Statistical coherence of educational attainment in population censuses: IPUMS-International integrated samples compared for 15 African countries

Lara Cleveland*
Minnesota Population Center, Minneapolis, USA – clevelan@umn.edu

Robert McCaa
Minnesota Population Center, Minneapolis, USA – rmccaa@umn.edu

Kristen Jeffers
Minnesota Population Center, Minneapolis, USA – kjeffers@umn.edu

Patricia Kelly-Hall
Minnesota Population Center, Minneapolis, USA – pkelly@umn.edu

Abstract

The IPUMS-International project, now in its fifteenth year, currently disseminates 277 integrated census microdata samples representing 82 countries to more than 10,000 researchers around the globe. As these numbers increase each year, there is a growing need to assess quality. This paper analyzes a rarely addressed dimension of quality, statistical coherence. We focus on primary schooling completed for 15 African countries in successive samples using the intra-cohort comparison method. Successive samples are coherent to the degree that the proportions completing primary schooling are similar for the range of birth cohorts in samples from, say, the 2010 round of censuses compared with the 2000. Nearly perfect coherence is attained by six sets of samples—those for Burkina Faso, Kenya, Morocco, South Africa, Tanzania and Zambia. These show a mean difference of less than one percentage point, R^2 => .93, and, for the regression coefficients, less than +/-0.08 deviations from unity. Overall these results are quite gratifying. The analysis is facilitated by the fact that the microdata are integrated, which is only possible thanks to the generous stewardship of the National Statistical Offices.

Keywords: statistical coherence; population census, integrated samples; Africa; Kenya; IPUMS-International.

1. Introduction

The IPUMS-International project, now in its fifteenth year, currently disseminates 277 integrated census microdata samples representing 82 countries to more than 10,000 researchers around the globe. The microdata are disseminated at no cost, but they are not "public." Access is restricted to researchers and policy makers who agree to the stringent conditions-of-use license. The website www.ipums.org/international disseminates extracts of samples—not the raw data as entrusted by the National Statistical Offices (NSO)—to a tiny, but important constituency—researchers and policy makers who require detailed data on individuals and households to measure and analyze complex relationships, often making comparisons over time and between nations. IPUMS offers a means of disseminating microdata which complements the dissemination activities of NSOs.

For both researchers and NSOs, questions of quality of the samples disseminated by IPUMS are of great concern. Baffour and Valente, in a recent review, define census quality as "fitness for use" and argue that it is characterized by six elements or dimensions: relevance, accuracy, timeliness, accessibility, interpretability, and coherence (2012:122). Coherence is our concern here, although accuracy and coherence are obviously interrelated.

We experiment with a single dimension of quality, coherence, for a single indicator, schooling completed (primary or secondary). We ask the question, how do sample statistics for educational attainment from the most recent census compare with a prior census for the same country? We test

schooling completed not only because universal primary education is a Millennium Development Goal but also because it is measured by most censuses and is the single most widely-available used variable in the IPUMS-International database—even more frequently than sex or age.

We use the demographic concept of birth cohort to generate a series of estimates for each individual year of age from 15 to 89 years for each sample. Figures from successive samples for a single country are then compared. Where statistics are coherent from one census to the next, they will show the same or closely similar percentages completing primary schooling, birth year-by-birth year.

2. Definitions: census coherence and primary schooling completed

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.

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. Baffour and Valente conclude that "ideally the [census] questions should keep the historical formulation to facilitate longitudinal comparison," and any unusual trends or inconsistencies in the data should be explained.

For the 2010-round of censuses, the United Nations Statistics Division recommended "educational attainment" as a core topic and, in post-census 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). "ISCED 1" constitutes primary education, typically 4-7 years completed with six years the most common (UNESCO 2012:17).

3. The intra-cohort comparison method

A population census contains the demographic history of a nation and its people. Successive, high quality censuses of a nation should tell similar, coherent stories. The population historian's tool kit—the principal investigators of IPUMS-International are historians—includes the intra-cohort comparison method, in which a statistic is measured by birth cohorts in successive censuses.

For external coherence we ask the simple question: For each birth year, is the proportion reported completing primary school in a 2010-round census similar to that for the 2000- or earlier rounds? Using Kenya as an example, we pose the question: Is the proportion completing primary school of those born in, say, 1965 the same in the sample for the 2009 census as for the census of 1999? As a matter of fact, the answer is yes, almost exactly: 71.4% of those born in 1965 completed primary schooling according to the 2009 sample compared with 71.3% for 1999—a minute difference and a remarkable testimony to the statistical coherence of the 2009 and 1999 censuses of Kenya.

