United Nations | Department of Economic and Social Affairs

Population Division
UN DESA/POP/2022/TP/No.5
December 2022

Method protocol for the evaluation of census population data by age and sex*

Peter Johnson**, Thomas Spoorenberg***, Sara Hertog*** and Patrick Gerland***

Abstract

As part of its work in revising population estimates and projections for the biennial publication the *World Population Prospects* (WPP), the Population Division of the United Nations Department of Economic and Social Affairs (United Nations Population Division) reconstructs the population changes of all countries and areas of the world starting from 1950 up to today. To assess the consistency of the population reconstruction, reference data sources, such as an existing population census, are used as population benchmarks. For many countries, these population benchmarks are affected by several inconsistencies that need to be examined and possibly adjusted.

This technical paper details the procedures that the United Nations Population Division has developed to assess the quality of the population benchmarks and adjust, if needed, the population data by age and sex. The workflows presented here extend from preliminary steps, such as the definition of the population included in the census and the determination of the territorial coverage of a census, to more advanced analytical methods related to the application of the results of a post-enumeration survey to adjust a population enumerated in a census, the extension of the age distribution up to age 100 or more, the correction of the age distribution due to age heaping or misstatement, and the adjustment for the under-enumeration of the child population.

The steps and procedures included in this method protocol are indeed sufficiently general and broad in nature that they can serve as a reference for the recommended steps to guide national practices in evaluating population by age and sex enumerated in a census.

Keywords: Population by age and sex, census data, survey data, data quality, adjustment, methods, population benchmarks.

Sustainable Development Goals: 17

* The authors wish to thank Stephen Kisambira for his comments and suggestions on the draft.

** Independent Consultant (Formerly with the U.S. Census Bureau).
*** Population Division, United Nations Department of Economic and Social Affairs.

Contents

The Population Division of the Department of Economic and Social Affairs provides the international community with timely and accessible population data and analysis of population trends and development outcomes. The Division undertakes studies of population size and characteristics and of the three components of population change (fertility, mortality and migration).

The purpose of the **Technical Paper series** is to publish substantive and methodological research on population issues carried out by experts both within and outside the United Nations system. The series promotes a scientific understanding of population issues among Governments, national and international organizations, research institutions and individuals engaged in social and economic planning, research and training.

Suggested citation: Peter Johnson, Thomas Spoorenberg, Sara Hertog and Patrick Gerland (2022). *Method protocol for the evaluation of census population data by age and sex.* UN DESA/POP/2022/TP/No.5. New York: United Nations Department of Economic and Social Affairs, Population Division.

This technical paper is available in electronic format on the Division's website at www.unpopulation.org. For further information, please contact the Population Division, Department of Economic and Social Affairs, Two United Nations Plaza, DC2-1950, New York, 10017, USA; phone: +1 212-963-3209; e-mail: population@un.org.

Copyright © United Nations, 2022, made available under a Creative Commons license (CC BY 3.0 IGO)

http://creativecommons.org/licenses/by/3.0/igo/.

EXPLANATORY NOTES

The following symbols have been used in the tables throughout this report:

A full stop (.) is used to indicate decimals.

References to countries, territories and areas:

The designations employed in this publication and the material presented in it do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. The term "country" as used in this publication also refers, as appropriate, to territories or areas.

The following abbreviations have been used:

AGESMTH	Smooths an age/sex distribution using several methods
AP	Adjusted population
BASEPOP	Estimates and smooths a base population consistent with past mortality and fertility
BP	BASEPOP method
BPA	Estimate the population under age 10 using the smoothed population and estimates of fertility and mortality
BPE	Estimate population under age 10 using the reported population ages 10 and above and estimates of fertility and mortality
BPSTRNG	Estimates and strongly smooths a base population consistent with past mortality and fertility
CCM	Cohort component method
CP	Census population
CPDA	Computer Programs for Demographic Analysis
CV	Coefficient of Variation
DA	Demographic Analysis
DHS	Demographic and Health Surveys
DYB	Demographic Yearbook
GBD	Global Burden of Disease
HFD	Human Fertility Database
HMD	Human Mortality Database
ID	Identifier
IHME	Institute for Health Metrics and Evaluation, University of Washington
IPUMS	Integrated Public Use Microdata Series
IT	Information Technology
LDI	Lag-Distributed Income
M49	Standard Country or Area Codes for Statistical Use (Series M, No. 49)
MICS	Multiple Indicator Cluster Survey
NA	Not Available
NCE	Net Census Error
OAG	Open-ended Age Group
OPAG	Estimates the population in the open-ended age group
PAS	Population Analysis System
PASEX	Newer version of the Population Analysis System

PES	Post-Enumeration Survey
Q5	Under-Five probability of dying
RSA	Redistribution Start Age
SDG	Sustainable Development Goals
SINGAGE	Analyzes the population by single years of age
TFR	Total fertility rate
UN	United Nations
UNDESA	United Nations Department of Economic and Social Affairs
UNSD	United Nations Statistics Division
US	United States of America
WHO	World Health Organization
WPP	World Population Prospects

I. INTRODUCTION

The Population Division of the United Nations Department of Economic and Social Affairs (UNDESA) releases every other year a set of population estimates and projections—the *World Population Prospects* (WPP).[1] They form a comprehensive set of demographic data to assess population trends at the global, regional and national levels. The WPP consists of a prospective population reconstruction for each country or area, starting from 1950 up to the present day (i.e., population estimates) and various scenarios of future population development (i.e., population projections) (United Nations, 2022a).

In the WPP, the cohort component method (CCM) is used both to estimate and to project populations by age and sex. The CCM offers a consistent framework for reconciling historical population estimates with estimated levels and trends in fertility, mortality and net international migration. CCM is based on the population balancing equation (Equation 1.1), whereby the national population can only increase or decrease between two points in time (e.g., t_0 and t_1) as the results of births, deaths, and movements of population across national boundaries (i.e. emigration and immigration). The population balancing equation is:

$$Pop(t_1) = Pop(t_0) + Births(t_0, t_1) - Deaths(t_0, t_1) + NetMigrants(t_0, t_1) \qquad (1.1)$$

where: t_0 is the initial time point or base year; t_1 is the second time point.

The use of CCM provides an appropriate framework to produce and validate population estimates by age and sex. Based on available population censuses, sample surveys, vital statistics population registers, analytical reports, and other sources, each component of demographic change can be estimated by age and sex. When applied to a base-year population distribution by age and sex, these estimates are used to reconstruct in an internally consistent way the population change by age and sex between two points in time. The reconstructed population can be compared against an available reference data source (called hereafter 'population benchmarks'), such as an existing population census, to assess the consistency of each estimated component of population change and the reconstructed population. In the event that the reconstructed population does not match closely the population benchmarks, revisions can be made in each component to improve the fit with the benchmarks.

For many countries, it remains difficult to fit the reconstructed population closely to the population benchmark, because the population benchmark presents inconsistencies that are impossible to reproduce with consistent demographic estimates. It is therefore important to first assess the quality of the population benchmark as an initial step to any population reconstruction.

In most of the countries of the world, the principal data source used for population benchmark is a population census. A series of international principles and recommendations on population censuses are available and it is recommended that a census is conducted once every 10 years (United Nations Statistics Division, 2017). Since 1950, more than 1,750 censuses have been conducted worldwide, with varying levels of population and territorial coverage. Population censuses can be affected by various issues, such as patterns of under- or over-enumeration varying by age and sex, as well as age misstatement. Under- or over-enumeration is usually determined based on the results of a post-enumeration survey[2] (PES) or

[1] For further details on the most recent revision of the World Population Prospects, see https://population.un.org/wpp/.

[2] A post-enumeration survey is a survey-based approach, that consists in creating an alternative estimate of the population totally independent of the census by surveying a sample of the population enumerated in the census (United Nations, 2010).

demographic analysis[3] (DA), and several indices have been developed to determine the quality of the declaration of ages.

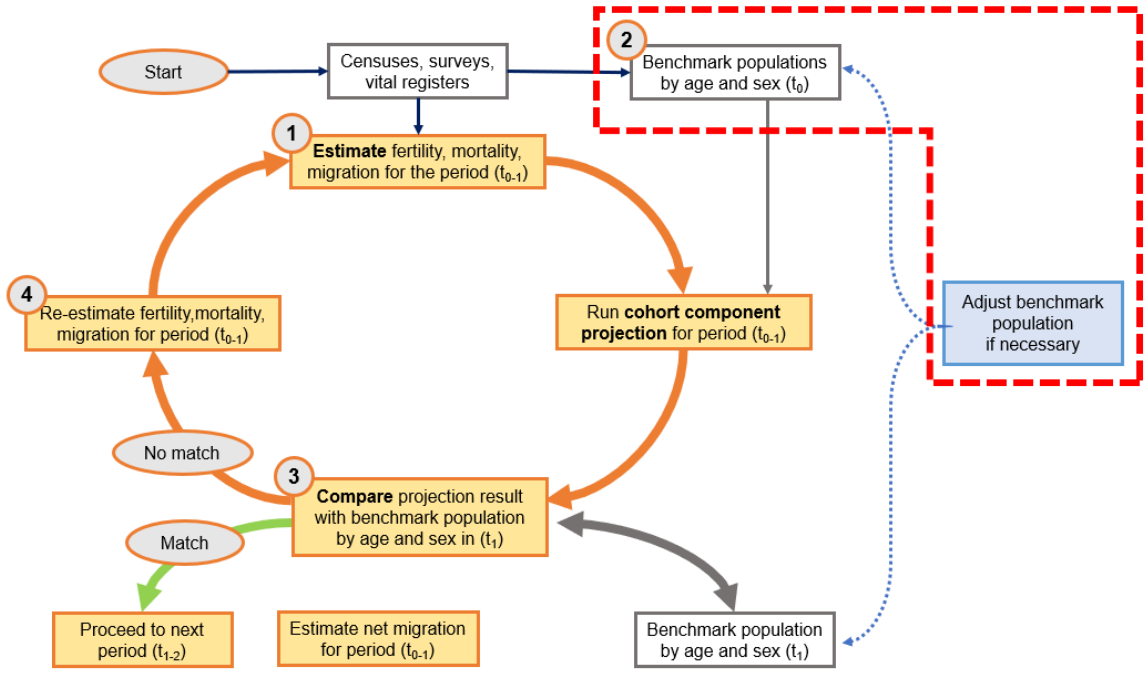
The combination of PES, DA and other assessments of the quality of the declaration of ages allows to address properly the various issues that affect population counts by age and sex across the world. More specifically, the under-enumeration of the child population (i.e., under age 10) needs to be examined carefully. In many cases a PES may not be able to fully capture the undercount of children due to respondent biases in both the census and PES reporting the information about the household composition. In addition, a series of other issues, such as the definition of the population included in the census, the territorial coverage of a census, the format of the population data (e.g. open-age group), need to be carefully assessed before using census data as population benchmarks.

This technical paper details the procedures developed to assess the quality and adjust, if needed, the population data by age and sex that serve as benchmarks in the overall process of the WPP (figure 1.1). The steps and procedures presented in this document were developed for the 2022 revision of the WPP as part of a series of new methodological enhancements, in particular a transition from the historical practice of estimating and projecting across five-year age groups and five-year periods of time to single year age groups and one-year periods of time (i.e. 1x1 framework) (for other enhancements, see United Nations 2022b). The transition to a new 1x1 framework required revisiting the WPP methods, procedures and workflows. An important change is that all the demographic components are now estimated by calendar year and the population estimates refer to January 1 of each year, making the calculations using the demographic balancing equation much easier to implement.

In the overall process of the WPP, the present method protocol covers step 2 related to the population benchmarks (indicated by the red-dotted frame on figure 1.1).

[3] The demographic analysis uses information from previous census(es), complemented by available estimates of births and deaths, data on international migration, as well as any other records (such as immunization records, school enrolment statistics, election listing, etc.) to independently estimate the population at the date of the new census to be evaluated (U.S. Bureau of the Census, 1985).

Figure 1.1. Process used to ensure intercensal consistency between demographic components and the total population

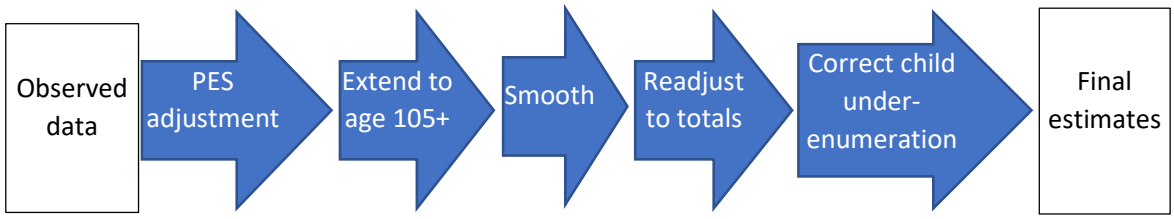


Source: United Nations, 2022b, p. 4.

A. Procedure to assess and adjust population benchmarks

Following several rounds of testing to evaluate over 1,800 national censuses conducted worldwide between 1945 and 2021, the United Nations Population Division determined an overall procedure to assess and adjust population benchmarks (figure 1.2). This workflow consists of several steps. A series of preliminary steps consists of checking the definition of the population included in the census, as well as determining the territorial coverage of the census. Then, starting with the observed data (e.g., a population enumerated during a census), the first step consists of applying an adjustment factor derived from a post-enumeration survey. Second, these adjusted data are extended up to age 105+ years, in the case the observed data are not available up to that age. The data are then smoothed, depending on the degree of the age heaping or misstatement. The resulting data are readjusted to match the total population. The final population estimates, or population benchmarks, are obtained after the application of the correction of the under-enumeration of the child population as applicable.

Figure 1.2. Overview of the population process workflow



Details of the steps, choices, and results of the application of this general workflow. Follow figure 1.3. In developing this workflow for the 2022 revision of the WPP, a series of tools and data infrastructure were developed. Those are briefly presented in boxes inserted in the text.

While developed specifically for the 2022 revision of the WPP, the steps and procedures included in this method protocol are indeed sufficiently general and broad in nature that they can serve as a reference for the recommended steps to guide national practices in evaluating population by age and sex enumerated in a census.

Box 1. Tools for demographic analysis

Over the last half century, the process of doing demographic analysis has been aided by the development of computerized tools. Notable examples include:

- *Computer Programs for Demographic Analysis* (CPDA) produced by the U.S. Census Bureau (Arriaga, and others, 1976). These were originally developed for use on main frame computers and later modified for use on microcomputers.

- *MortPak* (United Nations, 1988 and 2013). These were also originally developed for use on main frame computers and later modified for use on microcomputers.

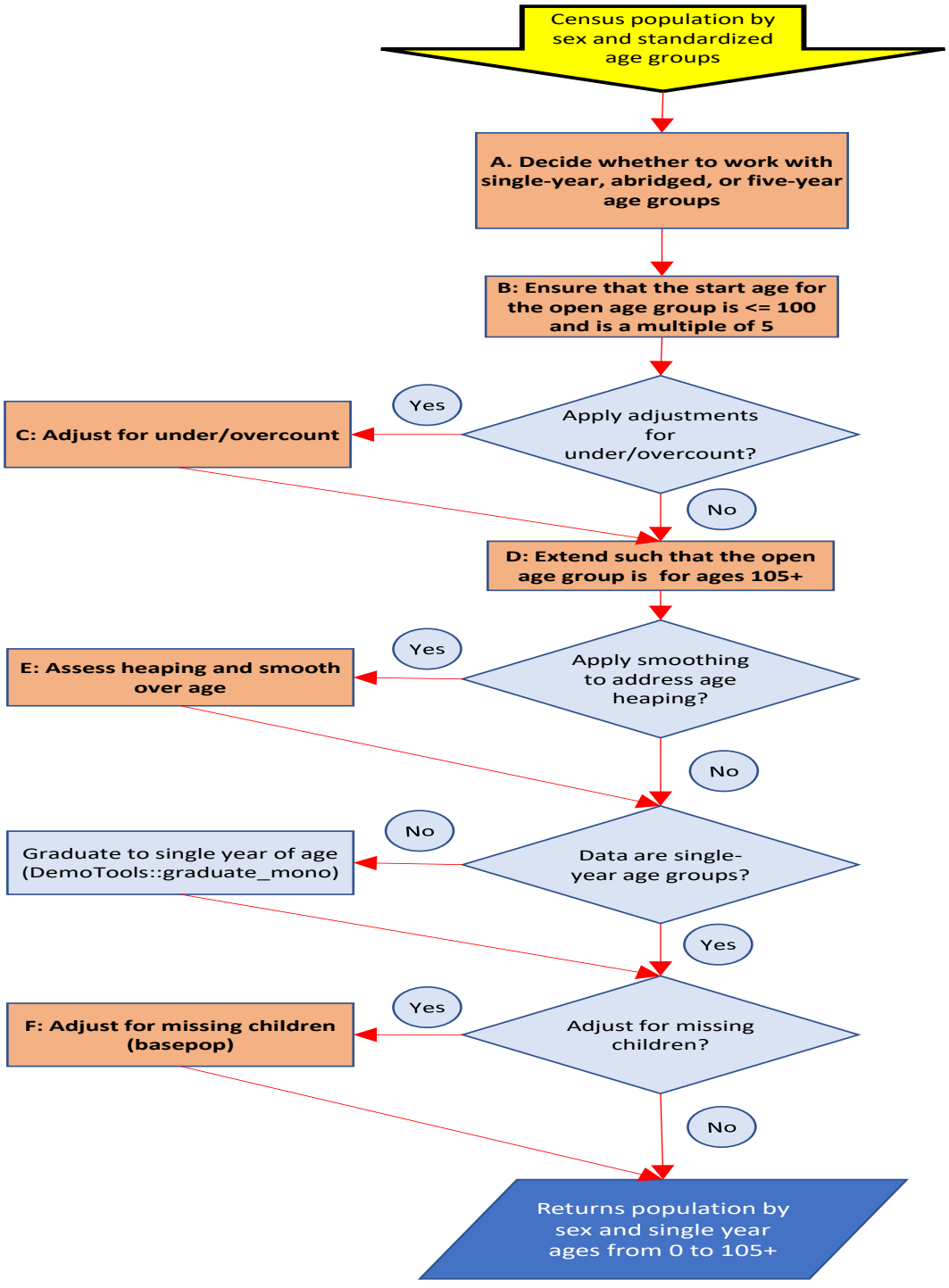
- *Population Analysis System* (originally PAS, now PASEX) (U.S. Census Bureau, 2019). These tools were implemented using spreadsheets (SuperCalc, then Lotus 1-2-3, then Excel where the name PASEX was applied). Since most demographers and statistician are familiar with spreadsheets, and the inputs and outputs are flexible, these tools were more accessible to many demographers.

- *DemoTools* (Riffe, 2021). This is a series of functions in the R programming language. These have the advantage over spreadsheets in being able to be applied on large groups of data. DemoTools is still in development and the implementation of certain methods have and will change over time.

The process or workflow developed for the 2022 revision was programmed using many of the *DemoTools* functions, customized to implement the protocol described in this document.

