Methods for Estimating and Projecting Population by Age and Sex for Subnational Areas of Countries in the HIV Spatial Data Repository

Peter D. Johnson and Robert M. Leddy, Jr.
Population Division
U.S. Census Bureau

Introduction

This paper describes the methods used to estimate and project population by age and sex for subnational areas for a country over a range of years, currently 2000 to 2015. The basic method used is sometimes called “raking” or the “method of iterative proportional fitting,” or the “contingency table method” (Arriaga and Associates, 1994, pp. 43-44). In general, the raking method estimates the contents of a two-dimensional table of values based on assumptions or independent estimates of the row and column totals (also called the marginals) and an initial distribution of the interior of the table (referred to here as the base population).

In this application of the raking method, the rows represent age/sex categories and columns represent subnational areas of a country for a particular year. The initial age/sex distributions or base populations for the subnational areas are based on the most recent census data for the country at the lowest available subnational level.

The subnational areas are sometimes referred to as being of different administrative levels, where the first level, or ADM1, represents the highest level of geographic breakdown (e.g., states or provinces). The second level, or ADM2, breaks down the geography to the next level (e.g., counties or districts) and the third level, or ADM3, represents the next level (e.g., municipalities or townships).

The independent row totals in this case are the total country age/sex distributions for 2000 to 2015 developed by the U.S. Census Bureau and presented in the International Data Base (IDB) <http://www.census.gov/ipc/www/idb/>.

Estimates for midyears 2000 through 2015 are made using one of two raking procedures:

Rake method L: Rake all the lowest level ADM data to the total. The ADM1 (and ADM2 if needed) values are then obtained by summing over the lower level units within each ADM1 (or ADM2).

Rake method 1: Rake the ADM1 data to the total then rake the lowest level ADMs (2 and/or 3) within each ADM1 to the ADM1 total. If the lowest level is ADM3, then the ADM3's are summed to get ADM2 totals**.**

Preliminary results indicate that the differences between these two methods are minimal.

General Rake Model Overview

The model is described in terms of rows (meaning age/sex categories) and columns (meaning ADM units). In general, the model assumes that the percent distribution by ADM unit does not change over time, and the problem is to find a set of populations that adds to the desired country totals by age and sex (from the IDB) and to the IDB total distributed by ADM unit based on the census (base) distribution.

(1) $P(t_0, s, x)$ = Base population for year $\mathbf{t_0}$, subarea **s**, and age/sex group **x**. This is usually the reported census population for a recent year ($\mathbf{t_0}$=the census date).

(2) $Pr(t,s,x)$ = Initial population for subarea **s**, age/sex group **x**. The initial estimate of the distribution for year **t** is assumed to be the same as the base population from the most recent census:

$$Pr(t, s, x) = P(t_0, s, x)$$

(3) $R(t, s)$ = Total population for subarea **s**. For this version, the distribution by area is assumed to be the same as in the census data, so:

$$R(t, s) = \sum_x Pr(t, s, x)$$

(4) $FT(t, x)$ = Final total population for year **t** and age/sex group **x**. In this case, the data come from the International Data Base (IDB).

(5) $FTT(t)$ = Final total population, all ages combined, for year **t**:

$$FTT(t) = \sum_x FT(t, x)$$

(6) k = Adjustment factor to compute the final revised total by subarea:

$$k = \frac{FTT(t)}{\sum_s R(t, s)}$$

(7) $F(t,s)$ = Final total population for subarea **s**:

$$F(t, s) = k \times R(t, s)$$

(8) cf(t,s) = Column factor adjustment for subarea **s**:

$$\mathrm{cf}\,(\mathrm{t},\mathrm{s}) = \frac{\mathrm{F}(\mathrm{t},\mathrm{s})}{\sum_{\mathrm{x}} \mathrm{Pr}(\mathrm{t},\mathrm{s},\mathrm{x})}$$

(9) Pc(t,s,x) = Population with column adjustment for subarea **s**, age/sex group **x**:

$$\mathrm{Pc}(\mathrm{t},\mathrm{s},\mathrm{x}) = \mathrm{cf}\,(\mathrm{t},\mathrm{s}) \times \mathrm{Pr}(\mathrm{t},\mathrm{s},\mathrm{x})$$

(10) rf(t,x) = Row adjustment factor for age/sex group **x**:

$$\mathrm{rf}\,(\mathrm{t},\mathrm{x}) = \frac{\mathrm{FT}(\mathrm{t},\mathrm{x})}{\sum_{\mathrm{s}} \mathrm{Pc}(\mathrm{t},\mathrm{s},\mathrm{x})}$$

(11) Pr(t,s,x) = Population of subarea **s**, age/sex group **x**, with row (age/sex) adjustment:

$$\mathrm{Pr}(\mathrm{t},\mathrm{s},\mathrm{x}) = \mathrm{rf}\,(\mathrm{t},\mathrm{x}) \times \mathrm{Pc}(\mathrm{t},\mathrm{s},\mathrm{x})$$

(12) cf(t,s) = Column factor adjustment for subarea **s**:

$$\mathrm{cf}\,(\mathrm{t},\mathrm{s}) = \frac{\mathrm{F}(\mathrm{t},\mathrm{s})}{\sum_{\mathrm{x}} \mathrm{Pr}(\mathrm{t},\mathrm{s},\mathrm{x})}$$

(13) Pc(t,s,x) = Population with column adjustment for subarea **s**, age/sex group **x**:

$$\mathrm{Pc}(\mathrm{t},\mathrm{s},\mathrm{x}) = \mathrm{cf}\,(\mathrm{t},\mathrm{s}) \times \mathrm{Pr}(\mathrm{t},\mathrm{s},\mathrm{x})$$

(14) Compute the sum of absolute differences between the independent totals by age/sex and the row sums of the population with column adjustment:

$$\mathrm{SAD}(\mathrm{t}) = \sum_{\mathrm{x}} \left| \mathrm{FT}(\mathrm{t},\mathrm{x}) - \sum_{\mathrm{s}} \mathrm{Pc}(\mathrm{t},\mathrm{s},\mathrm{x}) \right|$$

(15) If SAD(t) is "small enough," then the current set of estimates, Pc(t,s,x), appears to be a solution. The desired tolerance is the user specification of what is "small enough," and if it is not, then the program goes back to step (10) and repeats the process. If there is no conversion to the indicated tolerance after 100 iterations, the program stops iterations for that cycle.