We extend the question to encompass an entire series of birth cohorts, beginning 15 years before the census (very few individuals complete primary school at a more advanced age) and extending back in time until the absolute frequencies become too small to be reliable, say beyond age 89. We use the product moment correlation and the least-squares regression coefficients to measure the degree of coherence for each pair of series. The older census of the pair is used to predict the percentage for the more recent census. The sample counts of the first census are used as weights. For Kenya 1999 compared with 2009, we find R^2=.97 and b=1.02, which indicates an exceedingly high degree of coherence although not a perfect 1.0, as seen in Figure 1 and Table 1 below. In addition we must take into account sample error. The 95% confidence interval for our estimate of the regression coefficient is +/-.05, that is the range of fit is .97-1.07, indicating an outstanding degree of statistical coherence.

There are at least three caveats for assessing statistical coherence with the intra-cohort comparison method: census agency practices, IPUMS harmonization, and bias. First, the questions, definitions and categories posed in successive censuses and the training of the field enumerators must be taken into account as well as how the data were processed and edited by the national census authority. Second, since we are analyzing samples integrated by IPUMS-International, we must also consider how the IPUMS team harmonized the microdata, and whether the decisions taken to integrate coding schemes in successive censuses are harmonious or not. Third, the method assumes that there are no differentials in mortality, migration or reporting by level of educational attainment. In addition, note that our analysis is confined to sample microdata.[1] Using the full-count microdata would eliminate sample error and would enhance the analysis below. For additional details of the method see Feeney (2014).

4. Integrating educational attainment – the IPUMS-International approach

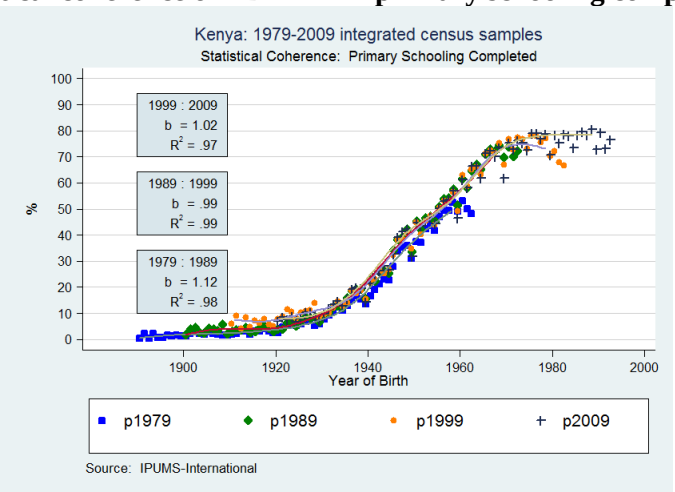
The principal benefit of IPUMS-International to researchers and National Statistics Offices alike is the integration of several decades of microdata samples for each country—typically beginning with the earliest census for which microdata exist or are recoverable and continuing through the 2010 round. When the project began 15 years ago, few NSOs disseminated census microdata. Today most do. Nonetheless, even today, few NSOs write new documentation to facilitate comparative analysis of two or more censuses. Even fewer re-examine earlier censuses to produce correspondence (cross-walk) tables to harmonize variables in successive samples.

The IPUMS-International team writes metadata for six types of information for each integrated variable in the database. These metadata plus links to the original source metadata are available via the seven "tabs" on the IPUMS-International variables pages: codes, general descriptions, comparability discussions, statements of universe, availability of concepts, detailed wording of the original texts ("Questionnaire text," which in turn links to the original source metadata in the official language and English translation), and links to the "source variables" used in constructing each integrated variable.

The goal of IPUMS integrated metadata is to facilitate informed analysis of the microdata by providing as much essential information as feasible—all readily accessible from the website by means of a few clicks. Note that the metadata are open access. Only the microdata must be restricted to respect the conditions of use agreed to by all participating statistical agencies.

5. Comparing integrated microdata for successive samples: primary schooling.

Figure 1. Kenya. Four census samples compared: 2009, 1999, 1989, and 1979 statistical coherence of EDATTAN primary schooling completed



[1] https://international.ipums.org/international/variance_estimation.shtml

Kenya. Figure 1 portrays the primary school completion rates by year of birth, as computed from the IPUMS integrated samples for the 1979, 1989, 1999 and 2009 censuses of Kenya. The curves reveal astonishing coherence, with product moment correlation coefficients of .97-.99 for comparisons of 1999/2009, 1989/99 and 1979/89. Regression coefficients attain near perfect coherence of 1.0 (.99 and 1.02, for the most recent and 1.12 for 1979/89). Perhaps there should be little surprise that the results are so nearly identical because all sets of data were produced by a single statistical agency. Nevertheless, the underlying data in each case were collected by 50-100 thousand field workers conducting face-to-face interviews at four different points in time, separated by intervals of ten years each. The data were processed and coded using increasingly sophisticated technologies that nonetheless offer many opportunities for error. Furthermore, the figures are computed from samples using integrated variables constructed by the IPUMS team with no knowledge that the microdata would be examined this way. Nonetheless the statistical coherence of the results in Figure 1 is striking. Researchers should take comfort in the outstanding coherence between successive census samples of Kenya. Space does not permit a more detailed discussion.