When applied to a particular country and census, a selection of the procedures included in the workflow can be specified depending on the format, quality, etc. of the data at hand. For example, in the case fluctuations in the population from age to age for a particular census are determined to be real and not the results of patterns of age heaping or misreporting, the smoothing procedures of the method protocol can be skipped.

Figure 1.3. WPP benchmark population adjustment workflow



II. PRELIMINARY STEPS: POPULATION DEFINITION AND CENSUS TERRITORIAL COVERAGE

A. POPULATION DEFINITION

As an initial step in assessing any population data, it is important to understand what population was intended to be enumerated in a census. Such examination is especially important when comparison over time or across data sources is made, because differences between population data may be produced by different population definitions and/or territorial coverages used in censuses over time.

A census is conducted based on a specific definition of the total population to be enumerated (UNSD, 2017: 186). Two broad definitions have been traditionally used:

- the total population may include all the peersons based on their regular or legal residence in a country generally referred to as the *de jure* population, or
- the total population may consist of all the persons present in a country at the time of the census generally referred to as the *de facto* population.

In practice, many countries have switched definitions one or more times over the course of their history collecting population censuses.[4] Also, countries do not usually fully achieve either type of count, because groups of the population may be included or excluded, depending on national circumstances.

The WPP framework uses the *de facto* population definition. Such choice is based on one fundamental idea. A *de facto* definition of a population offers a more inclusive and exhaustive definition of a population. All the persons present in a given country at the time of the census should be enumerated, regardless of their status. In this regard, the *de facto* definition aligns well on the principle of "Leaving no one behind" that is at the core of the Sustainable Development Goals (SDGs).

An increasing number of countries, mostly in statistically advanced countries, are indeed interested in estimates of the count and distribution of their usual resident population. Usual residence is indeed more interesting from planning and policy purposes, because it indicates where people live, and where they demand and consume services.

1. Census territorial coverage

A census usually covers all the areas of a national territory. However, for various reasons, this is not always the case, depending on national circumstances. Census enumerators may not be able to access some portions of the country, for example because of natural disasters, political disputes, or violent conflicts. A preliminary step in evaluating a census is therefore to verify that the whole population of the entire country has been enumerated. In the event that some areas of the territory were not enumerated, the enumerated population counts need to be adjusted to factor in the population in non-enumerated areas.

Several alternative approaches are available to adjust census population counts to account for non-enumerated areas. In some cases, such as when there are disputes over the governance of a territory, a separate census may be available. In other cases, a previous census can provide an estimate of the population size of the unenumerated areas relative to the total population, and that ratio used to adjust the enumerated total.

[4] Over the recent years, some countries have adopted more and more the use of a third definition: the usual resident population. For more details, see UNSD 2017.

The distribution by age and sex of the non-enumerated population also needs to be estimated. One option is to assume that the age and sex distribution of the non-enumerated population follows that of the enumerated population. Alternatively, it may be preferable to assume that the age and sex distribution of the non-enumerated population follows that of a neighboring region or one that is otherwise similar to the non-enumerated territory.

In some specific cases, the territory of a country may have increased or decreased over time to include, as a result, a larger or smaller population. For the WPP, the historical time series of estimates are constructed such that the reference territory is consistent with the present-day borders of each country. Thus, before any historical censuses can be used as benchmark populations in the WPP, they must be adjusted to account for any recent or historical changes in territory, by adding or removing population associated with the territorial adjustment. Failure to adjust the national population for change in territorial coverage can impede the comparability of censuses across time.

Box 2. The IT/Database infrastructure behind WPP including DemoData

Over the years of producing WPP estimates and projections, the United Nations Population Division has developed several databases and an infrastructure to help with the demographic analysis needed for their production.

DemoData is an online database of empirical data and selected tabulations on population by age and sex, fertility and mortality data compiled and used by the UN Population Division to derive estimates and projections from censuses, population and vital registers, surveys, and other sources, going back to the 1950s, depending on data availability. The database contains data with structured metadata from many different sources including:

1. The Demographic Yearbook systems maintained by the UN Statistics Division

2. Tabulations obtained from country publications such as census and survey reports

3. Tabulations generated from public use microdata from census or survey records, such as those maintained by IPUMS (https://international.ipums.org/international/), Demographic and Health Surveys (DHS) (https://dhsprogram.com/data/) and Multiple Indicator Cluster Surveys (MICS) (https://mics.unicef.org/surveys)

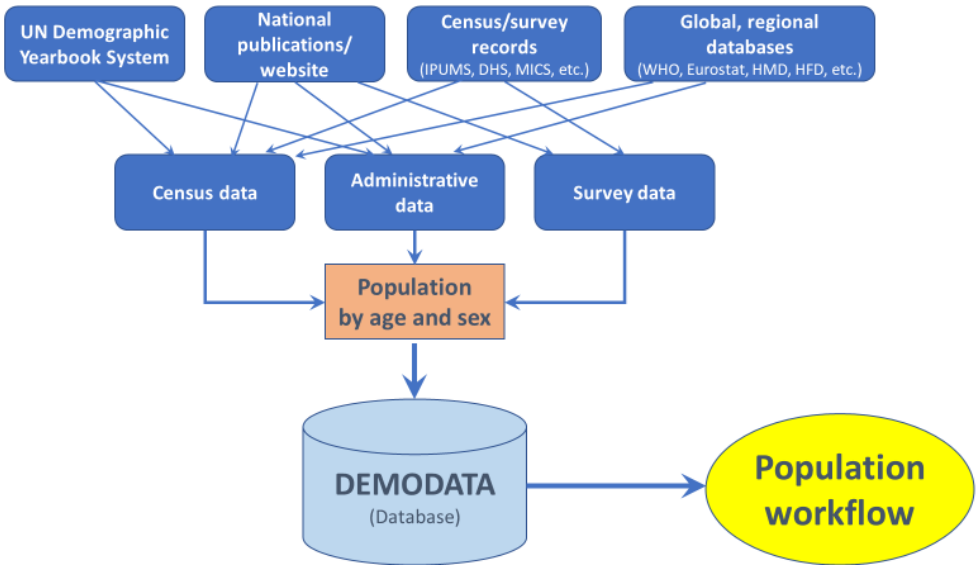
4. Data extracted from global or regional databases (e.g., WHO, Eurostat, HMD, HFD, etc.).

The list of indicators available in DemoData includes the total population by sex and the population by age and sex, including data by single years of age, by five-year age groups, abridged ages (with age 0 versus 1-4 broken out), or unconventional age break-down (e.g. 0-2, 3-6, 7-11, 12-17, 18-24, 25-40, etc.).

The database documents comprehensive metadata on data sources and estimation methods. These characteristics are important to understand what the data refer to, assess their consistency, and how they can help in developing the reconstruction of the population development of countries from 1950 to the present. Information such as the definition of the population (e.g., de jure vs. de facto), data coverage (geographic) and data representativity is available.

Box continued

Figure B2. Data sources and internal data process to produce population by age and sex



III. ADJUSTMENT OF POPULATION FOR ENUMERATION ERRORS

A. INTRODUCTION

A common problem with reported census data is the completeness of the enumeration, including differentials by age and sex. In some cases, a census may be affected by over-enumeration wherein some people are counted more than once. In other cases, a census under-enumerates the population by failing to count some people who should be included. Such errors may arise from deficiencies in various aspects of the census operation, such as field procedures, questionnaire design, data processing and editing, as well as from adverse conditions that may arise in the country while the census was being conducted, such as adverse weather conditions, political or social upheaval, and economic conditions.

Upon the completion of the main phase of a census data collection, it is recommended to conduct a Post-Enumeration Survey (PES), which is designed to reveal the accuracy of a particular census. It should be conducted independently from the census, that is after all census forms have been returned. A PES independently re-enumerates selected census enumeration areas, and then the population counts collected in the PES are compared to those from the census (UNSD, 2010).

The general idea of a PES is that its results are deemed of better quality, because it is conducted only on a smaller population sample size than the census. The results from the PES are compared to the results of the census and adjustment factors are computed and applied to adjust the census count to correct for any over- or under-enumeration.

If we specify that:

CP= Census total population

AP= Adjusted total population (or "true" population, i.e., the population estimated from the PES).

Then the ratio of CP/AP is called the completeness of enumeration, symbolized as μ.

$$\mu = CP / AP \tag{3.1}$$

The overall net enumeration (also called net census error or NCE) is just the relative difference from the "true" value, usually expressed as a percent. Negative values mean net underenumeration. The NCE or net census error (expressed as a proportion) is computed as:

$$NCE = \frac{CP}{AP} - 1 = \mu - 1 \tag{3.2}$$

and AP, the adjusted total population, corresponds to:

$$AP = \frac{CP}{(1 + NCE)} \tag{3.3}$$

Figure 3.1 summarizes the steps developed in the method protocol to adjust the population by age and sex. The sequence rests on two main steps: either a PES was conducted after a specific census (and it is available) and its results are used to correct the corresponding census count, or no PES was conducted and an expected value based on a statistical model is applied. Ultimately, the application of these analytical steps is conditional to the extent that the publicly available results of the census have already been evaluated, and adjusted by national statistical authorities (e.g. United Kingdom "one number census" approach5). In instances where available census results have already been adjusted for net enumeration errors either through PES and/or DA by national authorities, no additional PES adjustment is applied.

While recommended, many censuses are not followed by a PES, mostly due to budgetary and other constraints.[6] A common complementary, or alternative practice is to evaluated the census results using demographic analysis (DA) which relies on information from previous census(es), complemented by available estimates of births and deaths, data on international migration, as well as any other records (such as immunization records, school enrolment statistics and election listing) to independently estimate the population at the date of the new census to be evaluated (U.S. Bureau of the Census, 1985).

B. MODELS OF PES-LIKE ADJUSTMENT BY AGE AND SEX

Not every census has been followed by a corresponding PES. Moreover, not all PES results are publicly available outside of national statistical offices. For censuses without an available PES report, some assumptions are required to adjust population counts for enumeration errors. The Population Division has developed models to estimate country- and time period-specific enumeration adjustment factors. These models are based on results compiled from as many available PES as possible. Of the 1,758 census populations by age and sex referenced for the 2022 Revision, 310 had associated post-enumeration survey

[5] https://ons.gov.uk/census/2001censusandearlier/designandconduct/theonenumbercensus.
[6] https://unstats.un.org/unsd/demographic/meetings/egm/symposium2001/docs/symposium_10.htm#_Toc9220849

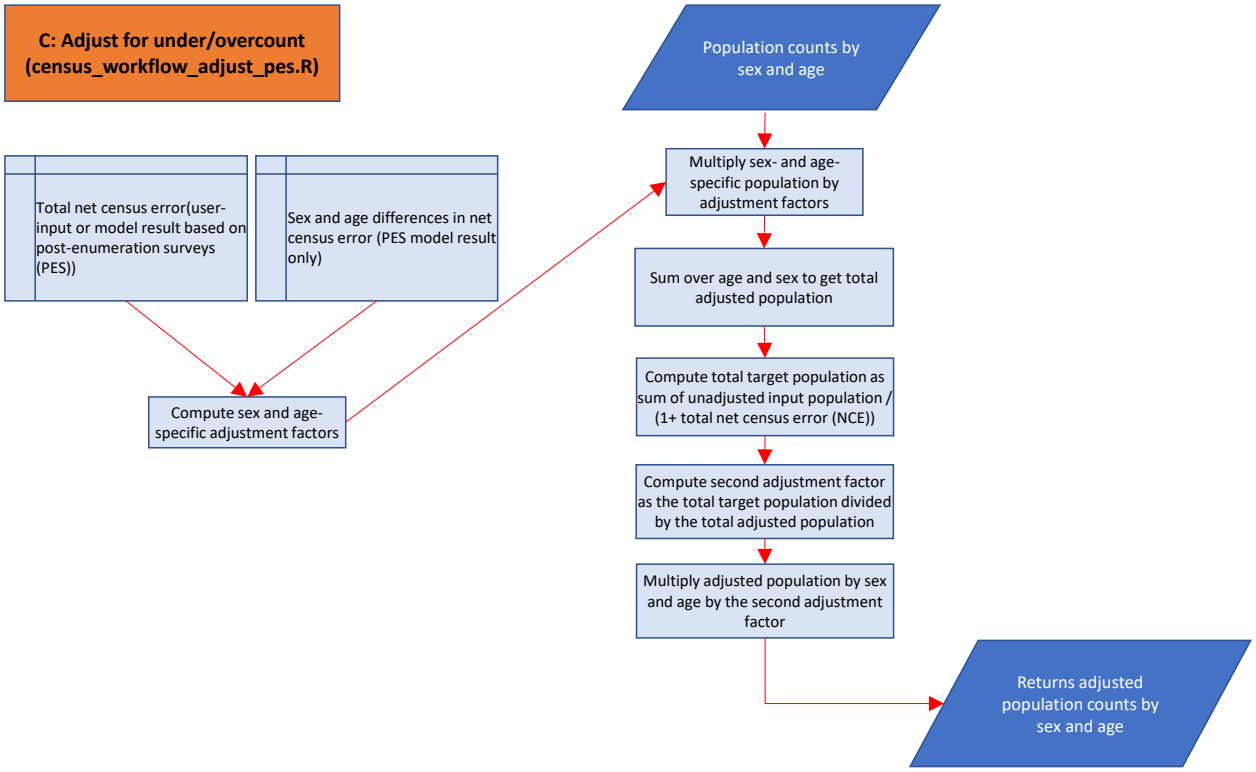
(PES) results available, and 126 of them had some demographic analysis (DA)[7]. These data cover 130 countries between 1946 and 2019 and together formed the basis of models to predict enumeration adjustments for all countries and time periods.

Three models were developed:

1. A model of the overall net enumeration error level (based on 310 PES and 126 DA estimates of the net census error for selected censuses covering 130 countries between 1946 and 2019).
2. A model of the differences in the level of net enumeration by sex and age (based on PES data from 56 censuses for 28 countries between 1950 and 2016). For further technical details, see the Technical Annex in Section IX.

The estimates returned by the first model describe the net enumeration error as a percentage of the total population after adjustment. Positive values indicate over-enumeration, while negative values indicate under-enumeration. The second model returns the sex- and age-specific differences in enumeration errors compared to the total. It is a difference of percentages. For example, if the net census error returned by the first model is -3.0 per cent and the estimate for males aged 35 returned by the second model is +1.0, then the counts of males aged 35 are adjusted to account for a net census error of -2.0.

Figure 3.1. Workflow to adjust for census under- or overcount



[7] The demographic analysis uses information from previous census(es), complemented by available estimates of births and deaths, data on international migration, as well as any other records (such as immunization records, school enrolment statistics, election listing, etc.) to independently estimate the population at the date of the new census to be evaluated (U.S. Bureau of the Census, 1985).

IV. EXTENDING THE POPULATION OPEN-ENDED AGE GROUP (OAG) UP TO AGE 105+

The population reconstruction that is conducted in the WPP requires the population data to be available until age 100 and over8. As an ever-growing percentage of the population is reaching older ages, it is important to have as precise population data and in a standardized format as possible for these age groups. In many countries, the population data by age are not available in the format required, however. Typically, census data are not available to the very advanced ages and the age distribution needs to be extended by some other means.

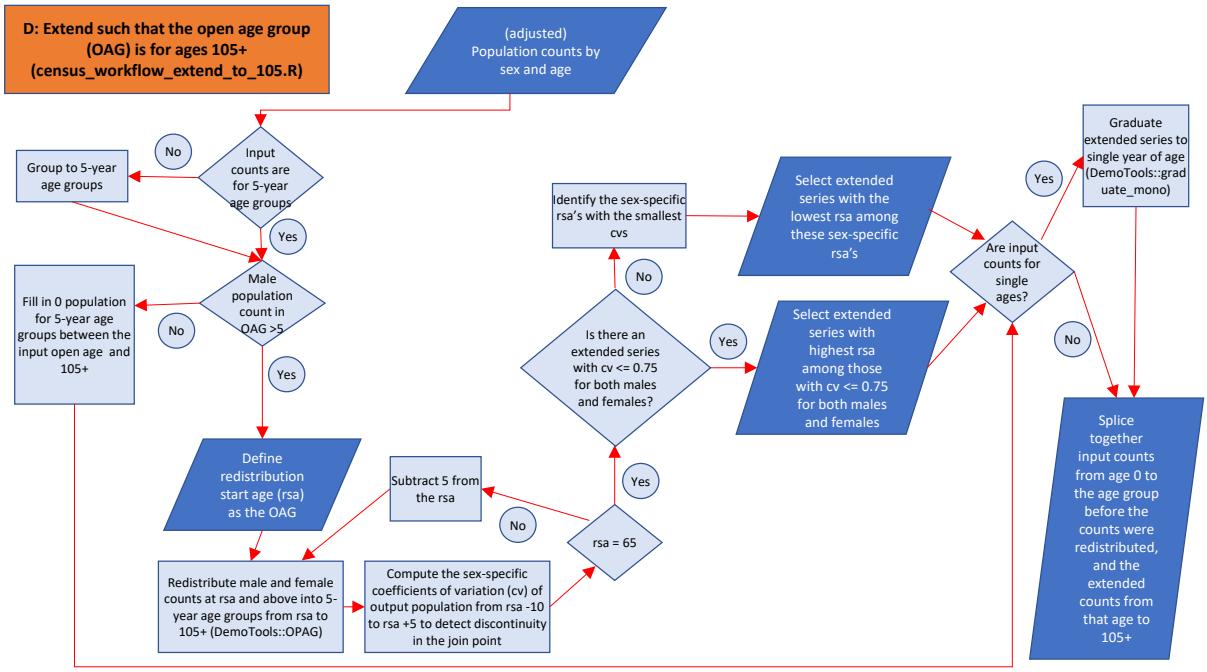
The procedure is based on the principle developed in a workbook called OPAG developed by The U.S. Census Bureau (PAS) that extends the open-ended age group to age 80+ by assuming that the older population was a stable population (Arriaga and others, 1994, pp. 45-47). A later version of this workbook, Pop100h, extends single-year age distributions up to 100+, and includes several options for how to do the extension.

For the WPP, census population age distributions were extended systematically to age 105+ by generalizing the OPAG approach. This was done by applying the life table estimated for a given country and census year as a stable standard. WPP used an updated version of OPAG available in the DemoTools R package that divides the task into two steps. First, the program finds the value of r that minimizes the sum of the absolute differences between the observed population and the stable population for the last two 10-year age groups before the open age group of the original census series. Second the population above a certain age is redistributed through age 105+, with the age from which redistribution begins selected so as to minimize the discontinuity between the empirical age distribution and the extended one (based on the coefficient of variation (cv) for both sexes).

Figure 4.1 summarizes the procedure and the steps developed to extend the population up to age 105+.

[8] The population is indeed extended up to age 105+ to guarantee a population as smooth as possible until the age 100+, and the final results are collapsed back to 100+ for the rest of the analytical steps.

Figure 4.1. Workflow for the extension of the population up to age 105+



V. SMOOTHING THE POPULATION TO REMOVE THE IMPACT OF AGE HEAPING AND MISREPORTING

A. INTRODUCTION

The assessment of the quality of the distribution of the population enumerated in a census by age and sex is one of the most basic and important steps to be taken before using any census data. Any unusual patterns in census age distributions that cannot be explained by extraordinary historical events (e.g. crisis, war, etc.), or any other known factors (e.g. sudden mortality and/or fertility changes, selective migration, etc.) that could affect the size of specific birth cohorts should be examined carefully.

