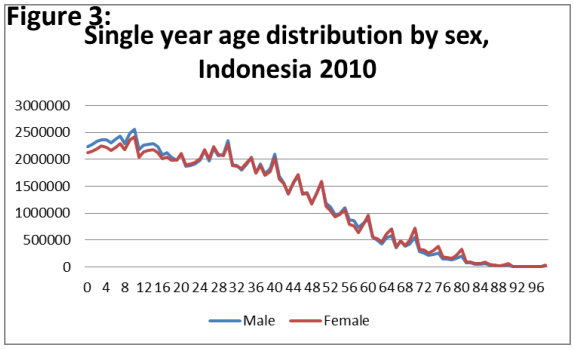
Knowing where these occur helps target research efforts toward investigating which causes are driving the anomalies. The index scores for the 1998 Cambodia sample appear in Table 1. The Whipple score corresponds to a mid-range evaluation of age reporting accuracy, the Myers score is associated with an evaluation of a relatively small amount of heaping, and the UN score is associated with a mid-range of inaccuracy in the data.

Table 1: Cambodia 1998 Age Heaping and Age-Sex Accuracy Indices

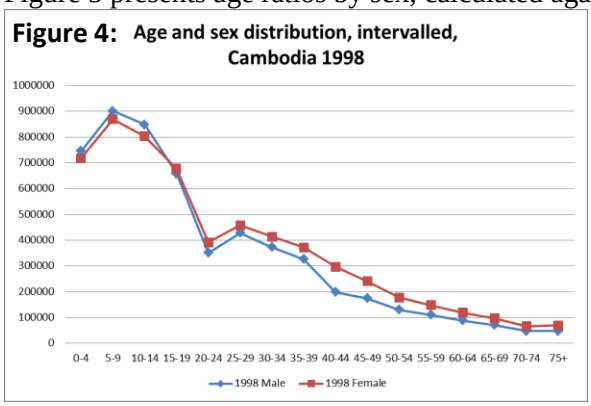
Whipple: 118	Myers Blended: 5	UN Age-Sex Accuracy: 35

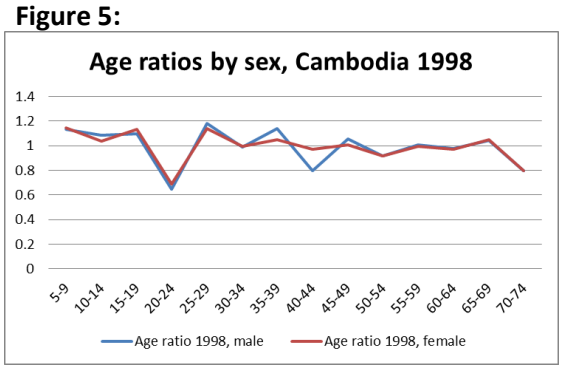
Difficulties in age reporting are all too real for census agencies, particularly in countries where exact age is not a salient social fact. Between 2000 and 2010, the Badan Pusat Statistik (BPS) Indonesia, responsible for administering the 2010 Census undertook enormous efforts to train their census enumerators and create temporal reference tools to facilitate more accurate age reporting for the 2010 Census. Evidence of the value of their efforts is readily visible in side-by-side comparisons of the age-sex distributions across samples from the two censuses (Figures 2 and 3).