Additional census comparisons. Table 1 summarizes our analysis for 15 African countries. Nearly perfect coherence is attained by six sets of samples—those for Burkina Faso, Kenya, Morocco, South Africa, Tanzania and Zambia. These show a mean difference of less than one percentage point, R^2 => .93, and, for the regression coefficients, less than +/-0.08 deviations from unity. A second group—with mean differences slightly greater and coefficients slightly larger—characterizes pairs for Ghana, Malawi, Nigeria (Post Enumeration Survey), and Uganda. A third cluster—Cameroon, Guinea, Liberia, Nigeria (General Household Survey), and Senegal—shows coefficients substantially different from one and wide ranging mean differences.

Digit attraction in ages, which has little to do with the quality or coherence of census design or execution, explains much of the deviation in statistical coherence. The distortion is accentuated when censuses are taken more than ten years apart, as in the case of all African samples analysed in this paper with correlations below .93. Take the case of Mali, for example. As Table 1 indicates, the three most recent censuses were conducted eleven years apart and the level of digit attraction is high with Whipple total indices exceeding 3.0. Shifting the year of birth back one year for the 1998 census and two years for the 2009 enumeration to synchronize digit attraction in the three censuses doubles R^2 to .91, lifts the regression coefficient within .03 of unity and shrinks the 95% confidence interval by two-thirds. The Malian microdata are certainly "fit for use" if number of publications is our yard-stick. The IPUMS-International bibliography yields 18 citations referencing Mali, including three chapters in the recently released Continuity and Change in Sub-Saharan African Demography (Clifford O. Odimegwu and John Kekovole, eds.) and a chapter in World Population and Human Capital in the Twenty-First Century (Wolfgang Lutz, William P. Butz and Samir K.C., eds.).[2]

7. Conclusions.

Coherence in successive censuses, as measured by the intra-cohort comparison method, is a strong statistical test. Once integrated into a single database, such as the case with samples disseminated by IPUMS-International, the method is easily applied to variables characterized by a "rite-of-passage," such as graduation from primary, secondary, or tertiary schooling.

Researchers must understand that the IPUMS-International integrations are performed *ex-post-facto.* The National Statistical Agency—owners of the data—are not responsible for the decisions taken to design the IPUMS system nor for the resulting integrated codes. In contrast, Eurostat's Census Hub dissemination platform was constructed by European statistical offices before the 2010-round censuses were taken so that integration of concepts, definitions, and categories was designed into the system prior to the actual taking of the 2011 censuses.[3] Thanks to the widespread acceptance of United

[2] https://bibliography.ipums.org/citations/search with keyword "Mali" and project "IPUMS-International" selected.

[3] https://ec.europa.eu/CensusHub2

Nations Statistics Division's *Principles and Recommendations for Population and Housing Censuses*, harmonization of census concepts and categories is possible to a greater or lesser degree.

Census microdata pose challenges for statistical offices with many priorities and a growing public demand for official statistics. The IPUMS partnership for disseminating integrated international census microdata offers substantial advantages at minimal cost or risk. Statistical offices are relieved of many of the most burdensome tasks and responsibilities for anonymizing and documenting samples. The isolated statistical office that disseminates microdata on an *ad hoc* basis incurs substantial risks as well as significant costs in human resources—all for a relatively small return with respect to users. The IPUMS project offers important economies of scale in anonymizing, integrating and disseminating series of census microdata under uniform protocols and with stringent safe-guards, while maintaining the highest standards of quality and coherence.

For the 2020-round of censuses statistical coherence is likely to be even greater than for the 2010-round thanks to the ever-increasing cooperation between the members of the African SSD, the African Centre for Statistics, the United Nations Statistics Division, the official statisticians of the National Statistics Offices, and—most of all—the citizens of the countries where the censuses are taken.

Finally, the IPUMS-International project expresses its gratitude to more than 100 National Statistical Offices world-wide that have embraced the project Memorandum of Cooperation and entrusted high-precision census microdata to the initiative. Those not yet cooperating with the project are invited to consider doing so by discussing participation with the authors of this paper.

References

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

Esteve, A. and M. Sobek. 2003. Challenges and methods of international census harmonization . *Historical Methods* 36: 66-79.

Feeney, G. 2014. Literacy and Gender: Development Success Stories. *Population and Development Revie*w *40*:545–552. DOI 10.1111/j.1728-4457.2014.00697.