One pattern that is widely found across population censuses is age heaping and age misreporting. Such pattern bears strong implications for demographic analysis (Ewbank, 1981).

Age misreporting means that the age reported in a data collection operation (e.g., census, survey, or death certificate) is not reported correctly. In most cases, the phenomenon of age misreporting is the result of a lack of accurate knowledge of age by the respondent. This can be because the respondent does not have enough education to accurately report the age or the society does not, in general, keep track of dates and ages. In other cases, a census enumerator may simply estimate the age based on the appearance of the person. In censuses and in most household surveys, the ages of all the members of a household are typically obtained thru proxy respondent, i.e., the person of reference or head of household. In such instances, the risk for age misreporting increases when another member of the household or even a neighbour is responding when household members are absent.

One factor that can affect age heaping is the form of the question asked in the census. If the question is simply "age" then the person might not understand the concept of age at last birthday and/or (depending on when they are answering) the concept of the reference date of the census. The use of the date of birth (DoB)

can help in areas where respondents are familiar with the calendar and likely to know this information. If the question only asks for the year of birth (YoB) then the age is ambiguous, and this is true even if the question asks information on month and year (depending on the reference date of the census). When both age and date of birth are asked, then the processing procedure should have decision rules of how to resolve inconsistencies between the two inputs. Sometimes, if only DoB is asked, there ends up being heaping on birth years that end in 0 (or 5), which may cause ages other than 0 and 5 to have heaping if the census year is not conducted in a year ending in 0 or 5. For example, the Central African Republic census of 1988 shows heaping on ages 0, 5, and 8, with digit 0 being the highest, but 8 being higher than 5. In the 1993 census of Gabon, the heaping on age 3 is highest for males and second highest for females.

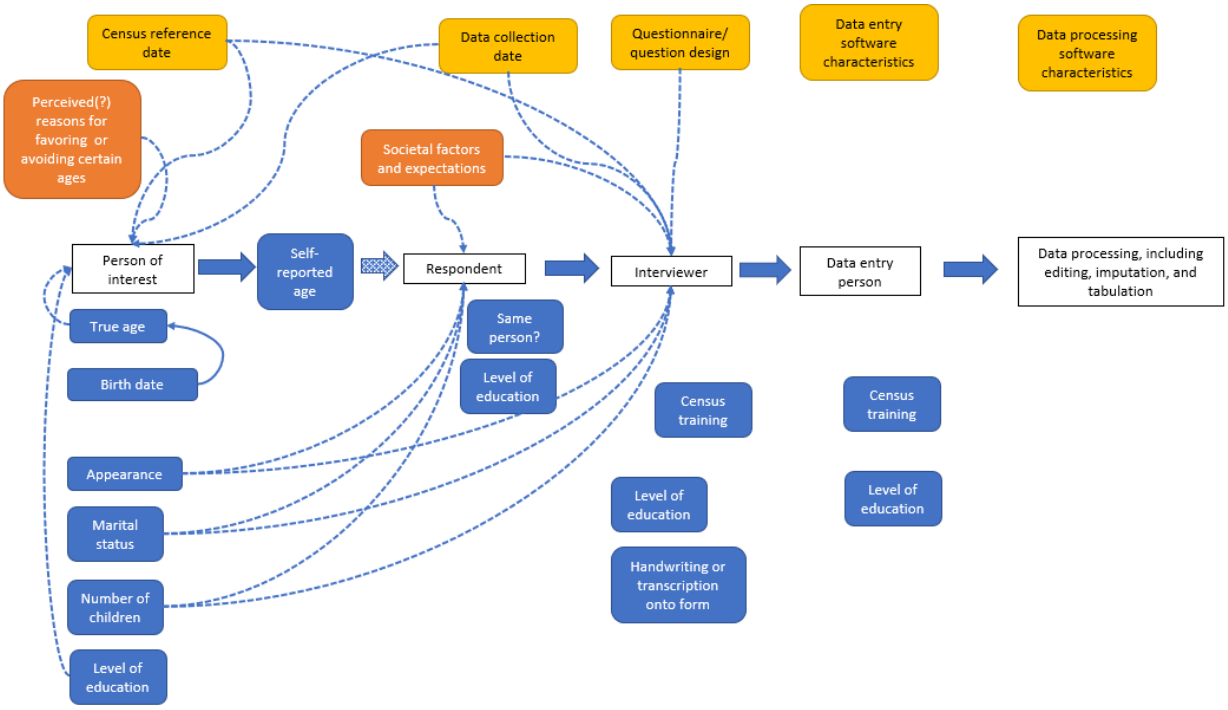
Thus far, no systematic efforts have been made to record in a central repository the form of the age question used in a census or survey and how, in the case the question was not just referring to age or date of birth, these responses were reconciled.

The reference date of the census can also have an impact on the accurate reporting of age. If respondents do a simple mental calculation of the census year minus the birth year, they will make more of an error the earlier the census is in the year. This will also tend to overstate the true age.

Problems with age reporting accuracy can be related to several causes:

- Respondents rounding the ages (or year of birth) to end with the digit 0 or 5 (or even numbers);
- Age overstatement (e.g., in part to gain prestige or pensions);
- Age understatement (e.g., to avoid military service);
- Age under- or over-statement of women depending on societal norms about childbearing (e.g., under-stating ages of childless women or over-stating ages of women with higher-than-normal numbers of children);
- Ages being reported by respondents without full knowledge (e.g., relatives or neighbors);
- Problems with census forms or procedures that end up recording the age incorrectly. For example, in the 1950 US census "a few of the cards were punched one column to the right of the proper position in at least some columns. The result is that numbers reported in certain rare categories—very young widowers and divorcés, and male Indians 10–14 or 20–24—were greatly exaggerated. " (Coale and Stephan, 1962).

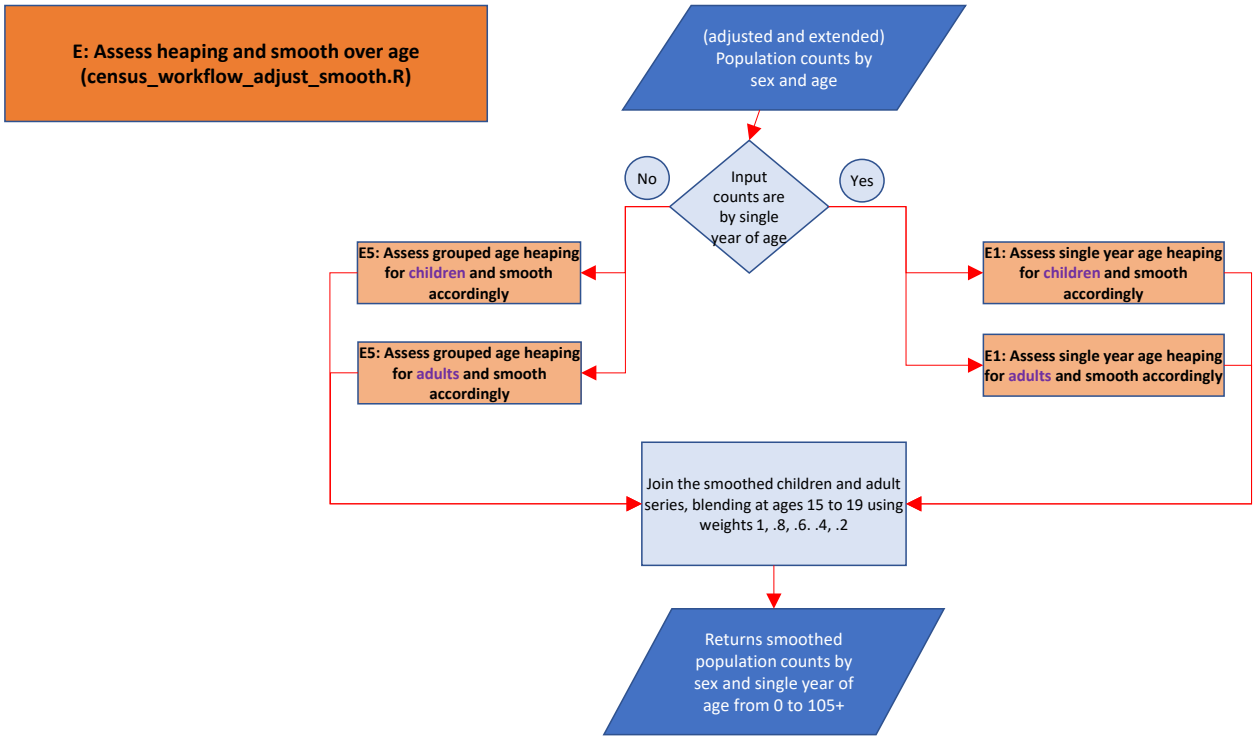
Figure 5.1. Factors affecting age reporting in census tabulations



One way to correct for age misreporting or age heaping is to smooth the population age distribution. The difficulty, however, is deciding what degree of smoothing is needed to remove the impact of age misreporting without altering real information about the size of the different birth cohorts. The degree of age misreporting needs to be estimated and that information should be used to inform the smoothing process.

For the WPP, a series of analytical steps were developed to determine the degree of age heaping and what amount of smoothness needed to be applied. Figure 5.2. provides an overview of the general process. The sections that follow refer to each step shown in that figure.

Figure 5.2. Workflow to assess age heaping and to apply a smoothing method



Population data are usually available either by single years of age, by five-year age groups, or abridged age groups (that separates age 0 from ages 1-4).[9] To correct problems with the age reporting, the degree of age misreporting or age heaping needs to be measured. Several measures have been proposed for this purpose. Depending on the availability of the age data, different measures apply.

For data by single years of age the measures include:

- The Whipple Index (Siegel and Swanson, 2004) which measures the degree of heaping on age digit ending in 0 and 5.
- The Whipple Index was extended to each ending digit by Noumbissi (1992) and a summary index was proposed by Spoorenberg (2007) based on Noumbissi's development.
- The Myers Index (1940 and 1954) estimates the population reporting by each ending digit and looks at the sum of the reported deviations.
- The Bachi Index (1951 and 1953) is similar to the Myers index. Bachi introduced the idea of using as the index half the sum of the absolute deviations by digit, which can be interpreted as the minimum proportion of the population that misreported age. This approach was later adopted by Myers (1954).

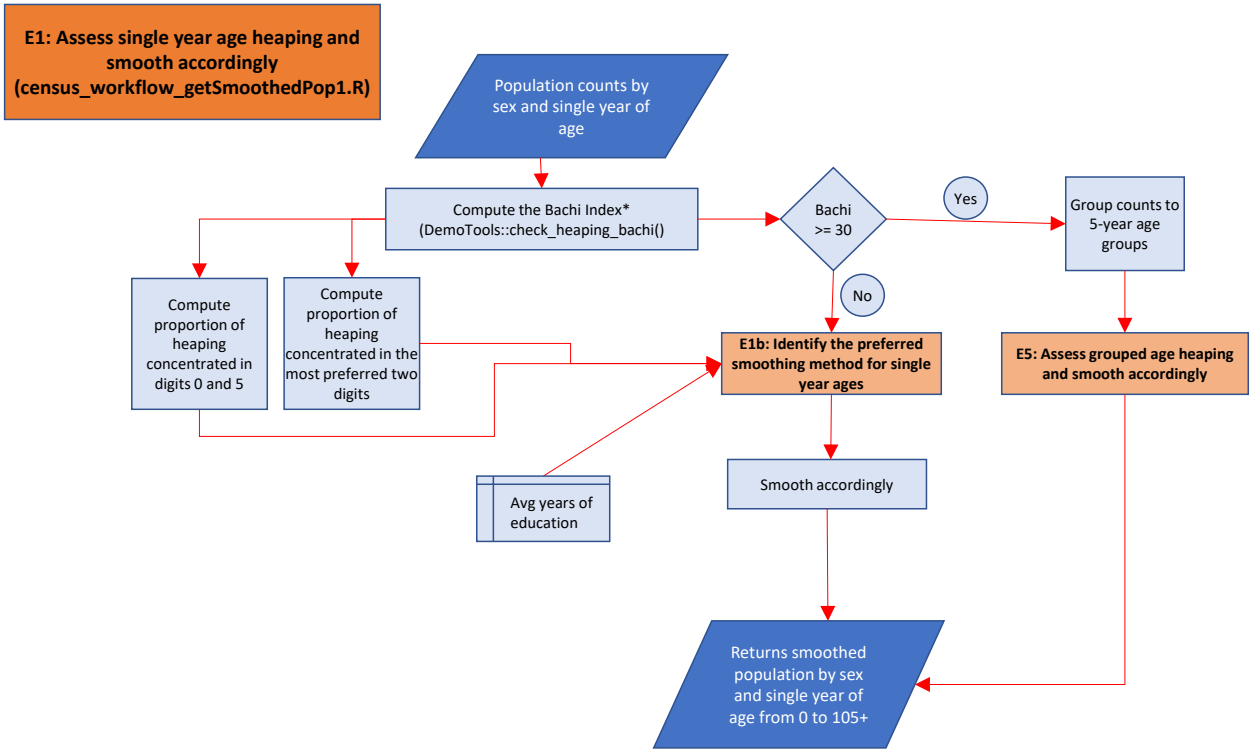
An analysis of each of these various measures applied to census population distributions by single year of age (de Oliveira Carnevali and Gerland 2021) showed that they are highly correlated. The Bachi index was the most highly correlated with the others. Based on this result and additional more detailed analysis,

[9] Rarely, and more often for censuses conducted in the 1950s, the population data can be available for unconventional age groups. In such case, the unconventional age distribution is interpolated or graduated into five-year age groups.

the Bachi index and its deviation by age were selected as the preferred measures of age heaping for WPP. Also, the Bachi index is based on the fact that it assesses deviation on a 10-year age range (like the Myers index), whereas the original Whipple index and its later extensions by Noumbissi and Spoorenberg are based on a five-year interval centred on the age of interest, precluding measurement of the relative strength of the preferences for age ending in 0 versus those ending in 5.

Figure 5.3. summarizes the main steps in assessing the data quality of the age declaration and the need to smooth the data. These steps are applied separately to the adult population and to the children population. The Bachi index and the smoothing procedures are first reviewed in general terms. The procedures for the adult population are then presented, both for the single year age distribution as well as the abridged (five-year) age distribution, before turning to the population of children.

Figure 5.3. Workflow to assess single-year age heaping and smoothing method



1. *Bachi index*

In the 1951 and 1953 sessions of the International Statistical Institute, Roberto Bachi, from Hebrew University and the Central Bureau of Statistics, Jerusalem, presented two papers (Bachi 1951 and 1953) that detailed the "measurement of the tendency to round off age returns." Although the index that bears his name is used throughout, it appears that much of the calculations were done by comparing the reported data to graduated data. The idea of the method (similar to Myers) is to compare the population with age ending with a certain digit to the average population over a specific range of ages. Bachi computed this two times, called Approximations I and II, where the age range used for Approximation II was shifted up by one year. Bachi computed the percentage reporting with digit d, as the ratio of the population in the range ending in that digit divided by the total over the whole range.

Looking at the description and examples presented by Bachi, it was not clear how to modify the approach if the minimum starting age, ageMin, is different from 23. The lower limit of ages to use for digit d is found to be as follows for approximation I:

$$Low(I,d) = ageMin + mod(int(ageMin/5)\times 5 - ageMin + d, 5) \tag{5.1}$$

The corresponding high values for the sums are simply a function of the number of 10-year groups of ages, T, to be used in the analysis:

$$High(I,d) = Low(I,d) + T\times 10 - 1 \tag{5.2}$$

The Low(II, d) and High(II, d) values are just one higher than the approximation I estimates:

$$Low(II,d) = Low(I,d) + 1 \tag{5.3}$$

$$High(II,d) = High(I,d) + 1 \tag{5.4}$$

The maximum age used in this analysis, ageMax, is ageMin + T*10 + 4.

If a maximum age, ageMax', is given (e.g., the highest single age for which there is data) then we can solve for (the maximum) value of T:

$$T = \text{int}((ageMax' - ageMin - 4)/10) \tag{5.5}$$

Combining these calculations, the percent reporting ages ending with digit d for approximation A (as a variable with values I or II) can be found to be:

$$Bpct(A,d) = \frac{\sum_{a\in([Low(A,d),High(A,d)]\wedge \text{mod}(a,10)=d)}^{High(A,d)} p(a)}{\sum_{a=Low(A,d)} p(a)} \tag{5.6}$$

Comparing the results of these calculations to the **PASEX** workbook **SINGAGE** and the initial version of the DemoTools **check_heaping_bachi** function (with the parameter pasex=TRUE) revealed that those two tools implemented the method slightly differently from how it is described above. Instead of computing

the two different distributions I and II and averaging them, they compute the average of the numerators divided by the average of the denominators. It turns out that this is equivalent to looking at the denominator as the sum of the centered 10-year moving average estimate of the population for each age in the numerator.

Further testing of how the Bachi formula performs in various situations showed that if the population under study is in fact linear over the range being analyzed, the Bachi method does not return exactly 10 percent for each digit. The **SINGAGE** and **check_heaping_bachi** tools, with the alternative method, did return equal portions of exactly 10 percent when the input population was linear. Indeed, it should always work for a linear population regardless of at what age you start at and how many 10-year age segments are used.

The **check_heaping_bachi** function has now replaced the pasex input with the choice of method="orig" or method="pasex". It is interesting that if the population is linear, the "pasex" method actually gives a value of zero, while the "orig" method gives a very small number. The "pasex" version is used in the following analysis.

Figure 5.4. shows the Bachi index values for the Philippines, a developing country with a large number of available censuses with single year age data. In the Philippines, significant improvement in the accuracy of age reporting (as measured by the Bachi index) is observed since the 1960s.

