Analyzing the quality of age reporting from population censuses for all countries of the world: 1950-2020 census rounds

Age is a fundamental variable in demography and the quality of age reporting reveals population aspects such as socio-economic development, education attainment, and civil registration quality. Within misstatement of age in census data, age heaping is very common. Age reporting tends to be rounded to certain digits such as even numbers. Age heaping is considered to be a measure of data quality and consistency (Pardeshi, 2010). The purpose of this paper was to assess quality for age reporting census data for 1950-2020 census rounds and it is centered around three main hypotheses:

1. We expect to observe a progressive improvement of the Age Heaping Indexes (AHI) through time; however, the rate of improvement for Europe and Africa might be smaller, than other Regions.
    - Europe: Considering that age reporting is strongly correlated to education attainment (or, more broadly, social-economic development), we expect that the European levels of the AHI should be already "good" and because of that, further improvement would be small;
    - Africa: In contrast, in the African case we will observe a smaller rate of improvement in the AHI, with significant differences between the continent sub-regions, lower socioeconomic development, education levels, and low civil registry quality ;

2. Considering we are working with the second half of the XXth century onwards, we expect the AHI for women's shows fewer disparities when compared to men's. However, this observation would be also highly correlated with countries where education attainment for women (as well as gender equality) is high. In those countries where there are larger gender inequalities, the age heaping for women is expected to be larger than for men.

3. When considering the rural and urban areas, we expect that the urban areas to have a larger rate of improvement in the AHI when compared to the rural areas. However, these disparities are expected to be smaller in European and North America countries.

Data and Methods

The census data used comes from the questionnaires filled by different countries that are dispatched annually to the United Nations Statistics Division and then published in the Demographic Yearbook (DYB) since 1948. This information was then compiled into a database that consists of census data for every country (current and former) for the 1950-2020 census rounds. The original data consists of 1061 unique census observations; however, after applying several validations to the dataset in order to guarantee, for example, that there was no age group missing or just one open age group, we end with 785 unique census entries. These census entries may appear multiple times as they contain different information regarding sex (“Male”, “Female” and “Both Sexes”) and area type (“Whole area”, “Urban” and “Rural”).

For estimating the age heaping, we selected four main indexes and then we re-scaled them into a 0 to 1 scale in order for their range to be comparable, being 0 no heaping and 1 meaning that all distribution is centered around a single digit. Those indexes were: 1) Traditional Whipple's index (Roger; Waltisperger; Corbille-Guitton, 1981), Myers' Index (Myers, 1954), Bachi’s Index (Bachi, 1951) and Spoorenberg’s Index (Spoorenberg; Dutreuilh, 2007). We also used the Noumbussi’s method (Noumbissi, 1992) as a heaping index as it improves on Whipple's method by extending its

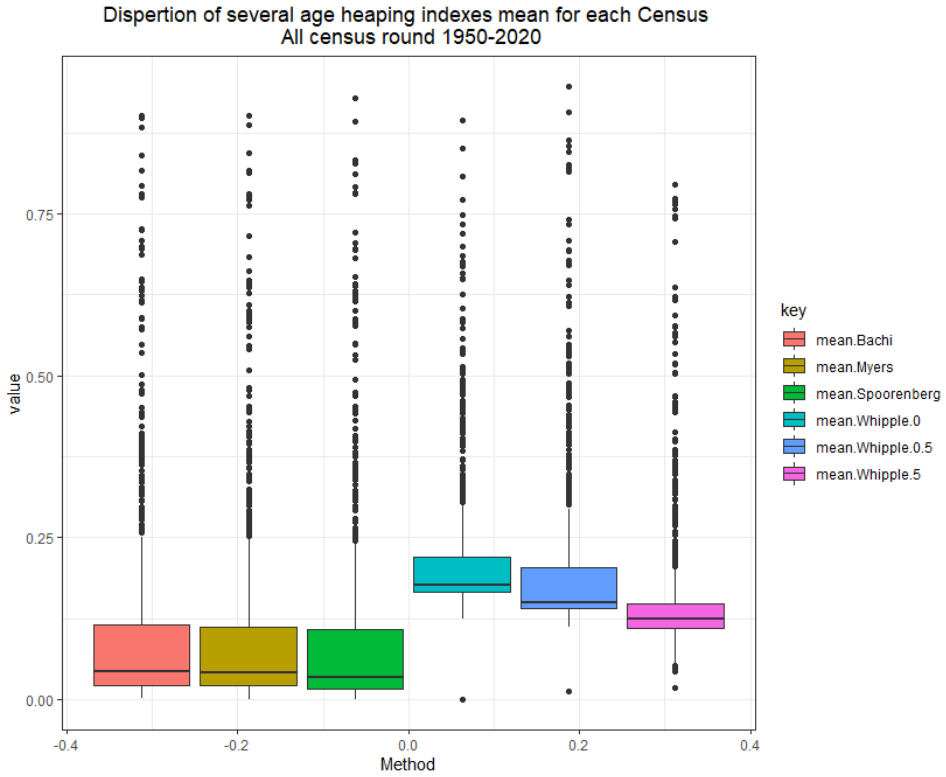
basic principle to all ten digits individually and values higher than 1 indicates a preference for the age's final digit, while values smaller than 1 indicates avoidance.