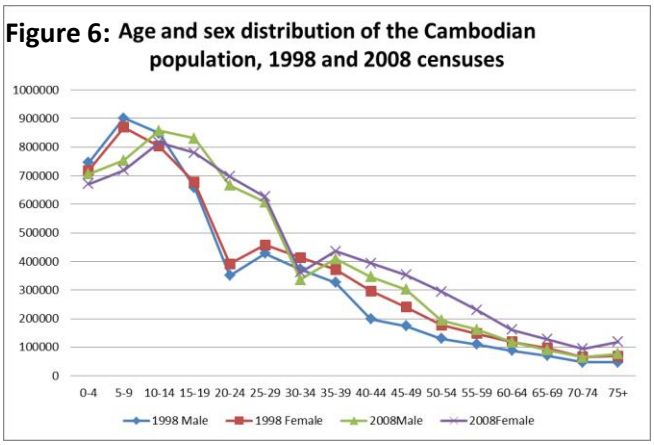


Age heaping persists, but at a much lower level. Even in highly accurately reported census samples, age distributions tend to be jagged or noisy. For this reason, Manual XI recommends grouping ages to smooth the distribution. Non-smoothed areas should be flagged and assessed for whether peaks and valleys reflect demographic events (conflict, migration, changes in fertility or related policy, etc.) or whether they reflect reporting or processing errors in the data. For example, in Figure 1, we see a dip in the presence of both males and females between in their early to mid-20s, which persists even when ages are grouped, shown in Figure 4. Figure 5 presents age ratios by sex, calculated against a smoothed expected value across adjacent

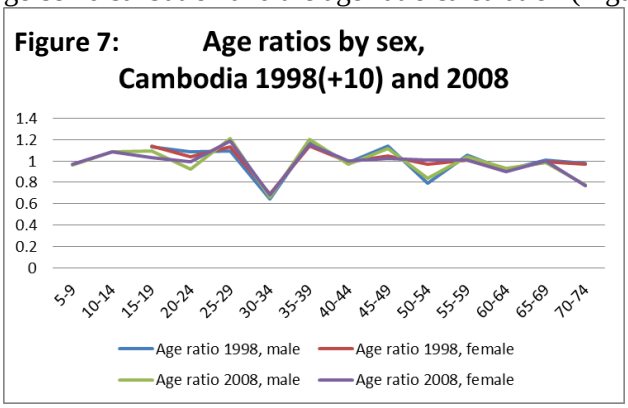




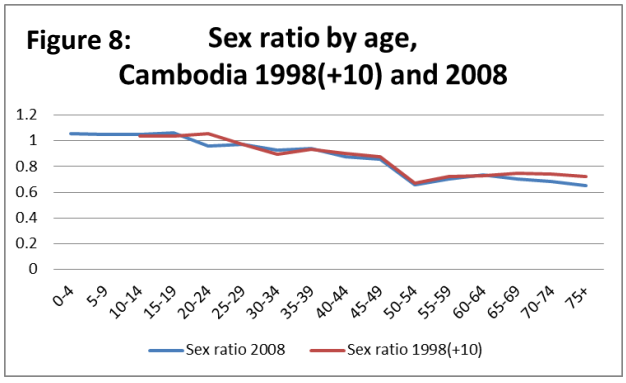
age intervals. The dip in the population of both males and females is even more evident in this diagnostic. The difference in male female rations for cohorts in their late 30s and early 40s is also highlighted here. Manual XI recommends that such anomalies be flagged as suspect; this dip, then, requires additional investigation. While many researchers may be aware of conflict and migration that contributed to the early-20s cohort dip, the availability of multiple years of data can help verify that the anomaly reflects the true underlying data structure rather than a data error, even absent knowledge of Cambodia's political history. Figure 6 includes age and sex distributions for both 1998 and 2008 samples from Cambodia.



Advancing the 1998 cohort by 10 years and overlaying is on the 2008 distribution confirms the similarity in both the age-sex distribution and the age-ratio calculation (Figure 7).



The third evaluation, sex ratios by age, illustrate a different potential problem, that of gender differences in the population at various ages. In many countries, males have been or still are favored over females and have higher survival rates at young ages. Since females often live to older ages, we expect to see lower male to female ratios at higher ages. Figure 8 shows the sex ratios by age in Cambodia 1998 (advanced by 10 years) and 2008. We can see that the gaps between males and females in the age trend lines from figures 4 through 7 are consistent across the two census years, suggesting that the gap may reflect a real difference in the population rather than a data error. We can target our investigation on possible reasons for the dip in the male-female ratio for this cohort.



Similar analyses for Indonesia 2000 and 2010, Vietnam 1989, 1999, and 2009 provide further evidence of the utility of this graph based review of data structures across census samples. In Indonesia, trends in age distribution by sex appear fairly stable across the two census years (Figure 9), but the graph of sex ratios by age (Figure 10) highlights areas for investigation.