Figure 5.4. Bachi Index, by sex and census year for the Philippines, 1960 to 2015

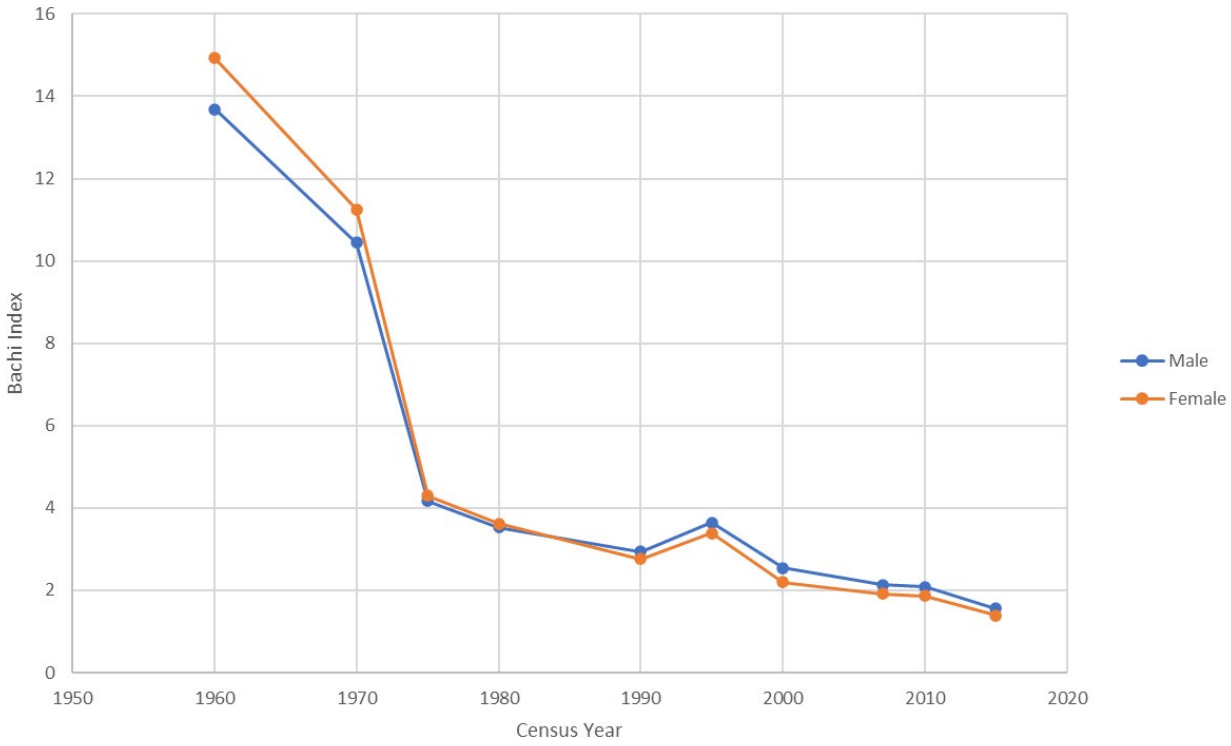


Table 5.1. presents a list of categories of age accuracy used in the United Nations Demographic Yearbook 1988. Added to this are the upper limits of equivalent levels of the Bachi index. According to this categorization, data quality in Philippines censuses went from 'Rough' in 1960 to 'Fairly Accurate' by 2015.

Table 5.1. UN DYB categorization of levels of heaping based on the Whipple and Bachi indices

			Upper limit values	
Category	***Whipple interval***	***Whipple***	***Bachi 0 = 5***	***Bachi 0 or 5 only***
Highly accurate	[0, 1.05)	1.05	1	1.125
Fairly accurate	[1.05, 1.1)	1.10	2	2.250
Approximate	[1.1, 1.25)	1.25	5	5.625
Rough	[1.25, 1.75)	1.75	15	16.875
Very rough	[1.75, 5)	5.00	80	90.000

Notes:
Whipple intervals from United Nations Demographic Yearbook 1988, p. 19.
Bachi 0 = 5 means that there is equal heaping on 0 and 5 only, and all other ages are distributed evenly.
Bachi 0 or 5 only means that only one of those ages has heaping and all other ages (including the non-heaped 5 or 0) are distributed evenly.

Two interesting observations can be made on figure 5.4.:

1. Females started with more age heaping, but by 1990 they were slightly lower than males.
2. The age heaping dropped fastest from 1970 to 1975, possibly resulting from a change in the age question, which seems to have changed from just asking the age in 1970 to asking year and month of birth as well as age in 1975 (Philippines National Census and Statistics Office, 1978). According to IPUMS (2021), from 1980 to 2010, questions on age and month and year of birth were included in the census questionnaire.

The detailed Bachi percent deviations for females show strong heaping on 0s and 5s from 1960-1970 (see figure 5.5. and figure 5.6.). In addition, the heaping was higher on 0s than 5s. This can also be seen from the percent distribution. The ratio of the percent reporting ages ending in 0s to those ending in 5s indicates that in census years ending in 5, the relative heaping of 0s is reduced. We can see that if we look at the percentages directly. The changes seemed to be lowering the per cent 0s in 1975, but raising the per cent 5s in 1995 and 2015.

Figure 5.5. Ratio of Bachi percent 0 / percent 5: Philippines females

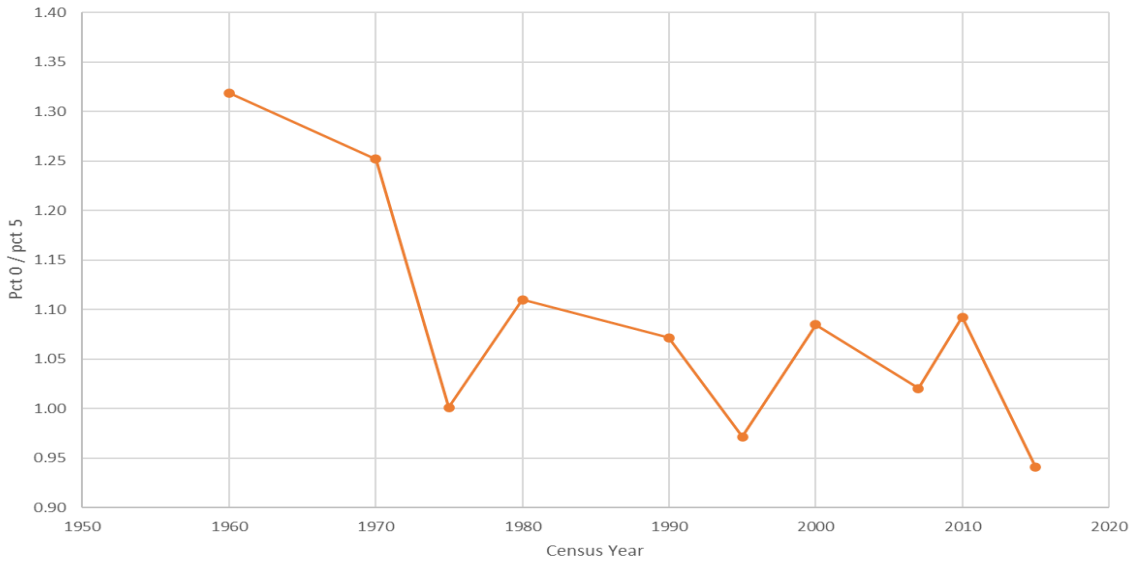
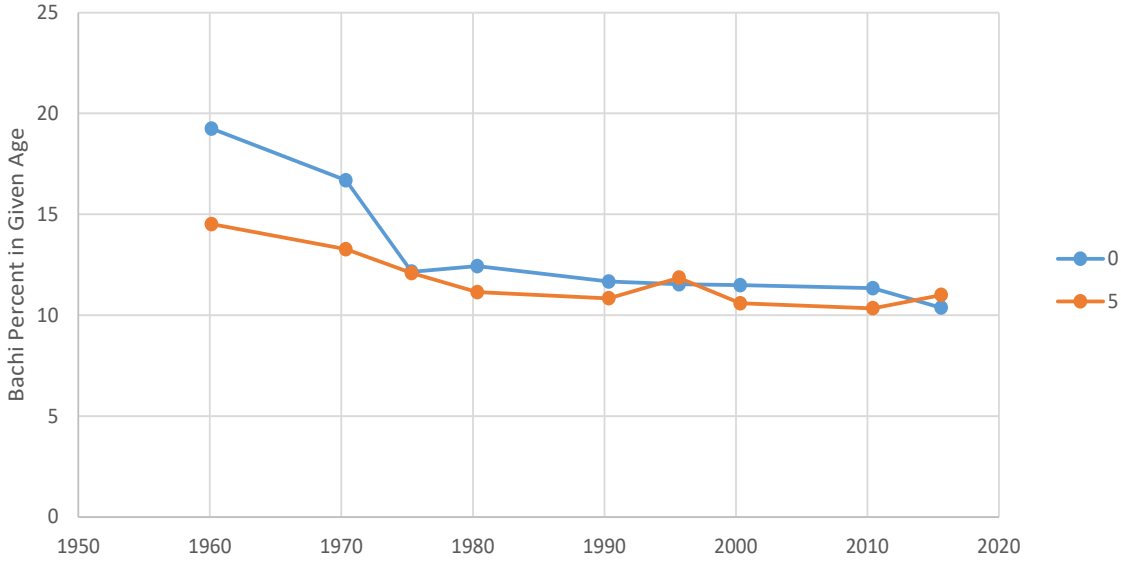


Figure 5.6. Bachi percent 0 and percent 5: Philippines females



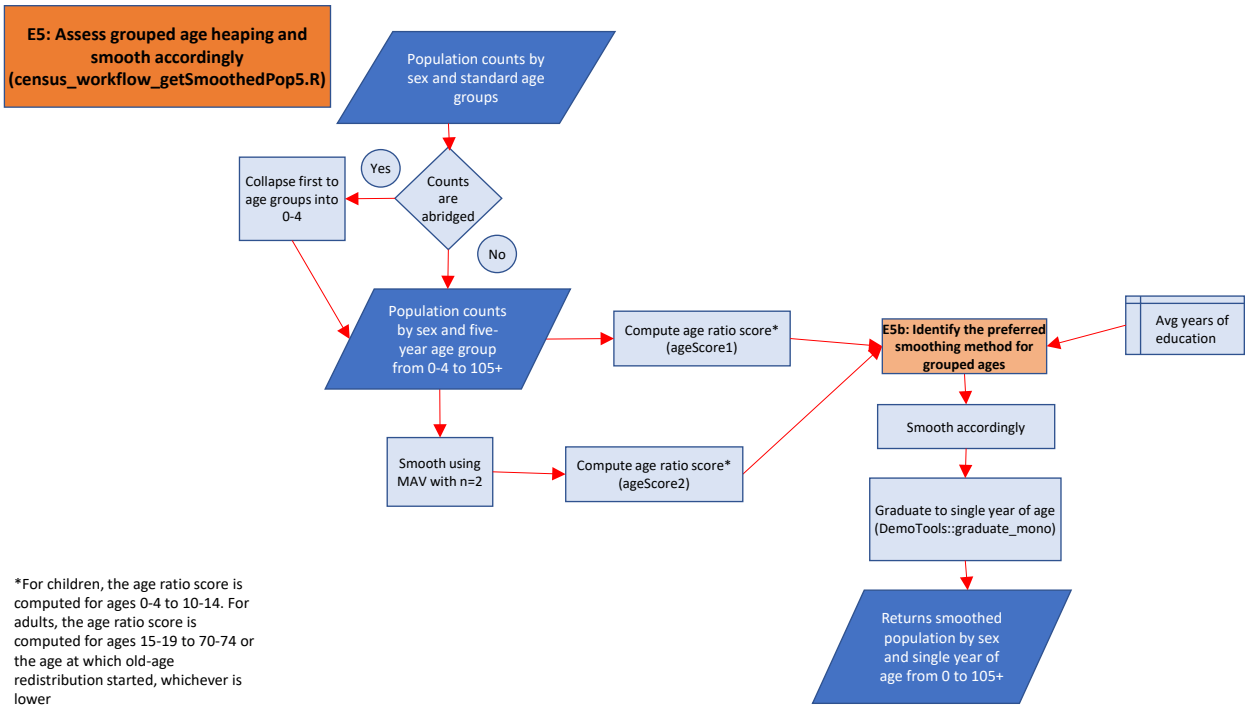
2. *Age ratio score*

When population data by single age are not available, or single age data are too deficient (i.e., Bachi levels of 30 and higher), data by five-year age groups are used and the age ratio score is applied to assess the quality of the age distribution (see figure 5.7.). The age ratio score is a component of the UN age-sex accuracy index (1952) used to measure the quality of population data classified into 5-year age groups, ${}_5P_x$. It is computed as the sum of the absolute differences between the reported age ratios and the expected value of 1.0.

$$ARS = \sum_{x=l}^{u} \left| \left[\frac{{}_5P_{x-1} + {}_5P_{x+1}}{2\,{}_5P_x} \right] - 1 \right| \quad (5.7)$$

As with the measures of quality for population by single years of age, the age ratio score (or the full UN age-sex accuracy index) is assuming that the population by age (and sex) has evolved rather smoothly over time.

Figure 5.7. Workflow to assess grouped age heaping and smoothing method



B. SMOOTHING METHODS

Once the quality of an age distribution has been assessed and it has been determined that the age distribution needs to be corrected for the age heaping, the data need to be smoothed.

Depending on the availability of the age distribution (i.e., single year of age or five-year age group), different smoothing methods are applied.

1. Smoothing of populations by single years of age

One of the simplest ways of smoothing a population is to use a moving average of the population by age. The moving average estimates, or mav, can be identified by N, the number of ages used in the smoothing, summarized as mavN. The moving average can be applied to the population by single years of age or the population in 5-year age groups.

For odd values of N, the formula is just the average of the population, p(x), over the ages (N-1)/2 ages on each side plus the age of interest, x.

$$mavN(x) = \frac{1}{N} \sum_{a=x-(N-1)/2}^{x+(N-1)/2} p(a) \tag{5.8}$$

For even values of n, a similar approach is used, summing over the age range x +/- (N/2 -1) ages plus half of the populations +/- (N/2) ages away. This is equivalent to:

$$mavN(x) = \frac{1}{2N} \left[\sum_{a=x-N/2}^{x+N/2-1} p(a) + \sum_{a=x-N/2+1}^{x+N/2} p(a) \right] \tag{5.9}$$

When used for N=2 to 10, the method produces a generally increasing degree of smoothing, but for some values of N, there seem to be new spikes created. Note that mav1 is equivalent to the original population.

Figure 5.8. illustrates the general smoothing procedure. The degree of smoothing is determined by the value of the proportion of heaping on 0 and 5 (prop05), the proportion of heaping on the favourite two digits (prop2), and the average number of years of education (Edu yrs).

Figure 5.8. General workflow procedure for determining the degree of smoothing for single year ages

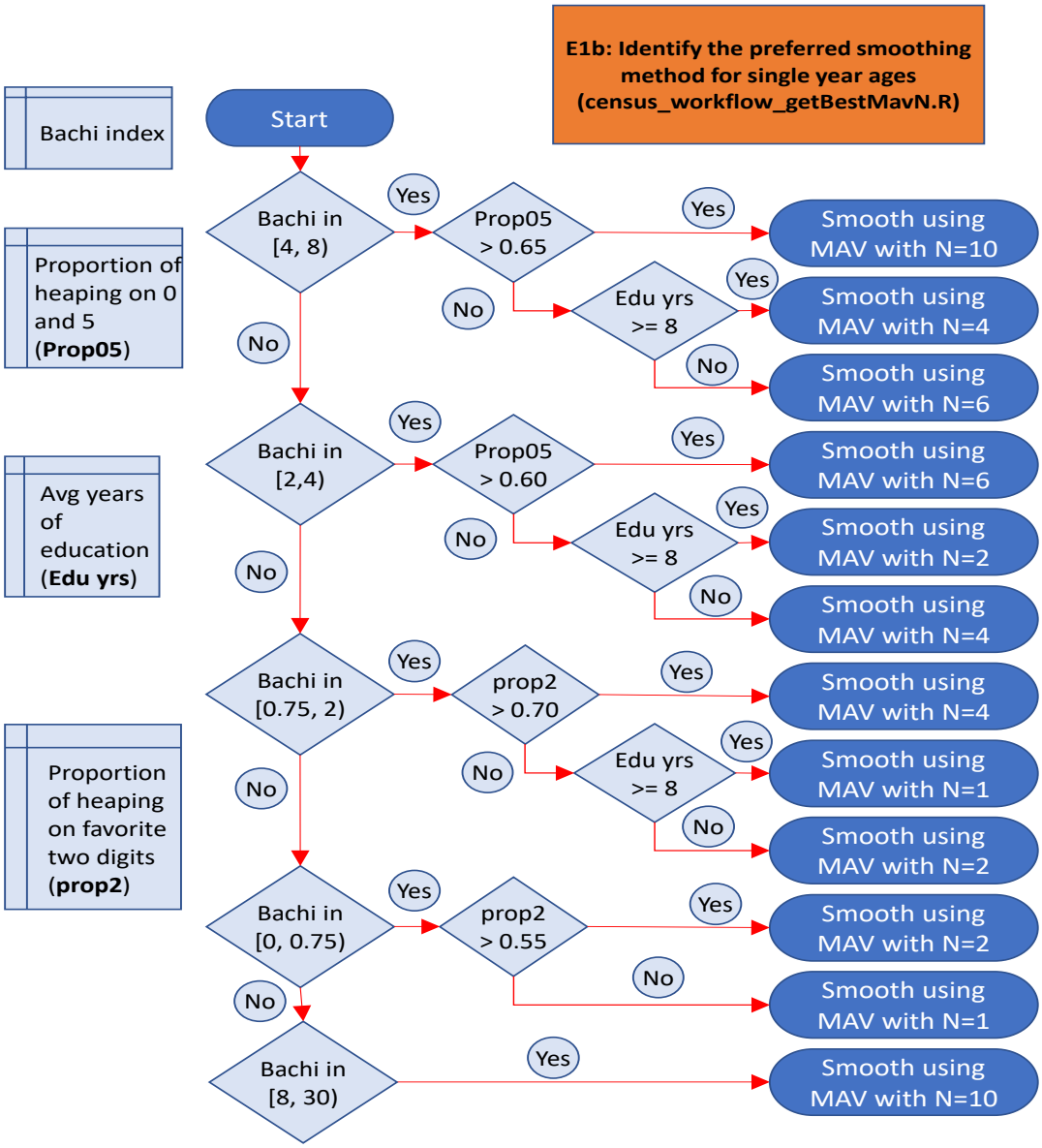
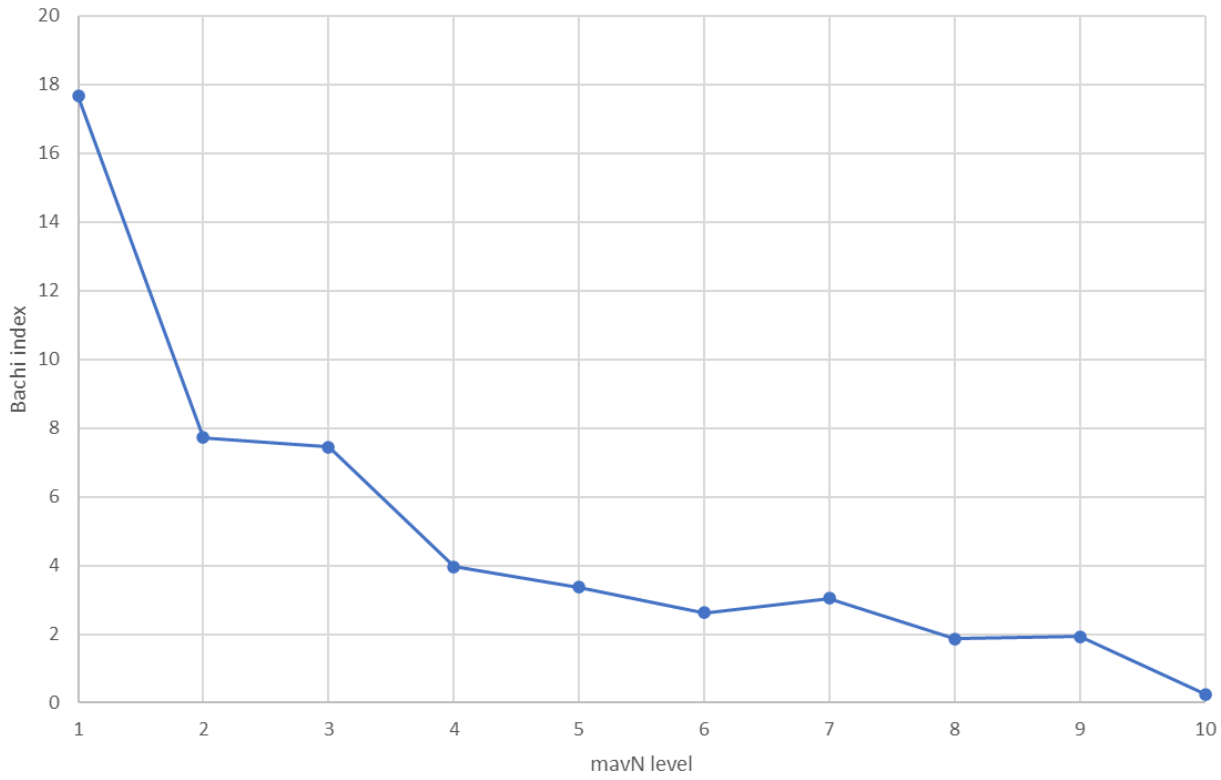


Figure 5.9. below provides an illustration of the Bachi index for the male population distribution by single age enumerated in the 2003 census of Haiti. The Bachi index value is 17.75, meaning that a minimum of about 17.75 per cent of the population (in the range 23-77) misreported their ages. According to Table 5.1. (above), such index corresponds to a "rough" age distribution. Figure 5.9. and table 5.2. indicates the value of the Bachi index corresponding to different levels of moving average. In the case of male population distribution by single year of age in Haiti in 2003, the use of mav2 brings the Bachi index under a value of 8. Higher levels of moving average contribute to smooth further the age distribution and reduce therefore the Bachi index until a minimum for mav10.