Preliminary Results

The data is presented here as the average of the AHIs by Sustainable Development Goals (SDG) Regions instead of continent or country for a matter of conciseness. The country groupings for SDG regions are based on the geographic regions defined by the United Nations Statistics Division. These groupings were created taking into consideration the Sustainable Development Goals in the 2030 agenda for Sustainable Development as well as similarities between the countries that make the region (United Nations, 2019).

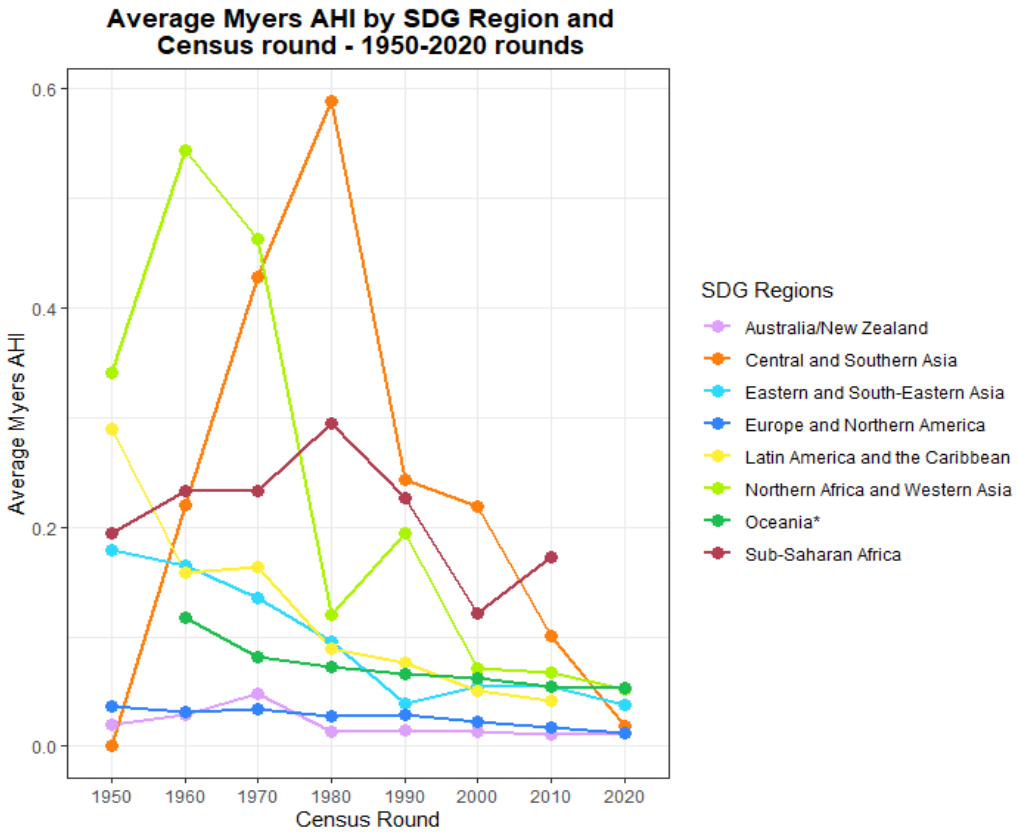
Dispersions

In figure 1 the box-plots show the dispersion of the mean of AHIs estimated for the 1950-2020 census rounds. We can see that as the Whipple's AHI are estimated for only zeros and fives (only zeros, only fives, and both zeros and fives) they show higher values indicating a more intense heaping. However, as they only take into consideration zeros and fives, we opted to use the Myers' AHI throughout the rest of this paper, as it considers all end digits when estimating the heaping and has fewer outliers' dispersion.