Figure 9: Age distribution by sex, Indonesia 2000(+10) and 2010

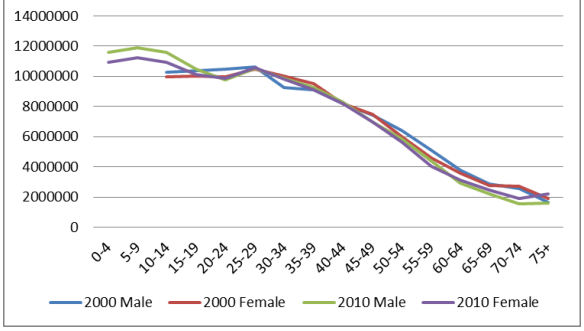
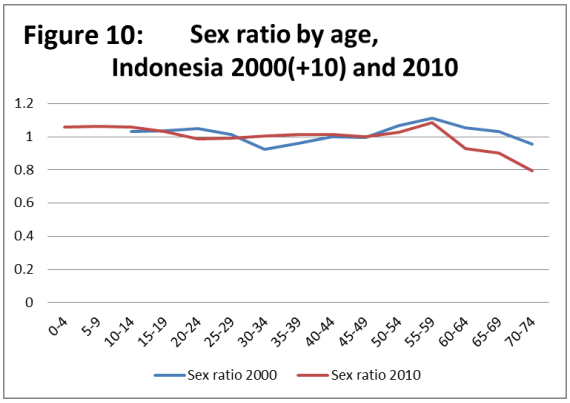
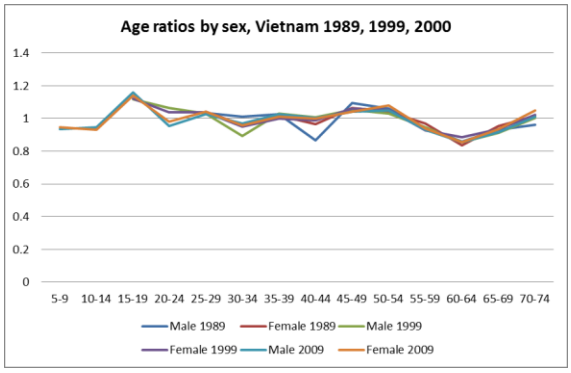
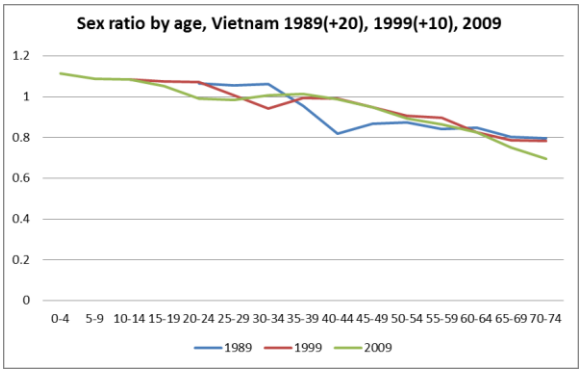


Figure 10: Sex ratio by age, Indonesia 2000(+10) and 2010



In Vietnam, variations in age ratios by sex are consistent across all three samples, but become smaller in magnitude in recent years, as seen in Figure 11. Differences that may warrant investigation are evident in the sex ratios by age. It is likely that the dips may be attributable to temporary male out-migration for labor or education.





The results of these analyses will be made available through the IPUMS International website in the future and will aid researchers in pinpointing structural data issues of interest to their particular type of analysis. Slight imbalances in age or sex ratios may have little or no effect in many social scientific multivariate analyses using the census samples. Other research projects, particularly those that isolate small cohorts or subsamples of the population may be especially sensitive to anomalies. Compared to age heaping assessments such as the Whipple Index or Myers Blended approach, the graphing techniques recommended in the Manual XI guidelines provide superior means of isolating potential structural issues in the census samples.