Figure 5.9. Bachi index by mavN level, Haiti 2003 census, males



The reason the Bachi index is relatively flat or rises when going from an even mavN to an odd value is that, as noted elsewhere, the odd mavN values equally weight the surrounding ages. This means that for mav3, digit 3 it is giving equal weight to digits 2 and 4, so twice the weight to even ages as odd. For this reason, it is probably generally best to use only even mavN values.

In some cases, the mavN smoothing can cause very strange patterns of "age heaping" that often result from the extra heaping on 0s vs other ages (even 5s). This means that for mav8 any digit that was smoothed including data for ages ending in 0 will tend to show "heaping" and others will show "avoidance." With mav8 that means that digit 5 does not include digit 0, so will probably show avoidance. Digits 4 and 6 include the population age 0 at half weight, so they may imply lower avoidance, and the remaining digits (0, 1, 2, 7, 8, 9) include 0 at full weight so they may show no significant deviation.

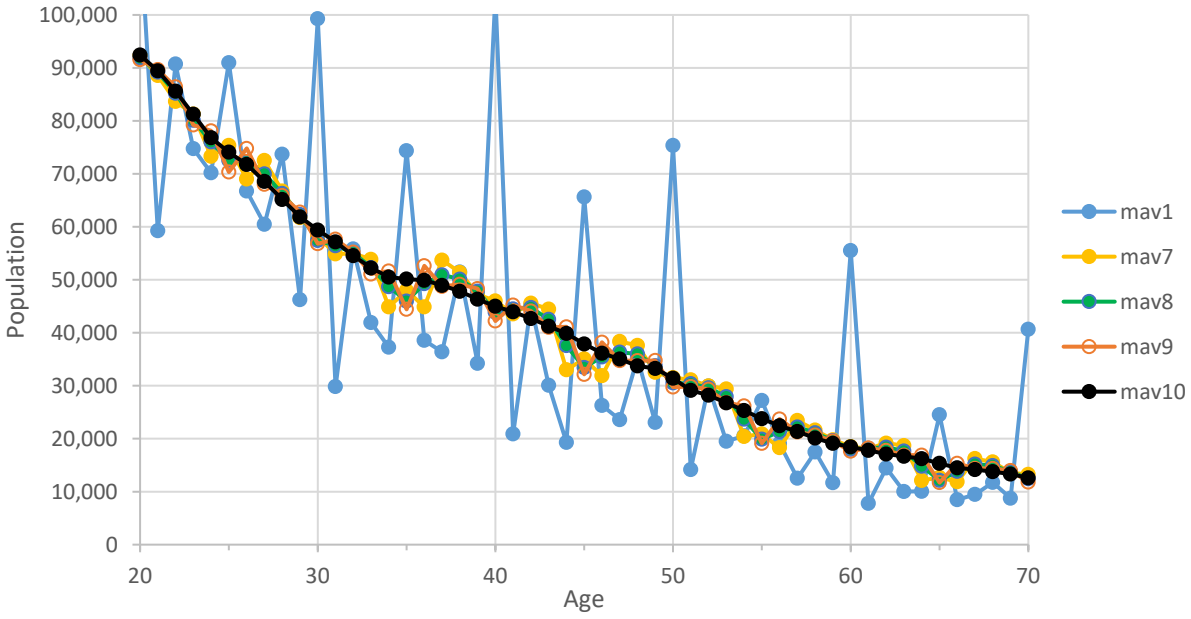
Figure 5.9. shows that both mav8 and mav9 have low values (anti-heaping) on ages ending in 5. This is because in Haiti there was stronger heaping on 0s than 5s, so with *mav8* and *mav9* the estimates for ages ending in 5 use all of the ages except 0.

Table 5.2. Selection of the best mavN level for smoothing single year population data

max (Bachi)	*Initial mavN*	*min (BachiProp0and5)*	*min (Max2prop)*	*min (Educ)*	*Final mavN*	*Notes*
[0,0.75)	1		> 0.55		2	
			<= 0.55		1	
[0.75,2)	2		> 0.70		4	
			<= 0.70	>= 8	1	
					2	
[2,4)	4	> 0.60			6	
		<= 0.60			2	
		<= 0.60			4	
[4,8)	6	> 0.65			10	
		<= 0.65		>= 8	4	
		<= 0.65		< 8	6	
[8,30)	10				10	
[30,Inf)	NA				NA	Use best Grad5

Notes:
- max and min refer to the max or min of the indicator among the two sexes;
- BachiProp0and5 is the sum of the Bachi percents for 0 and 5 minus 20 divided by the Bachi Index. Note that if there is avoidance of 0 and/or 5 the value could be negative;
- Max2prop is the proportion of the Bachi Index accounted for by the two digits with the highest percents. If the top two percents are 0 and 5 then this will equal BachiProp0and5;
- Educ is the average number of years of education by sex for the country and year based on the GBD modelling results.
- Grad5 is the level of smoothing for data by 5-year age groups (Grad5) to be graduated into single years of age as defined in Table 5.4.

Figure 5.10. Moving average smoothing of Haiti male census population, 2003



2. *Smoothing of populations in 5-year age groups*

Where population data are available only by 5-year age group or the quality of the data by single age is too low, they are evaluated and eventually smoothed by other methods (see figure 5.7.) available in DemoTools in particular the **smooth_age_5** method.

Some of the smooth_age_5 methods only smooth or redistribute the population within 10-year age groups, starting with age 0-9. As noted above, the 0-4 population is often under-counted, so it makes some sense to not use it in the smoothing process. Also, if the thought is that people are often over-stating their ages (e.g., rounding from, say 39 to 40), then we would want to allow redistribution from 40-44 back to 35-39.

Some preliminary tests indicated the following ranking of the methods (as measured by age ratio scores):

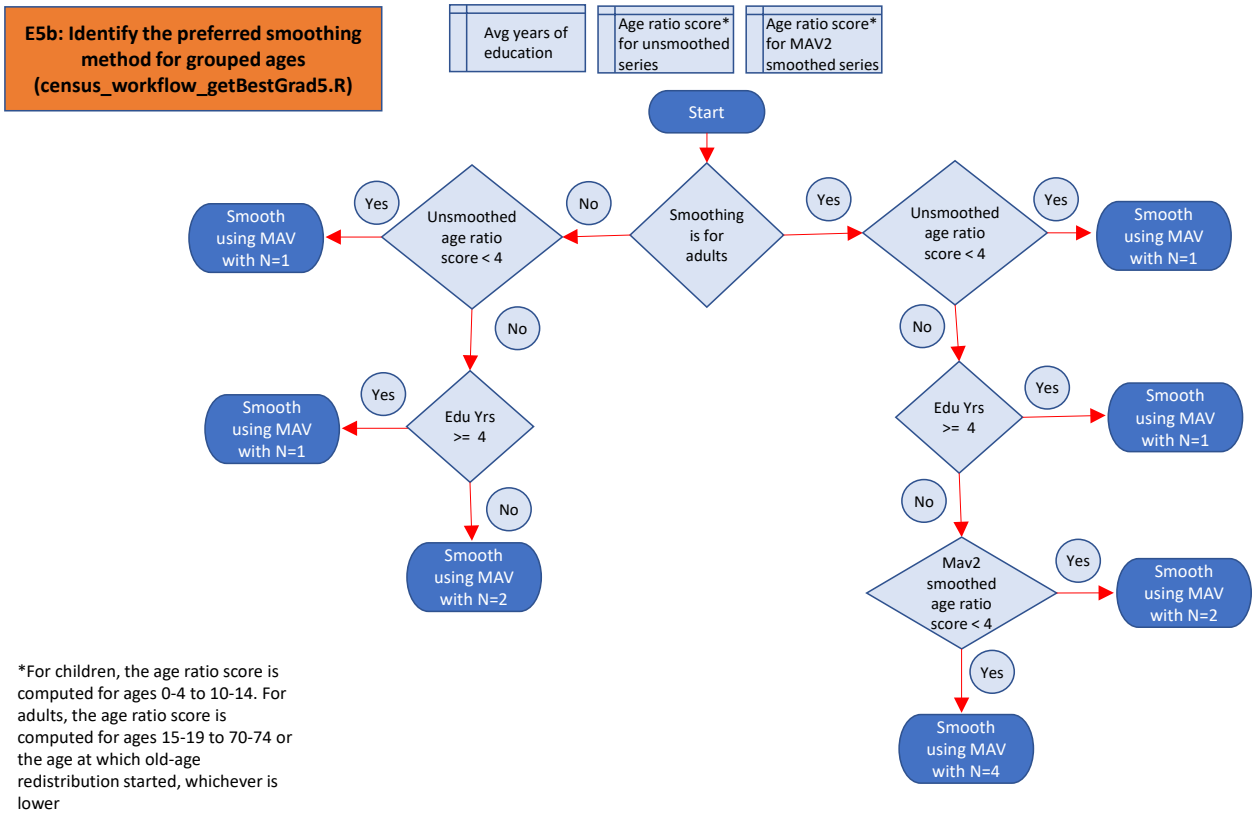
- The observed data will be the least smooth;
- Second, third, and fourth which tend to be close together are the Carrier-Farrag (1959), Karup-King-Newton (Carrier and Farrag, 1959), and Zigzag (Feeney, 2013);
- Fifth, but close, is the Arriaga (1968) method;
- Sixth is the United Nations method (Carrier and Farrag, 1959);
- Seventh is consistently the mav2;
- Eighth and ninth are always mav4 and strong.

Of the minimal smoothing options, the Arriaga method is preferred because it has a separate procedure for the youngest and oldest 10-year age groups.

The moving average methods (mav2 and mav4) are symmetrical around the estimated age group. Both mav2 and mav4 equally weight age groups starting with last digits 0 and 5, so if there is more heaping on 0s both measures will not overly contribute to any of the estimates.

Figure 5.11. summarizes the various steps and criteria selected to apply a smoothing method for data by five-year age group, and complements figure 5.7.

Figure 5.11. Workflow to identify the preferred smoothing method for grouped ages



Due to the rather different types of biases potentially affecting the quality of the reporting of young children compared to adult ages, the analysis and smoothing methods are applied separately for adult and child populations.

C. *ADULT POPULATION SMOOTHING*

1. Smoothing of adult populations by single years of age

It is expected that census populations by single years of age can reveal more about the true age patterns than the populations grouped into 5-year age groups. There are several possible smoothing methods that could be used, but the moving average is simple and easy to apply.

The current procedure uses the Bachi index, the proportion of heaping on 0 and 5, the proportion of heaping on the favourite two digits, and the average level of education to select the *mavN* level (i.e., degree of smoothing). The current mapping from the Bachi index and other measures to determine the *mavN* level is shown in Table 5.3.

The maximum Bachi level of the two sexes for the same census, is used to keep the level of smoothing the same for the two sexes. Since at higher levels of the Bachi index there is more likelihood that there is significant heaping on 0s and 5s, it is better not to not use mav8 (see the example above for Haiti). This is because mav10 uses the ages from 5 years below to 5 years above and balances the number of digit 0 and digit 5 populations (as well as the number of evens and odds, but that is true for all even mavNs).

For Bachi levels of 30 and higher, the quality of the single age distribution is too problematic, and the specific information that can be drawn from the single ages is probably minimal. In such cases, the data by five-year age groups are used and the level of smoothing needed is derived based on the age ratio scores (see section 5.2.3). The smoothed 5-year age groups are then graduated into single years of age (see section 5.5 and Table 5.4).

D. GRADUATION OF ADULT POPULATIONS IN 5-YEAR AGE GROUPS

Graduation of population data in 5-year age groups is another way of smoothing the population distribution. Graduation is the process of fitting a function to the 5-year data and then using that function to estimate the population by single years of age. Graduation has been done for many years by actuaries creating life tables for use in determining life insurance rates using osculatory interpolation methods, such as those developed by Beers (1945) and Sprague (1880). These were easy to apply prior to the use of computers. More recently, these have been replaced by spline methods that are more flexible.

To graduate a population in 5-year age groups (or abridged ages) into single years of age, the DemoTools function in R, **graduate_mono,** was used. The program separately looks at the available 5-year only data and does the smoothing and graduating for censuses that did not have single year data.

The original implementation of the procedure graduate_mono in DemoTools (with spline method="monoH.FC") produced estimates where the population approaching age 0 from above had a slope near zero at age 0. The graduate_beers procedure (especially with the Johnson adjustment if there is an estimate for age 0) does not seem to do this, so is probably preferred for smoothing the child population. The **graduate_mono** function in DemoTools has been updated to use the "hyman" method which produces results similar to the Beers/Johnson method. The graduation method is applied on the transformed cumulative distribution of the population by age using some monotonicity constraint to prevent any implausible negative values, and the graduated values are obtained using the decumulated values interpolated to the relevant standard ages. This procedure is also applied to all the single age censuses (single-age data is returned as-is) along with the different *mavN* smoothing of that data.

This procedure is done for all the single age censuses along with the different mavN smoothing of that data. The program separately looks at the available 5-year only data and does the smoothing and graduating for censuses that did not have single year data.

1. *Smoothing of adult populations by 5-year age groups*

Table 5.3 summarizes characteristics of different smoothing methods for data, including the number of original ages used and whether (some of) the original totals are retained.

Table 5.3. Number of ages used to get estimate for a given age for different smoothing methods

Method	*Original ages used*	*Keeps 10-year population*	*Notes*
Arriaga	30	Yes	Formulas for 0-4 and 5-9
Carrier-Farrag	30	Yes	
Karup-King-Newton	30	Yes	
Strong	30	No	AGESMTH adjusts total 10-69 to agree with observed and Arriaga for youngest and oldest
United Nations	25	No	is there a formula for 0-4 and 5-9
5Mav4-5	25	No	Not in AGESMTH
5Mav2-3	15	No	Not in AGESMTH
1Mav10	11	No	Requires single year data
1Mav8-9	9	No	Requires single year data
1Mav6-7	7	No	Requires single year data
1Mav4-5	5	No	Requires single year data
1Mav2-3	3	No	Requires single year data

The mapping from the age ratio score (again, the maximum of the two sexes) and the level of education to the mavN level is given in Table 5.4.

Table 5.4. Determination of the best level of smoothing for data by 5-year age groups (Grad5) to be graduated into single years of age

max(ageScore1)	min(Educ)	max(ageScore2)	Final level of smoothing for Grad5
[0,4)			1
[4, Inf)	>= 4		1
[4, Inf)	< 4	[0,4)	2
[4, Inf)	< 4	[4, Inf)	4

Notes: - **max** and **min** refer to the max or min of the indicator among the two sexes;
- **ageScore1** is the age ratio score (based on the observed population in 5-year age groups adjusted for net census error);
- **ageScore2** is the age ratio score (based on mav2 smoothing of the observed population in 5-year age groups adjusted for net census error);
- **Educ** is the average number of years of education (by sex) for the country and year based on the GBD modeling results.

E. CHILD POPULATION SMOOTHING

1. *Child age misreporting patterns*

The patterns of age misreporting for children (age 0-17 years) are quite different than those for adults. There are definite patterns that seem to be relatively consistent by country and common across countries (often in the same region).

The overall pattern of child age reporting can be examined by looking at the ratio of the population at each age to another age, for example age 0. Figure 5.12. shows the average pattern of relative population

by age for censuses with single year data in the least developed countries for which census age distribution was available.

Figure 5.12. Average size of reported census populations by age relative to age 0 for least developed countries

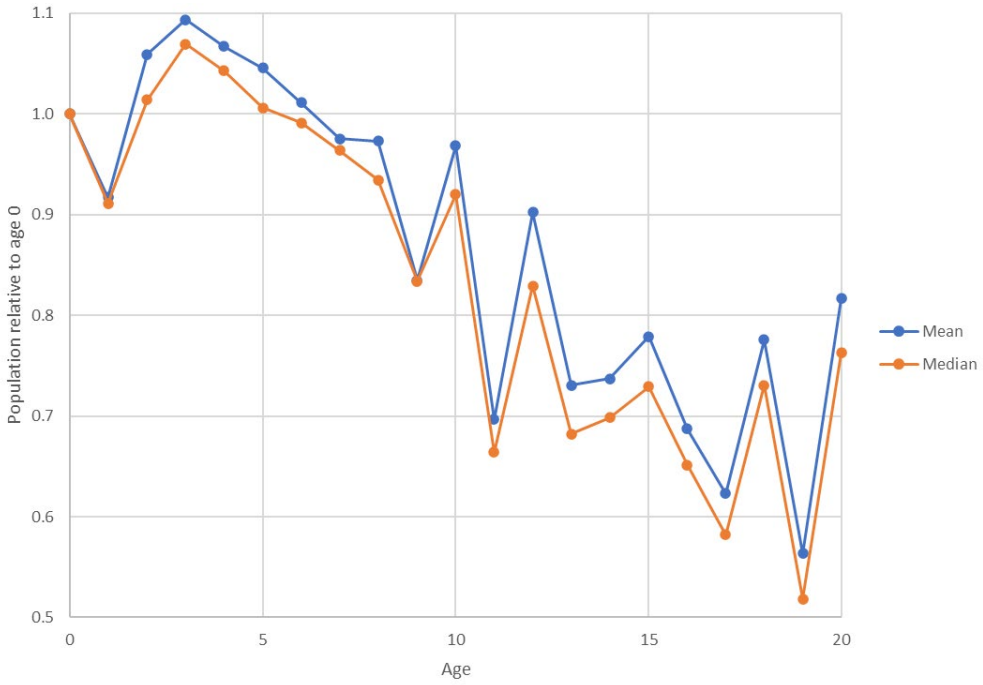


Figure 5.12. indicates that the peak child population tends to be at age 3, with the minimum being for age 1. The figure also shows the tendency toward heaping on even digits, starting at age 8, but with a transition to more preference for age 15 (for example, age 14 tends to be slightly lower than age 15).