Lutz, W. W. P. Butz and Samir K.C. (eds.) 2014. *World Population and Human Capital in the Twenty-First Century*. Oxford: Oxford University Press.

McCaa, R. 2013. The Big Data Revolution: IPUMS-International. Trans-Border Access to Decades of Census Microdata Samples for Three-fourths of the World and more. *Revista de Demografía Histórica 30*: 69-87.

McCaa, R. and A. Esteve. 2006. IPUMS-Europe: Confidentiality measures for licensing and disseminating restricted access census microdata extracts to academic users. *Monographs of official statistics: Work session on statistical data confidentiality*. Luxembourg: Office for Official Publications of the European Communities, 37-46.

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

Odimegwu, Clifford and John Kekovole (eds.) 2014. Continuity and Change in Sub-Saharan African Demography. New York: Routledge.

Spoorenberg, T and C. Dutreuilh. 2007. Quality of Age Reporting: Extension and Application of Modified Whipple Index. *Population* (English Edition), 62(4):729-741.

United Nations European Economic Commission (UNECE). 2014. Group of Experts on Population and Housing Censuses. Quality Management – Draft Text. *Conference of European Statisticians Recommendations for the 2020 Census Round*. Geneva.

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.

Table 1. Statistical Coherence in Primary Schooling For Pairs of Samples: 15 African Countries

			Harmonized Education Variables			Primary Schooling Completed (EDATTAN)					
		Whipple	SCHOOL	YRSCHL	EDATTAN	Born	~55 Over-lapping Birth Years				
		Index	Attendance	Years	Levels	1965	Mean	Difference		R2	b
Country	Year	(Total)	Number of Categories			%	%	Mean	Median		95% ci
Burkina Faso	2006	2.13	5	-	9	11.1	7.4	0.1	0.1	.98	1.03
	1996	2.74	-	-	8	12.1	7.3				+/-.04
Cameroon	2005	3.16	4	25	12	44.4	40.2	11.8	11.5	.64	.68
	1987	3.08	-	26	12	59.5	28.3				+/-.14
Ghana	2010	2.62	4	20	10	47.2	51.9	5.0	5.5	.93	1.02
	2000	3.56	4	20	10	42.9	46.9				+/-.07
Guinea (Conakry)	1996	4.43	4	25	11	24.5	10.5	1.0	0.9	.43	.65
	1983	4.79	4	25	10	19.1	9.5				+/-.20
Kenya	2009	2.11	5	19	11	71.4	46.3	0.2	0.0	.97	1.02
	1999	2.25	4	16	8	71.3	46.1				+/-.05
Liberia	2008	2.09	4	20	9	40.2	18.5	4.5	3.3	.73	1.37
	1974	3.34	3	19	8	27.1	14.0				+/-.27
Malawi	2008	1.39	4	23	9	29.8	21.3	1.3	0.7	.90	1.19
	1998	2.45	3	20	8	28.7	20.0				+/-.10
Mali	2009	3.49	5	26	11	15.3	9.5	1.2	1.2	.45	.71
	1998	3.57	3	27	12	13.3	8.2				+/-.20
Morocco	2004	1.24	-	22	8	29.4	16.6	0.0	0.3	.99	.92
	1994	1.35	-	20	8	30.1	16.7				+/-.02
Nigeria PES (unofficial,	2006	5.69	6	-	9	57.8	42.2	2.6	1.4	.83	.92
unedited, unweighted)	1991	7.70	-	-	7	57.8	39.7				+/-.11
Nigeria (GHS)	2010/11	5.86	5	19	10	48.0	51.4	5.9	4.4	.38	.46
	2006/07	5.33	3	16	8	47.2	45.6				+/-.14
Senegal	2002	3.63	3	20	8	26.3	16.8	2.7	3.3	.39	.67
	1998	2.11	4	22	9	28.5	14.1				+/-.22
South Africa	2001	0.66	2	15	7	74.2	60.6	0.3	0.1	.99	1.01
	1996	0.69	4	17	8	74.1	60.3				+/-.02
Tanzania	2002	2.83	5	18	9	73.4	30.2	0.5	0.7	.93	.92
	1988	3.84	5	18	9	77.5	29.7				+/-.06
Uganda	2002	1.80	4	16	10	46.1	36.8	4.7	3.6	.89	.89
	1991	2.95	4	21	10	44.2	31.9				+/-.08
Zambia	2010	1.56	4	15	9	61.9	49.0	3.7	3.2	.99	.97
	2000	1.71	4	15	9	57.3	45.4				+/-.04

Source: www.ipums.org/international Note: Whipple total index: 0 = best; 16 = worst--preference for a single digit such as zero; see Spoorenberg and Dutreuilh (2007)