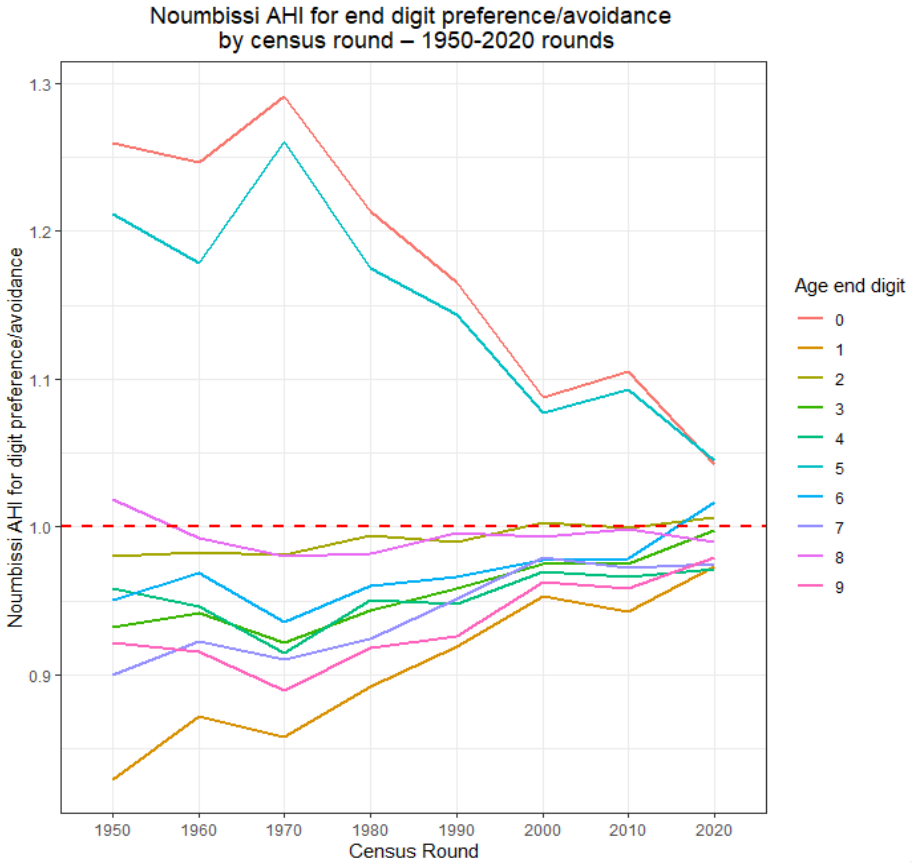


After selecting the Myers' AHI as the succinct representation of the other indexes, Figure 2 shows the average of the Myers by SDG region[1] and by census rounds. The results show that there is a trend toward lower proportions of age heaped. However, we can observe that the highest rate of change between the rounds is not only experienced by African countries but also by Central and Southern Asia countries. Sub-Saharan Africa shows the smallest rate of change considering its high results in the heaping indexes, while Europe and Northern America show a very consistent low rate of change of their low heaping values.

[1] Oceania* refers to Oceania excluding Australia and New Zealand



When we examine digit preference and avoidance by census rounds in figure 3, it is very promising the strong trend of decline for the preference of zeros and fives along with the rounds (measured by its proximity to one, marked by the red dotted line). The digit that 'surrounds' zeros and fives, such as nine or one, shows the opposite behavior as at the beginning of the period they were highly avoided and this avoidance diminishes along the period.