Children are often omitted from censuses for various reasons (Ewbank, 1981). The lower population for age 1, even lower than age 0, is hard to understand, and needs more research. It is possible that some age fields that were left blank were read as zero, inflating that population. It could also be the result of misreporting children age 1 to age 0. Age 0 would therefore be artificially inflated by the (down) transfer from age 1 and deflated by children who are not reported.

2. *Child population smoothing procedures*

The smoothing based on the adult Bachi index (based on ages 23-77) or the age ratio score based on 5-year data (using ages 15-19 to 70-74) tends to be too extreme for the child estimates. A separate smoothing procedure was developed for the children based on the child Bachi index (ages 3-17). The conversion table 5.5. from child Bachi level to mavN level is the same as for adults as shown in Table 5.4.

The smoothed child population is then blended with the adult smoothed populations over the age range 15-25 assuming a linear weighting assumption. The final population is therefore obtained as:

BestSmthAdj(x) = BestSmthChild(x) for x <= 15

BestSmthAdj(16) = 0.8*BestSmthChild(16) + 0.2*BestSmthAdult(16)

BestSmthAdj(19) = 0.2*BestSmthChild(19) + 0.8*BestSmthAdult(19)

BestSmthAdj(x) = BestSmthAdult(x) for x >= 20

When 5-year age distributions are being smoothed (whether due to lack of single year data or the need for stronger smoothing), the smoothing is done based on the child age ratio score (using ages 0-4 to 20-24) according to Table 5.5.

Table 5.5. Mapping of child age score to best Grad5 mavN level

max(ageScore1)	Min(Educ)	Final Best Grad5
[0,4)		1
[4, Inf)	>= 4	1
[4, Inf)	< 4	2

The maximum smoothing of the child 5-year data is mav2.

Smoothed populations (especially based on 5-year age groups) that include relatively undercounted populations under age 5 (or maybe just under 3) can cause an artificial bulge in the 5-9 populations, which would need to be addressed later on.

F. POST-SMOOTHING ADJUSTMENT

When the smoothed adult and children's populations are blended the population that is obtained does not match the total population figure. The smoothed population is therefore prorated at all ages to get a smoothed population that sums to the adjusted totals by sex.

VI. CORRECTION OF UNDER-ENUMERATION OF YOUNG CHILDREN AND INTEGRATION WITH SMOOTHED CHILD AND ADULT POPULATION ESTIMATES

As noted above (see figure 5.12.), the population at younger ages is often presenting a pattern of under-enumeration (Ewbank, 1981). The observed population of children at age 0, 1, and/or 2 are lower than expected, in particular in the developing countries with high fertility, where a continuous increase in the population is expected at younger ages. But patterns of under-enumeration of young children is found in more statistically advanced countries as well. Some of these under-enumeration problems also affect the post-enumerations surveys because the information provided on the household composition, and children in particular, is provided by the same proxy respondent[10].

Patterns of child under-enumeration can be corrected through demographic analysis by estimating the population at ages 0, 1-4 and 5-9 based on the smoothed adult female population from a census and estimates of fertility and mortality for the various periods preceding the census. In other words, the population below age 10 can be reconstructed based on estimates of past fertility and mortality. Such procedure is available in the **PAS** workbooks **BASEPOP** and **BPSTRNG** (U.S. Census Bureau, 2019; Arriaga, and others, 1994).

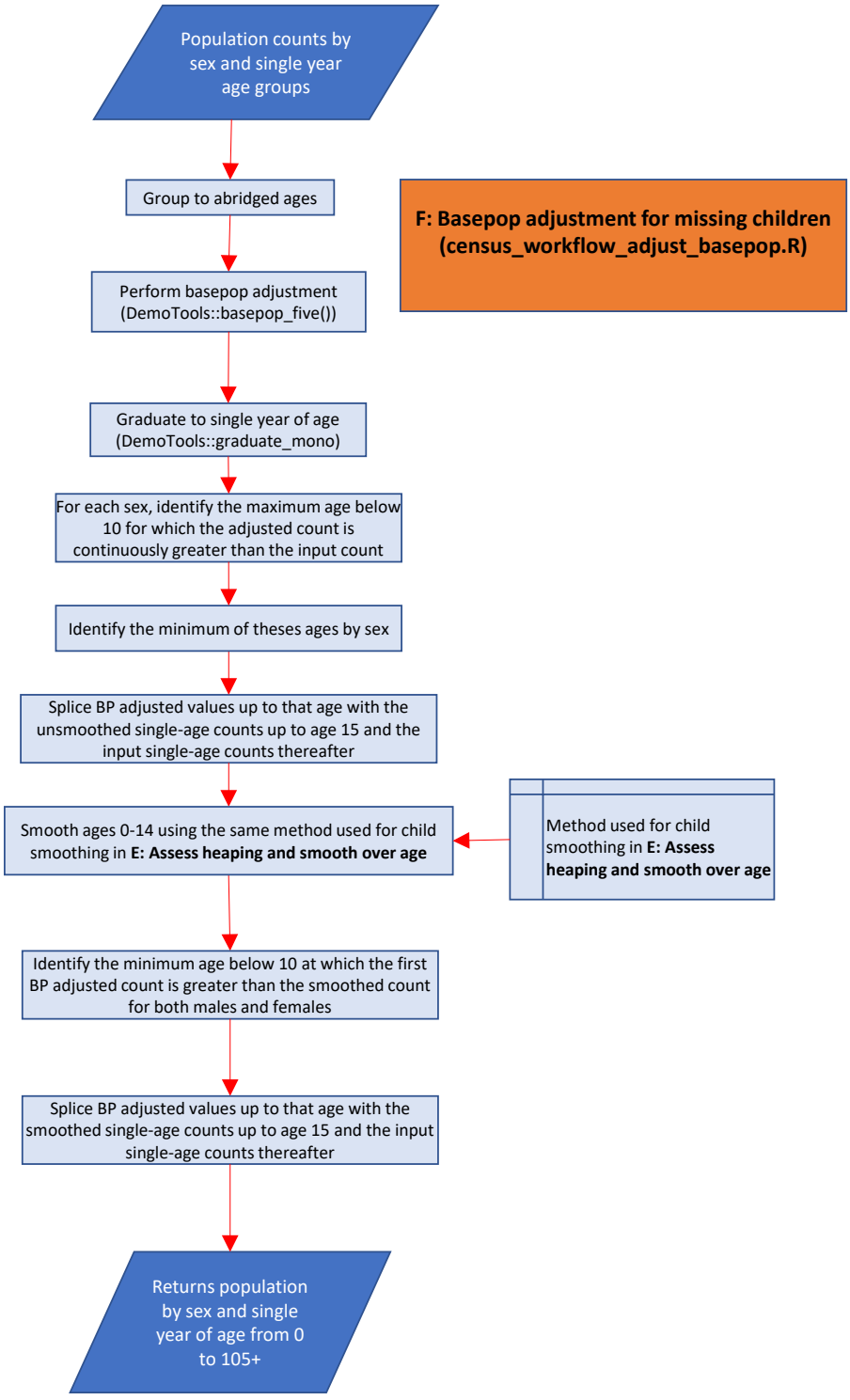
A new version of these initial methods was created in order to address some problems with the initial versions. The **NewPAS** workbook **BPA** (U.S. Census Bureau, no date) is similar to **BASEPOP**, but uses the Arriaga end formula to get the smoothed population 10-14 and 15-19 in order to prevent a circular reference. The workbook **BPE** allows the user to enter the population by age, and that can be either the observed population or a population separately smoothed (e.g., using the **AGESMTH PAS** workbook).

The **BPE** method has been included in **DemoTools** in R as function **basepop_five**. The method was applied to the adjusted and smoothed population by single years of age aggregated into 5-year age groups and using fertility and mortality estimates from the previous revision of the WPP, the 2019 revision. Note that since the WPP estimates were only done in 5-year intervals, estimates prior to 1953 were held constant for earlier years, which will affect the adjusted estimates for censuses from about 1950 to 1959. Future WPP revisions will use instead new annual time series estimated since 1950, with single-year and single-age data for countries with sufficiently reliable single age distributions.

The BASEPOP procedure consists of reconstructing the population aged 0-10 using consistent estimates of fertility and mortality. The reconstructed or adjusted population present birth cohorts that are consistent with recent fertility and mortality levels and patterns. The adjusted population can later be compared to the population enumerated in the census allowing identification of inconsistencies in the enumeration of the children in the census.

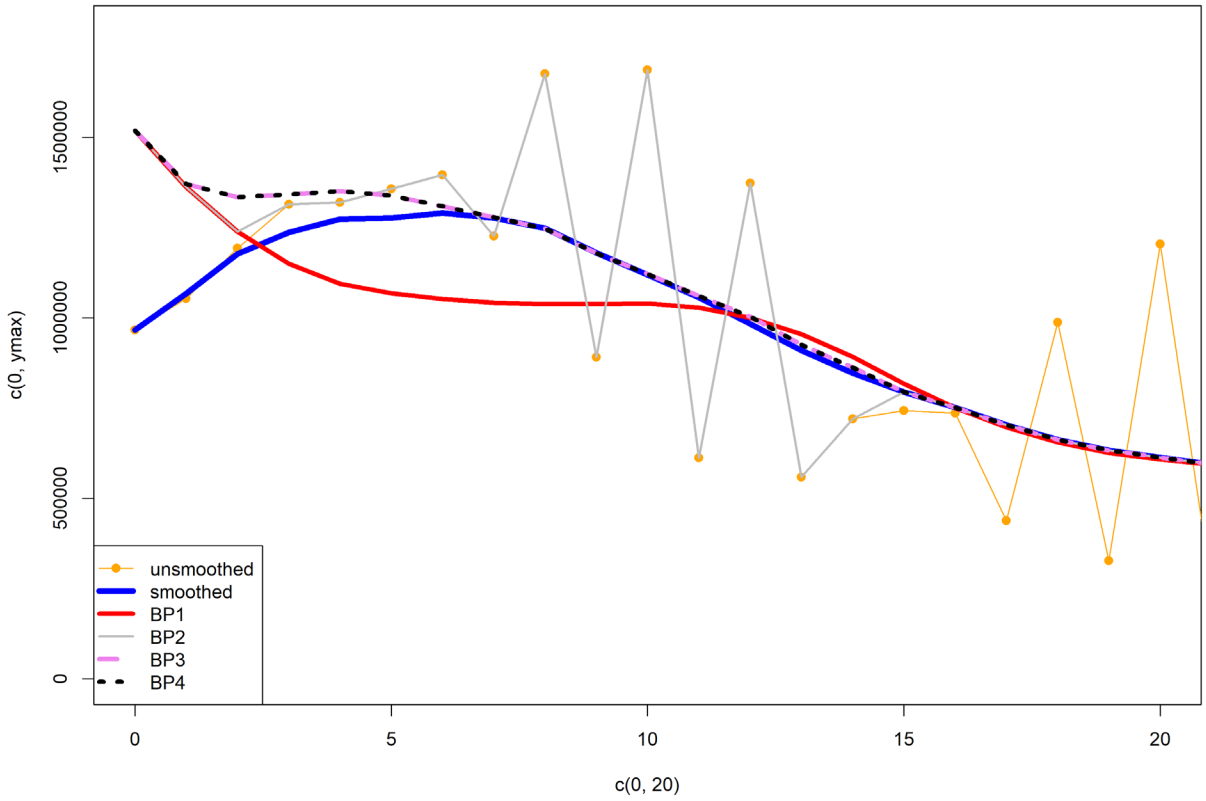
[10] For example, see U.S. Census Bureau (2022), "Despite Efforts, Census Undercount of Young Children Persists", available online (last accessed 18 October 2022): https://census.gov/library/stories/2022/03/despite-efforts-census-undercount-of-young-children-persists.html.

Figure 6.1. BASEPOP procedure to adjust for missing children



The adjusted population estimates are given by sex and abridged age groups for ages 0, 1-4, and 5-9. These estimates are then merged with the smoothed populations and graduated using the graduate_mono function to get estimates by single years of age. This estimate is called BP1 (see figure 6.2). In this example, the BP1 estimate (red line) starts higher at younger ages, but crosses over the smoothed estimate (blue line) after age 2.

Figure 6.2. BASEPOP (BP) Estimates for female population, Bangladesh, 1974 census



In some cases, the BP estimate can cross over the smoothed estimates at different ages for each sex. This sometimes happens when the BP and smoothed curves get close for a few ages and then diverge again. If for one sex the BP dips below the smoothed series then the rest of the BP results are not used for that sex, but if the BP is always higher for the other sex, the BP results 0-9 are used for that sec. This happens for Algeria 1954 (see figures 6.3 and 6.4). The solution is to look at the last age where the BP estimates is greater than the smoothed population, and pick the minimum of this age for males and females to create a combination of the BP estimates for young ages with the smoothed population.

Figure 6.3. Female population adjusted with BASEPOP procedure, Algeria, 1954 census

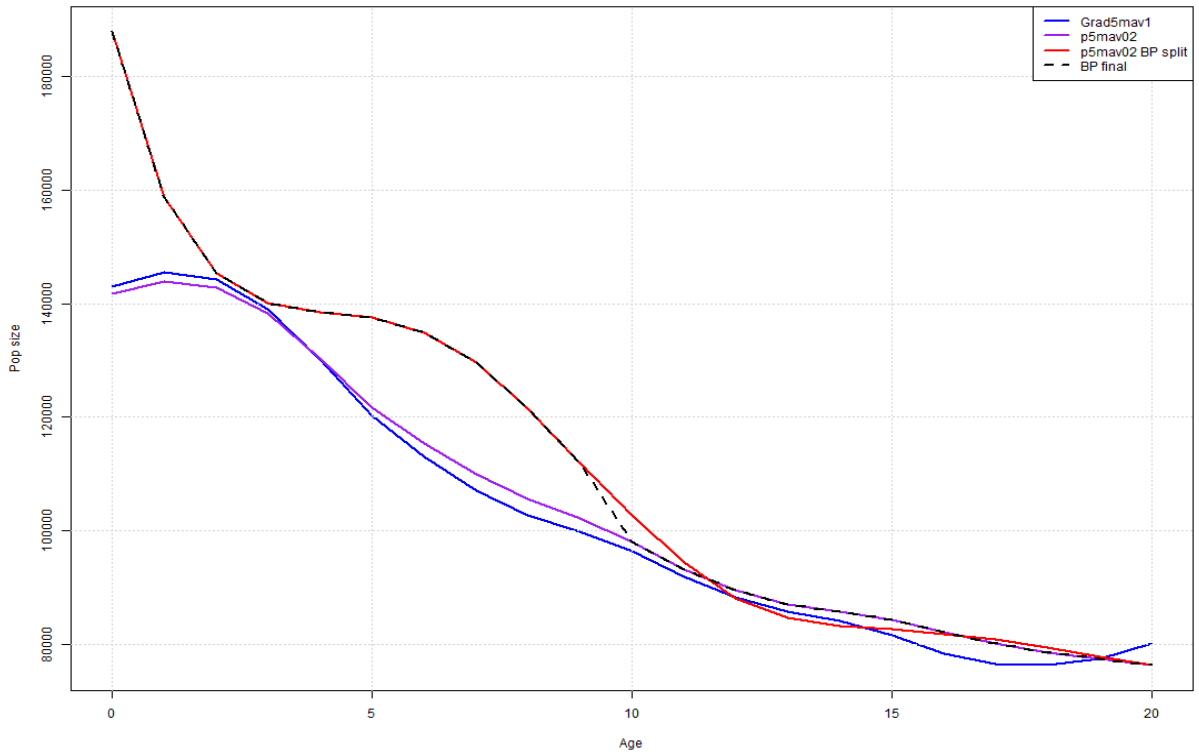
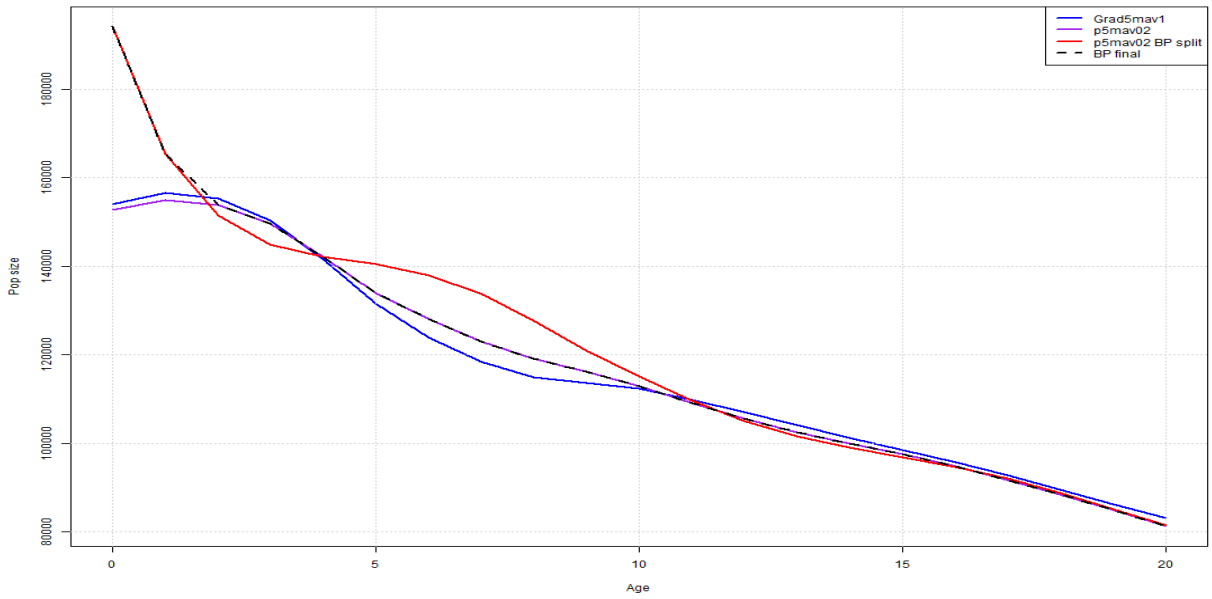


Figure 6.4. Male population adjusted by BASEPOP procedure, Algeria, 1954 census



The BP and smoothed results sometimes do not fit together very well, creating a dip around the point where the BP transitions to the smoothed estimates. Some of this may be the result of the inflated populations 5-9 due to the lower population ages 0-4, caused by the undercount of young children. In other cases it could be because the fertility levels used to get the BP results were too low.