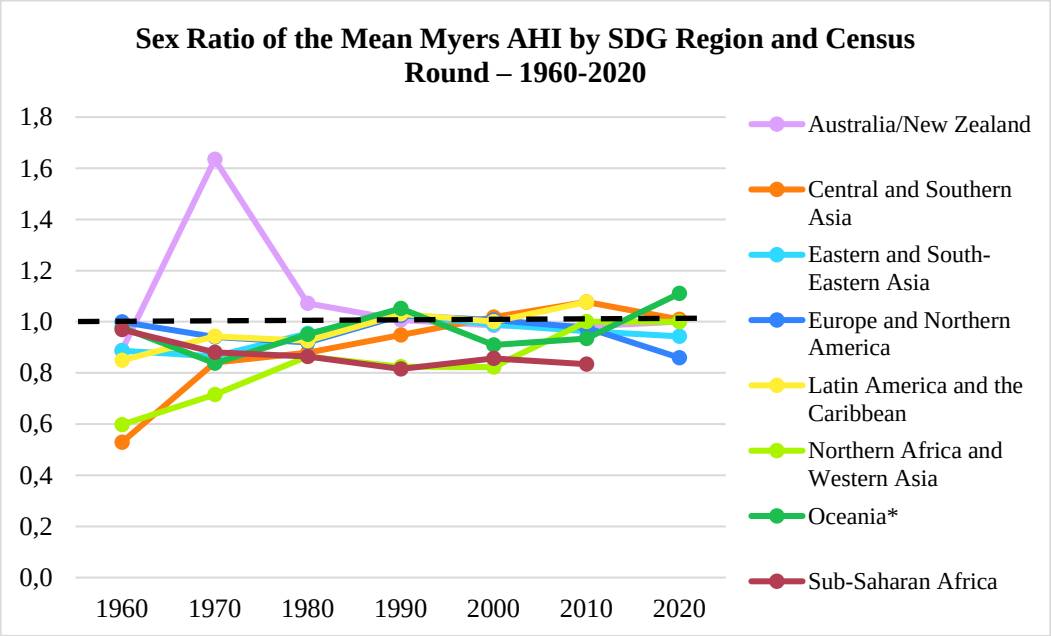
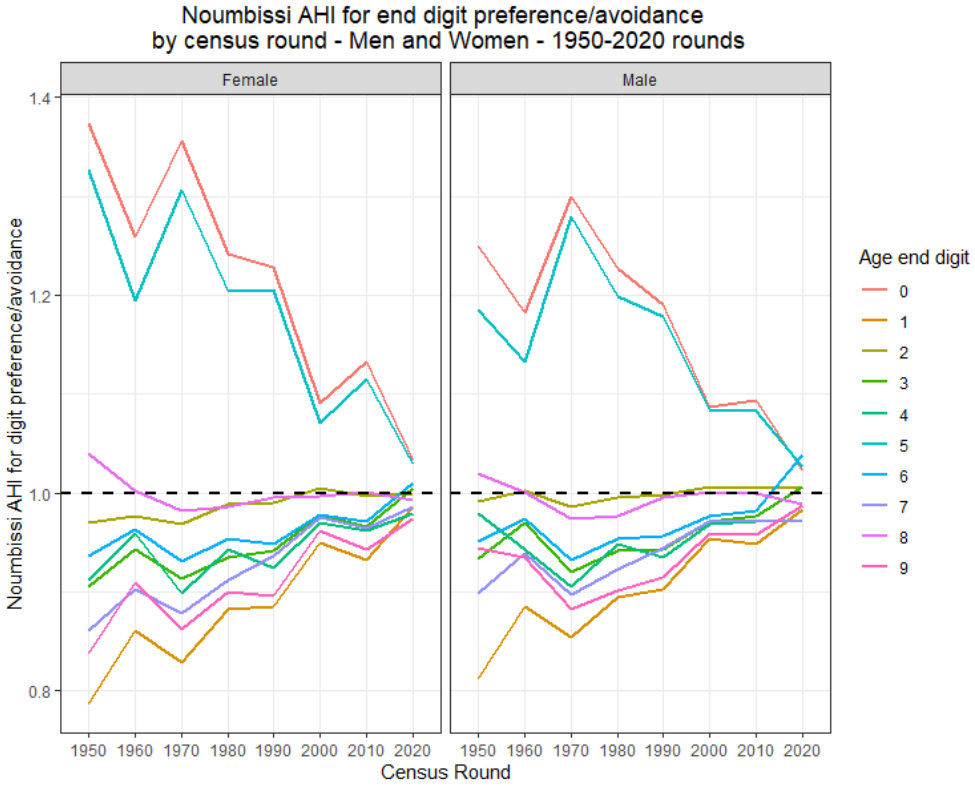
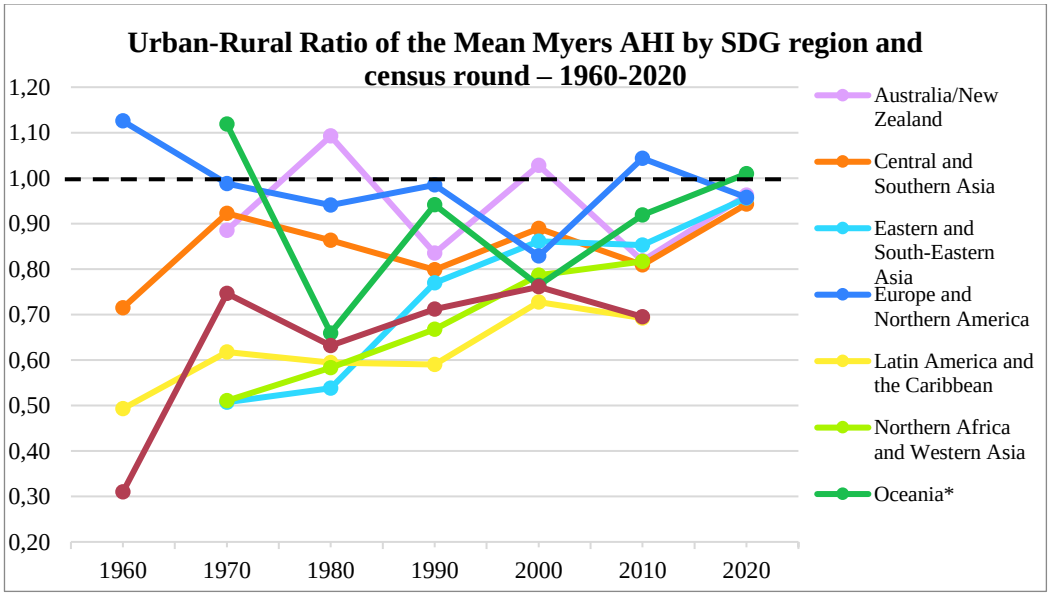


Figure 4 shows the ratio between the average Myers' AHI estimated for men and estimated for women. A ratio larger than 1 (the black dotted line) indicates a heaping more intense towards men, while a ratio smaller than 1 indicates a heaping more intense towards women. The same behavior is seen in figure 5 for the Noumbissi's AHI estimates for both men and women, where women show a stronger digit preference compared to men, but at the same time, a more noticeable fall. As our second hypothesis, the age heaping indexes estimated for women was indeed larger than those estimated for men. This behavior may be because, due to gender inequalities that constrain women, they would be less likely to receive formal education and to the husband being the one that would mostly answer census in the household.



The same as the ratio estimated for men and women, the ratio presented in figure 6 shows the ratio of the mean of the Myers' AHI between urban and rural areas. A ratio larger than 1 indicates a heaping more intense towards urban areas while a ratio smaller than 1 indicates a heaping more

intense towards rural areas. The age heaping for rural areas were more intense than for urban areas, especially in less developed regions There is a slight trend towards reducing the difference in age heaping between urban and rural areas



Urban and Rural areas comparison needs further investigation considering that it is not the same sample of censuses, as not every country has sent this information throughout the questionnaires. Also, this information is only available from 1960 onwards. Therefore, it is important to take a closer look at which countries have and those witch do not have this information, and see if there is a pattern among them.

The R code and an example database will be provided along with the final paper for it to be easily replicable.

References

Bachi R (1951). "The tendency to round off age returns: measurement and correction." *Bulletin of the International Statistical Institute*, 33(4), 195–222.

Myers R J (1954). "Accuracy of age reporting in the 1950 United States census." *Journal of the American Statistical Association*, 49(268), 826–831.

Noumbissi A (1992). "L'indice de Whipple modifiée: une application aux données du Cameroun, de la Suède et de la Belgique." *Population (french edition)*, 1038–1041.

Pardeshi G. S. (2010). Age heaping and accuracy of age data collected during a community survey in the yavatmal district, maharashtra. *Indian journal of community medicine: official publication of Indian Association of Preventive & Social Medicine*, *35*(3), 391–395.

Roger G, Waltisperger D, Corbille-Guitton C (1981). *Les structures par sexe et Âge en Afrique*. GDA, Paris, France. United States Census Bureau (2017). "Population Analysis System (PAS) Software." https://www.census.gov/data/software/pas.html

Shryock HS, Siegel JS, Larmon EA (1973). *The methods and materials of demography*. US Bureau of the Census.

Spoorenberg T, Dutreuilh C (2007). "Quality of age reporting: extension and application of the modified Whipple's index." *Population*, 62(4), 729–741.

United Nations (2019). The Sustainable Development Goals Report 2019. Department of Economic and Social Affairs. https://unstats.un.org/sdgs/report/2019/The-Sustainable-Development-Goals-Report-2019.pdf