The initial method is to compare BP1 (i.e. the graduated population age 0-9 based on the reconstructed population age 0, 1-4 and 5-9) to the smoothed adjusted pop to get the last age where BP1 is higher the smoothed estimates for each sex, and get the minimum of those two values, called minLastBPage1.

The next step is to combine the BP1 values for ages <= minLastBPage1 with the original (unsmoothed) adjusted population, to redo the smoothing of the child population without the highly under-reported population at the youngest ages.

Let **BP2** = **BP1** for age <= **minLastBPage1** then switch to **Pop1** (if available) or **Grad5mav1** up to age 15, then **BestSmthAdj**. This process is summarized in Table 6.1.

Table 6.1. BP2 formulas

BP2(x)=	***x value***	***Data available***
BP1(x)	x <= minLastBPage1	
Pop1(x)	minLastBPage1 < x <= 15	Single ages
Grad5Mav1(x)	minLastBPage1 < x <= 15	5-year ages
BestSmthAdj(x)	x > 15	

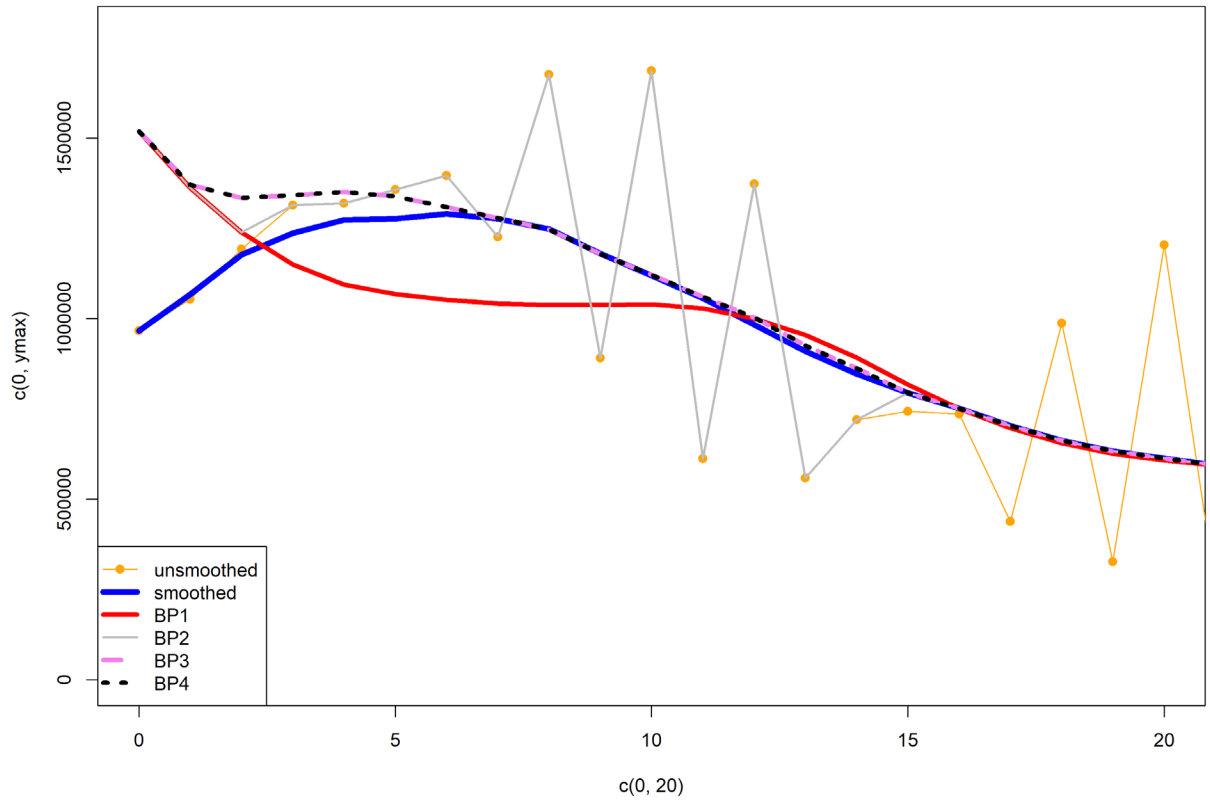
After computing **BP2**, the child smoothing process on the **BP2** results is repeated, but only for ages 0-14, using the original child Bachi/ageRatioScore criteria, calling this BP3. Age 9 would only affect up to age 14 at the highest (when using **mav10** or **Grad5mav2**).

The comparison of **BP1** is repeated, but this time to the **BP3** results to get **minLastBPage3**, trying to fix the problem of undercounted populations at the youngest ages causing a bump in the 5-9 age group. The **BP3** smoothed population may have a lower population aged 5-9 that blends better with the BP results. In most cases **minLastBPage3** should be greater than or equal to **minLastBPage1**.

Finally, **BP1** is combined up to age **minLastBPage3** and then switch to **BP3**, and is called **BP4**.

Figure 6.5 shows the various steps for the BASEPOP estimates for females from the 1974 census of Bangladesh.

Figure 6.5. BASEPOP (BP) female estimates, Bangladesh, 1974 census



The current steps for the BP procedure are as follow:

1. Get the BP estimates of the population 0, 1-4, and 5-9 using the female **BestSmthAdj** population.
2. Graduate results using **graduate_mono** and call this **BP1**.
3. Do the comparison of **BP1** to the smoothed adjusted pop (**BestSmthAdj**) to get the last age where **BP1** > **BestSmthAdj** for each sex and getting the minimum of those two values, **minLastBPage1**.
4. Let **BP2** = **BP1** for age <= **minLastBPage1** then switch to **Pop1** (if available) or **Grad5mav1** up to age 15, then **BestSmthAdj**.
5. Repeat the child smoothing process but only for ages 0-14, using the original child Bachi/ageRatioScore criteria on the **BP2** results, calling this **BP3**. Age 9 would only affect up to age 14 at the highest (when using **mav10** or **Grad5mav2**).
6. Repeat the comparison of **BP1** to the **BP3** results to get revised **minLastBPage3**. This is trying to fix the problem of undercounted populations at the youngest ages causing a bump in the 5-9 age group. The **BP3** smoothed population may have a lower pop 5-9 that blends better with the BP results. In most cases **minLastBPage3** should be greater than or equal to **minLastBPage1**.
7. Combine **BP1** up to age **minLastBPage3** then switch to **BP3**, and call this **BP4**.

If necessary, some of the alternatives for smoothing the BP combined results (using various averages) could be used if there are still significant discontinuities in the **BP4** results.

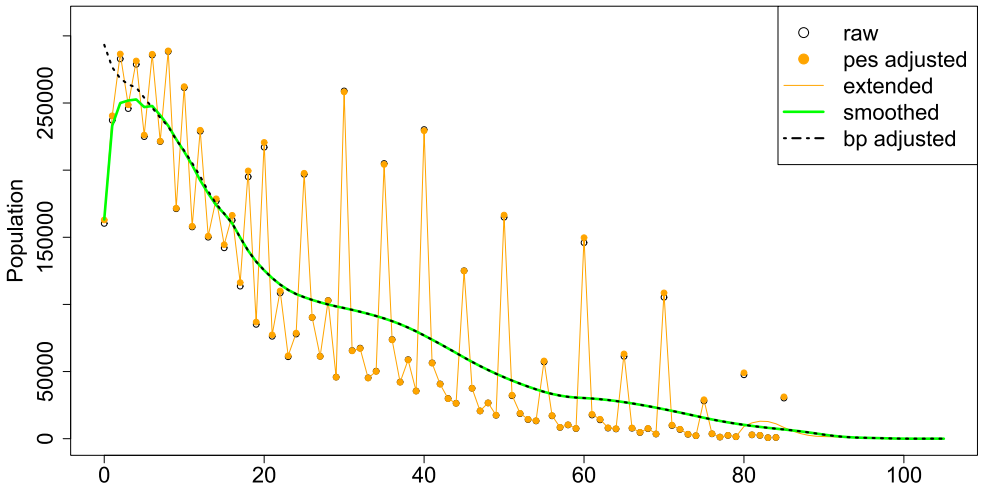
Once the population has been adjusted for the missing children, the adjusted child population is spliced with the smoothed single age counts up to age 15 and combined with the single age counts above age 15, yielding a population by sex and single year of age from age 0 to 105+.

VII. ILLUSTRATIVE EXAMPLE

The previous sections presented in detail the series of steps and procedures that were developed and followed to implement a standardized, reproducible, and transparent method to assess and adjust population distributions by age and sex. The present section provides an illustration of those steps Morocco, which has conducted six population censuses starting from 1960. Two of those censuses, the 1971 and the 2004 censuses, were selected to illustrate the different cases presented in the previous sections of this method protocol (i.e., heaping, smoothing, adjustment for the under-enumeration of children, etc.).

Figure 7.1 presents the steps of census workflow to arrive to an adjusted population by single age and sex for the female population of Morocco enumerated in the 1971 census. The black emptied dots (series 'raw' on figure 7.1) represent the female population by single year of age enumerated in the census from age 0 to the open-ended age group of 85+. These raw data indicate a rather poor quality of the census data, with strong heaping on various ending age digits. The yellow dots ('pes adjusted' on figure 7.1) is the raw data adjusted by the post-enumeration survey. The 1971 population census of Morocco was not followed by a PES, so the adjustment made is based on the PES statistical model that was developed as described in the technical annex. The age distribution of the 1971 census of Morocco is available only to the open-ended age group of 85+, so it needed to be extended to age 105+ (yellow line 'extended' series in figure 7.1). The extension of the age distribution to age 105+ modified only the age distribution starting from age 80 years, without altering the population count at other ages.

Figure 7.1. Census workflow adjustments, Morocco, female population, 1971 census



Strong age heaping is affecting the population age distribution enumerated in 1971. For the adult population, the Bachi index reaches a value of 34.93. Given the value of the Bachi index, the single age distribution was deemed too rough to work with and was collapsed into five-year age group data, as indicated in Table 5.2. Based on the value of the age ratio score (33.17), the mav2 method was determined to be the best graduation method to obtain a single age distribution for adult ages. For the child population,

the single age data were smoothed using a mav10 method. The resulting joint child and adult smoothed population appears in green in figure 7.1.

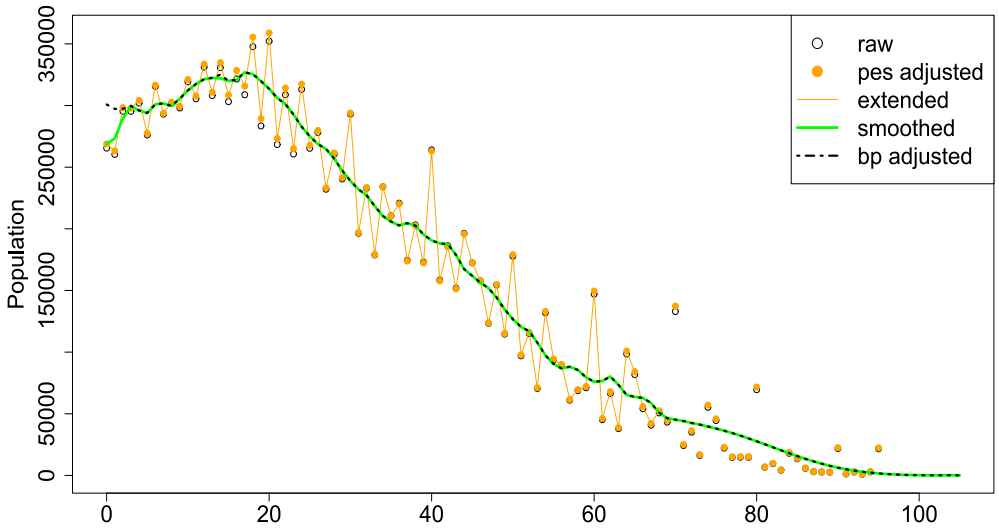
The last step consisted of assessing the extent of under-enumeration in the child population. The reconstructed child population indicates larger number of children under age 7 years.

With the application of the adjustments and smoothing discussed here, the black dashed line corresponds to the final adjusted female population by single age for the 1971 census of Morocco ('bp adjusted' series in figure 7.1). It is this series that is used as population benchmark for 1971 to assess the population reconstruction performed in the World Population Prospects.

Through time, the quality of the population censuses in Morocco has improved and at the 2004 census the quality of the age declaration is much better. Figure 7.2 shows the age distribution of the female population enumerated in the 2004 census of Morocco, together with the various series adjusting for coverage, smoothing and adjustments.

The population by single year of age that was enumerated in the 2004 census (with data available up to open age group 95+) shown the black empty circles ('raw' series in figure 7.2). Compared to the 1971 census, the quality of the age declaration has significantly improved. The application of the PES statistical model produces an adjusted population ('pes adjusted' series in figure 7.2) that is very close to the enumerated population. The Bachi index for the adult population reached 6.37—a substantial improvement from 1971—and a mav6 method was determined the best method to smooth the single age data. For the child population, the Bachi index was 1.56 and a mav2 smoothing method was applied. The joint smoothed population is indicated in green in figure 7.2.

Figure 7.2. Census workflow adjustments, Morocco, female population, 2004 census



The application of the adjustment for the under-enumeration of children returns the black dashed series ('bp adjusted' series in figure 7.2). This series is used as benchmark population to assess the consistency of the population reconstruction made in WPP. The under-enumeration of the children is not too important in the 2004 census; only the population at age 0 and age 1 were adjusted for under-enumeration.

VIII. CONCLUSIONS

This technical paper documents a series of choices made by the United Nations Population Division in its efforts to develop the implementation of a standardized, reproducible and transparent method to assess and potentially adjust population distributions by age and sex. In reconstructing the population change and its demographic components from 1950 to today in the *World Population Prospects*, the population estimates and each estimated demographic component of population change need to be assessed for consistency against an available reference data source, a population benchmark, usually a population by age and sex enumerated in a census. The quality of the population benchmarks that serve as 'target' needs to be first assessed as an initial step before being used for comparison purposes.

First, efforts are needed to assure that the definition of population and territorial coverage are consistent before using such benchmarks for the WPP. Second, no population census is perfect. All censuses are affected by potential issues with over- or under-enumeration. Proper care needs to be taken to assess the extent of such issues. A post-enumeration survey is recommended to be conducted after the completion of each census enumeration to determine the coverage of the census count (both overall, and by age and sex). Given that many census do not have an associated PES, a statistical model was developed to estimate net enumeration errors based on the currently available set of PES. Third, the enumeration of population by age and sex needs to be extended up to the required open-ended age. This step is required to compare in a consistent way the population benchmarks (i.e., population censuses) at older ages with the population that is reconstructed starting from 1950. Fourth, the population distribution by age in many censuses suffers from various issues related to age heaping that artificially distort the population distribution across the ages. A special effort was made to develop procedures to assess and smooth the population distribution while preserving as much as possible real changes across cohorts from artificial distortions due to poor age declaration. These procedures target both the adult and the child population, as well as the distribution by single year of age or by abridged age group. Finally, under-enumeration of young children is common in statistically less advanced as well as in statistically more advanced countries. It is therefore imperative to assess the extent of the under-enumeration of children in population census. Such step is rarely conducted with the required diligence in many countries such that when a recent population census that is affected by serious under-enumeration of children is used as base population in national population projections, this deficiency carries serious implications on planning efforts, in particular for health and education services. The method proposed here is to assess the under-enumeration in the child population based on a comparison between the population under age 10 enumerated in a census and an adjusted population under age 10 reconstructed using recent mortality and fertility trends and levels. The method protocol also describes how the adjusted and smoothed child and adult populations are blended together to produce a final population estimate by (single) age and sex.

The procedures included in this method protocol were developed specifically to serve the needs of the United Nations Population Division in its analytical work on revising population estimates for the World Population Prospects. These procedures are sufficiently general and broad in nature to serve as a reference for the recommended steps to guide national practices in evaluating population by age and sex enumerated in a population census.

IX. TECHNICAL ANNEX: STATISTICAL MODEL OF POPULATION ADJUSTMENT

A. MODEL OF OVERALL NET CENSUS ERRORS

Several functional forms were initially explored to fit the data without the use of covariates (i.e., using only time, including census round decades), but the use of time-dependent covariates was found preferrable to deal with the mixture of available data (i.e., variable number of observations by country and between regions covering different time periods), and requirement to predict Net Census Errors (NCE) for locations and time periods without PES while allowing to control for various country characteristics in a multilevel linear mixed-effects model.

In developing a statistical model of population adjustment, the goal was to find generalized patterns of under- and over-reporting of population that are not specific to a particular census operation. These estimates were modeled to be functions of several time-dependent covariates from 1950 up to 2020:

a. Average number of years of education, by sex based on IHME GBD 2019 annual estimates;
b. Lag-Distributed Income (LDI) in log scale based on IHME GBD 2019 annual estimates;
c. Under-Five probability of dying (Q5) in log scale based on UN WPP 2019 revision (annually interpolated).

Additional time-dependent covariates like total fertility rate (TFR), based on annually interpolated estimates based on the WPP 2019 revision and completeness of vital registration for births or deaths were also considered but were not found statistically significant as explanatory covariates.

In addition, the following time-invariant regional grouping covariates (used for hierarchical regional nested levels) as defined in WPP 2022 list of locations[11] were also taken into account:

d. UN SDG region
e. UN sub-region

We modeled the relationship between NCE (net census error as the outcome of interest) and a set of covariates used as dependent variables using the following regression model:

$$NCE_{ijkl} = \beta_0 + \beta_1 PES_{ijkl} + \beta_2 EducYrsM_{ijkl} + \beta_3 EducYrsF_{ijkl} + \beta_4 logLDI_{ijkl} + \beta_5 logQ5_{ijkl} + u_j + v_k + w_l + e_{ijkl} \quad (9.1)$$

where ijkl is an i observation nested within country j nested within sub-region k and SDG region l, and $e_{ijkl} \sim N(0, \sigma_e^2)$, $u_j \sim N(0, \sigma_u^2)$, $v_k \sim N(0, \sigma_v^2)$, $w_l \sim N(0, \sigma_w^2)$

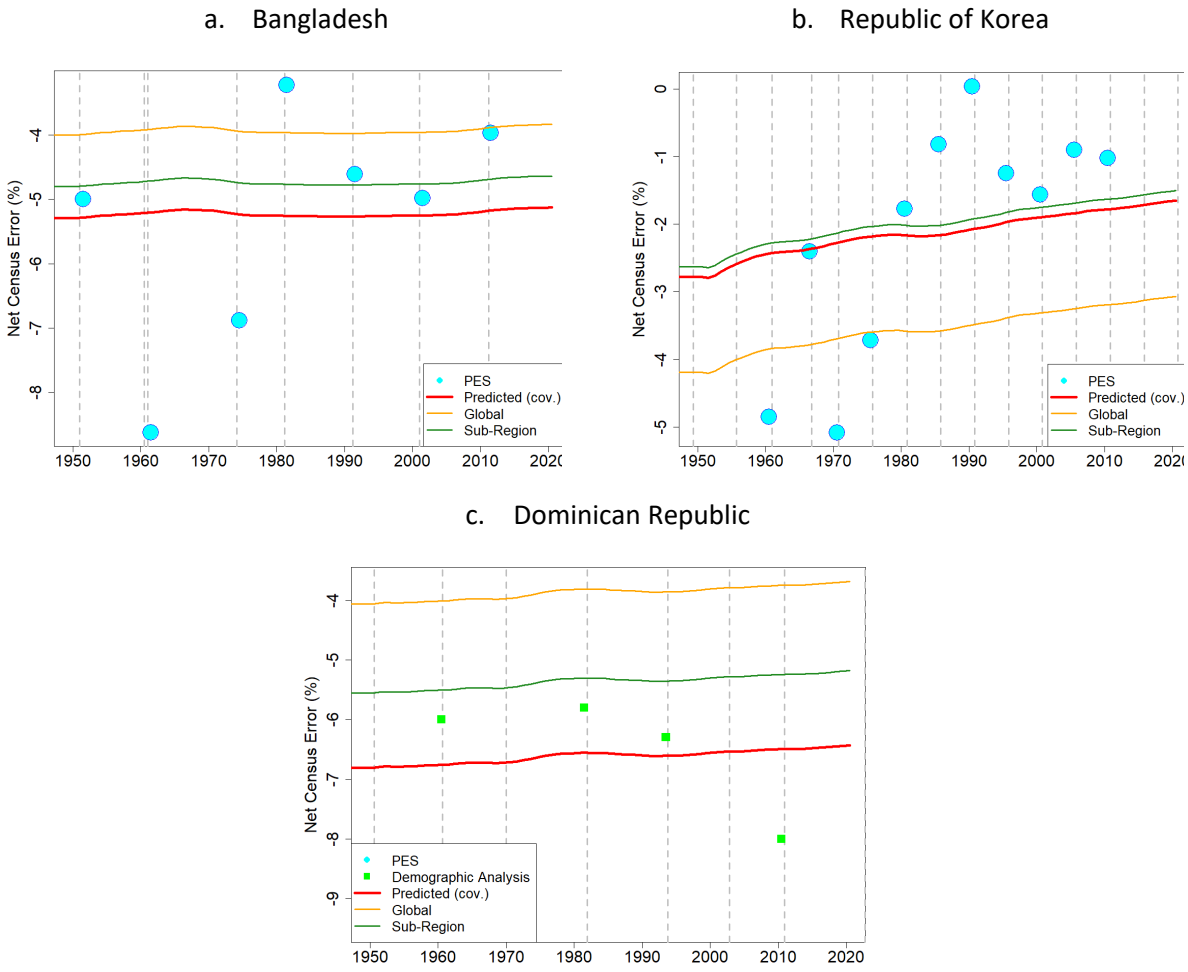
with PES a dummy variable equal to 1 if the observation is based on PES or 0 otherwise if it is based on Demographic Analysis (DA), EducYrsM and EducYrsF are respectively the male and female average number of years of education, logLDI is the log transformed Lag-Distributed Income, and logQ5 is the log transformed Under-Five probability of dying between birth and age 5. The model is fitted using the lme function of the nlme R package.[12] The random intercept is specified as 1 | SDGRegID / SubRegID / LocID

[11] https://population.un.org/wpp/Download/Files/4_Metadata/WPP2022_F01_LOCATIONS.XLSX
[12] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

where LocID is the location-specific M49 code, SubRegID is the UN sub-regional code, and SDGRegID is the SDG regional code.

Based on the set of time-dependent covariates several time series of expected NCE values can be predicted from 1950 to 2021 for all locations, as illustrated for three selected countries in figure 9.1. Each panel shows a country. The respective years with censuses are plotted with vertical gray dash lines, the PES estimates are shown as blue circles, and DA estimates as green squares (upon availability). The bold red line shows the predicted (or expected) net census error (NCE) for (1) the country-specific expected values; the green line indicates the UN sub-region model and the yellow line the overall global model. These last two series are shown for information purposes only.

Figure 9.1. Observed and predicted PES net census errors, selected countries



B. Model of differences in Net Census Errors from the overall level, by sex and age

The model uses the differences in NCE values by age from the total (rather than the ratios). This has several advantages over an approach using relative values:

1. Even when the overall level of NCE is changed, there is no need to readjust to get a total consistent with that implied by the overall level;
2. The measures are easier to understand;
3. The method does not assume that if the original NCE was zero then the values for all ages must be zero;
4. It does not have numerical problems (with potentially extreme ratios) when the total NCE is close to zero.

The new measure DiffNCE(x) is defined as:

$$DiffNCE(x) = NCE(x) - NCE \tag{9.2}$$

If we, for example, use the NCE from the model, say NCE', then we can get revised estimates by age:

$$NCE'(x) = DiffNCE(x) + NCE' \tag{9.3}$$

We can then adjust the census populations by age using these values:

$$AP'(x) = CP(x) / [1 + NCE'(x)] \tag{9.4}$$

One of the challenges is the paucity of publicly available PES results that are disaggregated by age and sex. However, more information is available only for overall Net Census Errors by sex.

Information on sex-specific net census errors was available for about 100 censuses. To leverage these extra set of empirical observations, a second model was estimated building on the first model of overall net census errors, this time to fit the sex-specific differences and to predict them for all countries from 1950 to 2021. The functional form is similar to the first model, but the overall net enumeration error was included as an additional covariate.

Knowing the sex-specific NCE (respectively NCE_M and NCE_F for males and females) for a smaller subset of observations than for the overall total, the sex-specific difference (e.g., NCE_M_Diff = (NCE_M – NCE) for male) can be computed for this subset (about 100 censuses), and the following analytical forms were fitted on the data by sex:

$$\text{For males: NCE\_M\_Diff}_{ijkl} = \beta_0 + \beta_1 NCE_{ijkl} + \beta_2 PES_{ijkl} + \beta_3 EducYrsM_{ijkl} + \beta_4 logLDI_{ijkl} + \beta_5 logQ5_{ijkl} + u_j + v_k + w_l + e_{ijkl} \tag{9.5}$$

$$\text{For females: NCE}_{\text{F}_{\text{Diff}}ijkl} = \beta_0 + \beta_1 NCE_{ijkl} + \beta_2 PES_{ijkl} + \beta_3 EducYrsF_{ijkl} + \beta_4 logLDI_{ijkl} + \beta_5 logQ5_{ijkl} + u_j + v_k + w_l + e_{ijkl} \tag{9.6}$$

Finally, based on the availability of PES data for 56 censuses for 28 countries, a model of the differences in the level of net enumeration by sex and age was fitted on this subset of data to predict net census errors by sex and age for all locations from 1950 up to 2021. This model was sex-specific, with the same covariates as in the first model, but with an age interaction, as well as the overall net census error and the sex-specific difference. The following analytical forms were fitted on the data by sex:

- For males: $\text{NCE\_M\_Diff\_x}_{ijkl} = \beta_0 + \beta_1 NCE_{ijkl} + \beta_2 \text{NCE\_M\_Diff}_{ijkl} + \beta_3 Age5 + \beta_4 PES_{ijkl} + \beta_5 (Age5 * EducYrsM_{ijkl)} + \beta_6 (Age5 * logLDI)_{ijkl} + \beta_5 (Age5 * logQ5)_{ijkl} + u_j + v_k + w_l + e_{ijkl}$ (9.7)

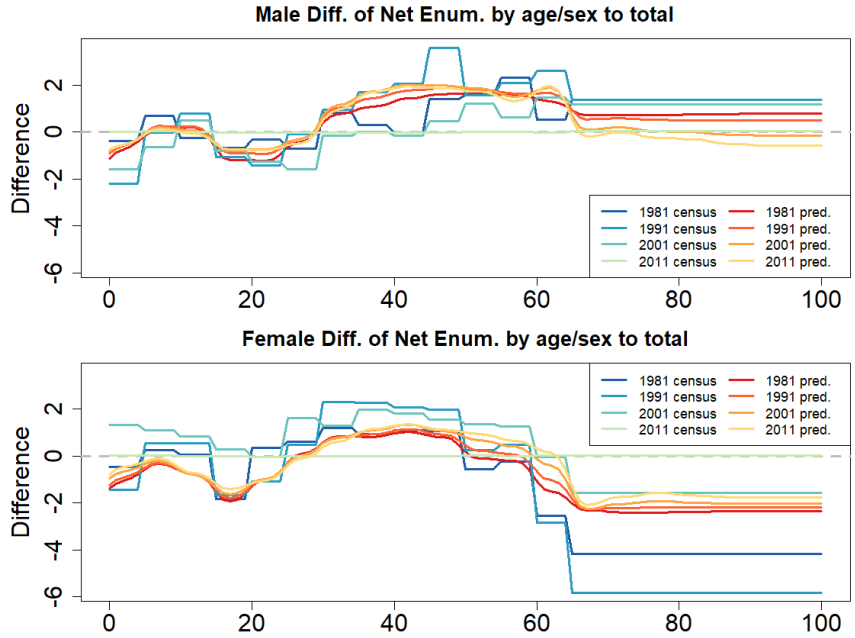
- For females: $\text{NCE\_F\_Diff\_x}_{ijkl} = \beta_0 + \beta_1 NCE_{ijkl} + \beta_2 \text{NCE\_F\_Diff}_{ijkl} + \beta_3 Age5 + \beta_4 PES_{ijkl} + \beta_5 (Age5 * EducYrsM_{ijkl)} + \beta_6 (Age5 * logLDI)_{ijkl} + \beta_5 (Age5 * logQ5)_{ijkl} + u_j + v_k + w_l + e_{ijkl}$ (9.8)

where NCE is the overall NCE for both sexes, NCE_M_Diff and NCE_F_Diff are the respective sex-specific overall differences (all ages), x is the age varying from 0 to 100, and Age5 is the associated 5-year abridged age group (i.e., 0, 1-4, 5-9, 10-14, etc.) used as covariate and interaction into the model. The initial predicted values for abridged age groups are smoothed using a spline function to obtain the values by single ages.

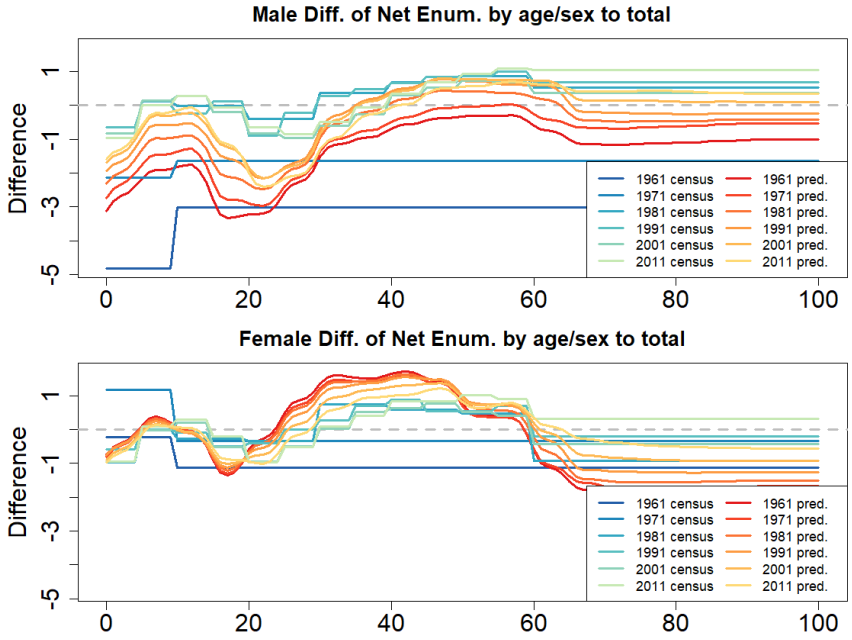
In the series of figures below (figure 9.2), the input values, are shown in blue/green, the predicted model values are shown in and in red/yellow for the respective censuses with available PES results. PES results are sometimes reported in broader age groups in which case, all age distributions are first standardized into single age distributions based on uniform assumptions within each age group (shown as extended horizontal blue lines in the plots below). The resulting predicted values from the modelling are by single age.

Figure 9.2. Difference in net census error by sex and age, selected countries

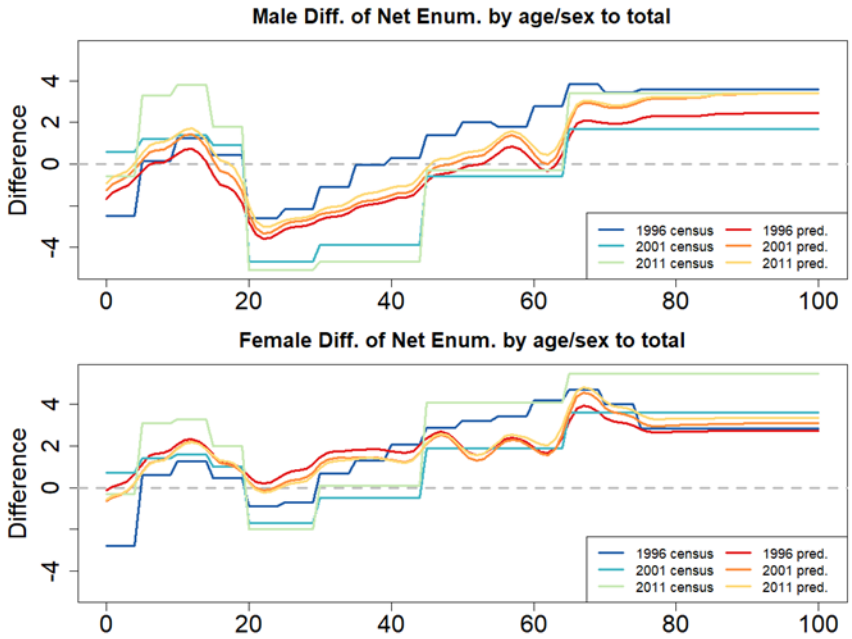
a. Bangladesh



India



b. South Africa



X. REFERENCES

Arriaga E., and others (1968). *New life tables for Latin American populations in the nineteenth and twentieth centuries*. Population Monograph Series, No. 3, appendix 3, pp. 295. University of California, Berkeley.

Arriaga, E., P. Anderson, and L. Heligman (1976). *Computer Programs for Demographic Analysis*. (Washington, D.C., Government Printing Office).

Arriaga, E., P. Johnson, and E. Jamison (1994). *Population Analysis with Microcomputers,* vol. I. Washington, DC: U.S. Census Bureau. https://2.census.gov/software/pas/documentation/pamvi-archive.pdf.

Bachi, R. (1951). The Tendency of Round Off Age Returns: Measurement and Corrections. *Bulletin of the International statistical Institute*, vol. 33, No. 4, pp. 195-222.

_____ (1953). Measurement of the Tendency of Round Off Age Returns. *Bulletin of the International Statistical Institute*, vol. 34, No. 3, pp. 129-137.

Beers, H. S. (1945). Modified Interpolation Formulas That Minimize Fourth Differences. *Record of the American Institute of Actuaries*, vol. 34, pp. 14-61.

Carrier, N.H., and A.M. Farrag. (1959). The reduction of errors in census populations for statistically underdeveloped countries. *Population Studies*, vol. 12, No. 3, pp. 240–285.

de Oliveira Carnevali, R. and P. Gerland. (2021). *Analyzing the Quality of Age Reporting from Population Censuses for All Countries of the World: 1950-2020 Census Rounds*. Paper presented in Session 113. Measurement of Age and Age Structures, IUSSP International Population Conference 2021, Hyderabad, India, 5-10 December 2021.

Ewbank, D. C. (1981). *Age Misreporting and Age-Selective Underenumeration: Sources, Patterns and Consequences for Demographic Analysis*. Washington, D.C., National Academy Press, Committee on Population and Demography, Report 4.

Feeney, G. (2013). *Removing "Zigzag" from Age Data*. http://demographer.com/white-papers/2013-removing-zigzag-from-age-data/.

IPUMS International (2021). World Population Census Forms, 1955-Present. https://international.ipums.org/international/census_forms.shtml.

_____ (2021). Source Documents. https://international.ipums.org/international/enum_materials.shtml.

_____(2019). *Census questionnaire for the 1971 Census of Haiti* (translated into English). https://international.ipums.org/international-action/source_documents/enum_instruct_ht1971a_tag.xml.

Myers, R. (1954). Accuracy of Age Reporting in the 1950 United States Census. *Journal of the American Statistical Association,* vol. 49 No. 268, pp. 826-831.

_____(1940). Errors and Bias in the Reporting of Ages in the Census data. *Transactions of the Actuarial Society of America*, vol. 41, No.2, pp. 411-415.

Noumbissi, A. (1992). L'indice de Whipple modifié: une application aux données du Cameroun, de la Suède et de la Belgique. *Population,* vol. 47, No. 4, pp. 1038-1041.

Philippines National Census and Statistics Office (1978). *1975 Integrated Census of Population and Its Economic Activities,. vol. II. National Summary. Phase I. Population. Appendix*. https://psa.gov.ph/sites/default/files/1975%20ICPEA%20Philippines.pdf.

Riffe, T. (2021). *DemoTools: Standardize, Evaluate, and Adjust Demographic Data*. https://rdrr.io/github/timriffe/DemoTools/.

Siegel, J. S., and D. A. Swanson (2004). *The Methods and Materials of Demography*. Second edition. San Diego, CA: Elsevier Academic Press.

Spoorenberg, T. (2007). Quality of age reporting: extension and application of the modified Whipple's index. *Population (English Edition)*, vol. 62, No. 4, pp. 729-741.

Sprague, T. B. (1880). Explanation of a new formula for interpolation. *Journal of the Institute of Actuaries*, vol. 22, No. 4, pp. 270-285.

United Nations (2022a). *World Population Prospects 2022: Summary of Results*. New York, United Nations, Department of Economic and Social Affairs, Population Division, UN DESA/POP/2022/TR/NO. 3.

_____(2022b). *World Population Prospects 2022: Methodology of the United Nations Population Estimates and Projections*. New York, United Nations, Department of Economic and Social Affairs, Population Division, UN DESA/POP/2022/TR/NO. 4.

_____(2013). *MORTPAK for Windows* (version 4.3). New York, United Nations, Department of Economic and Social Affairs, Population Division.

_____(2010). *Post Enumeration Surveys. Operational Guidelines*. Technical Report, New York, United Nations.

_____(1988). *MORTPAK-LITE - The United Nations Software Package for Mortality Measurement*. Interactive Software for the IBM PC and Compatibles. Sales No. E.88.XIII.2. New York.

_____(1952). Accuracy tests for census age distributions tabulated in five-year and ten-year groups. *Population Bulletin*. No. 2, ST/SOA/Ser.N/2, pp. 59-79. New York.

United Nations Statistics Division (UNSD). (2017). *Principles and Recommendations for Population and Housing Censuses, Revision 3*. New York, United Nations, Department of Economic and Social Affairs, Statistics Division, ST/ESA/STAT/SER.M/67/Rev.3.

U.S. Census Bureau (2019). *Population Analysis System*. https://census.gov/data/software/pas.html.

_____ (1985). *Evaluating Censuses of Population and Housing*. Statistical Training Document. ISP-TR-5. Washington, DC.

_____ (no date). *NewPAS* unpublished workbooks.