Currently, the default tolerance is set to 1, meaning that the sum over all age groups of the absolute difference between the independent estimate and the sum over all the subareas must be less than 1.

(16) After the unrounded final results have been obtained, then the row adjustment factors are re-estimated, rf(t,x):

$$\mathrm{rf}(t,x) = \frac{\mathrm{FT}(t,x)}{\sum_{s} \mathrm{Pc}(t,s,x)}$$

(17) Then the population figures are re-estimated with row adjustment, but this time with progressive rounding:

$$\mathrm{Pr}(t,s,x) = \mathrm{Round}\left[\mathrm{rf}(t,x) \times \sum_{r=0}^{s} \mathrm{Pc}(t,r,x)\right] - \sum_{r=0}^{s-1} \mathrm{Pr}(t,r,x)$$

Additional models, assumptions, and issues

The rake model above assumes that the base (census) data are available by sex and 5-year age groups up to the open-ended group “80 years and over” (sometimes designated as “80+”). This section describes some methods that are used to transform the raw census data into the form that is needed by the rake model.

Unknown age

In spite of improvements in data processing (such as implementing the “hot deck” method, which tries to impute characteristics for a respondent with missing data based on a previously processed respondent with other characteristics in common), census results often are presented with an “Unknown” or “Not stated” age category. Fortunately, the percent of the population that falls into these categories is usually small. The general procedure for dealing with this is to inflate the population for each sex by the ratio of the “total” over the “total” minus the “not stated ages”:

$$\mathrm{Factor}(g) = \frac{\mathrm{P}(g)}{\mathrm{P}(g) - \mathrm{NS}(g)}$$

This factor can then be used to inflate the population in each age group:

$$\hat{\mathrm{P}}(g,a) = \mathrm{P}(g,a) \times \mathrm{Factor}(g)$$

The progressive rounding method (see formula 17 above) is used to ensure that the rounded numbers add up to the desired totals by sex.

Broad age groups

Census data are often presented in broader age groups than the standard for this project, often with a younger open-ended age group (e.g. as low as 65+) and/or 10-year age groups (e.g. 50-59, 60-69). This may be done for a number of reasons, including:

1. Known problems with the quality of age reporting. This is often worst at the oldest ages.
2. Issues of space in the census report (especially for subnational data)
3. Issues of disclosure and/or size of the population. If the data are shown for very small ADM areas, there may not be many (any?) people in some age groups.

Regardless of the reasons for the lack of the full set of desired data, there are a number of methods that can be used to approximate the missing data.

If the full age range is available at a higher ADM level, these data can be used to adjust the lower level data. For example, if the full range of data is available for regions, then this information can be used to split the lower level data in age groups 55-64, 65-74, and 75+:

$$\hat{P}(s,55-59) = P(s,55-64) \times \frac{P(r,55-59)}{P(r,55-64)}$$

$$\hat{P}(s,60-64) = P(s,55-64) - \hat{P}(s,55-59)$$

...

If the data are only presented up to age group 75+, even at the national level, then the simplest solution is to use the IDB data for the census year to split the 75+ into 75-79 and 80+.

Census problems

If it is known that a census was not completed in part of a country (usually due to wars, other violence, or natural disasters), then attempts are made to adjust for this. This may involve looking at earlier census data as well as other sources of data. These situations are fully discussed in the country-specific metadata.

Boundary issues

Boundary issues may create census problems if areas claimed by a country are either not counted, e.g., due to lack of or limited access, or estimated without adequate discussion of the methods used. These issues are documented in the country-specific metadata.

Trend Methods

When populations for a country's subareas are available from two censuses, it may be feasible to determine final populations for the subareas for a range of years based on the areas' intercensal growth (or declining) rates relative to the country. These final populations can be used as controls along with the national age/sex group populations from the IDB to calculate the subareas' age/sex group estimates with the General Rake Model. For subarea **s** for year **t**, this is substituting the subarea's final population in the expression F(t,s) from step (7) in the General Rake Model Overview section (p. 2). In basing final populations on intercensal growth rates, the subarea boundaries must have remained constant during the period spanning the two censuses and the years for which the estimates are to be calculated.

Various trend methods can be applied to calculate values for F(t,s), including the following which have been used to calculate the 2000-2015 subnational 5-year age/sex group population estimates for some countries uploaded on the HIV Spatial Data Repository.

Shift-Share

Using the Shift-Share method (Siegel and Swanson, 2004, pp. 570-571), the rate of change of the ratios of the population of each subarea to that of the country during the intercensal period is first calculated. This rate of change can be extrapolated to calculate estimates for the subarea for a range of years using the national projected populations from the IDB as the base numbers for the country. These estimates can serve as the subarea's final populations to be used as controls for calculating the subarea's age/sex group estimates with the General Rake Model. The Shift-Share equation is as follows:

$$\mathrm{SE}(t_E, s) = \left\{ \left[\frac{\dfrac{\mathrm{SC}(t_2,s)}{\mathrm{NC}(t_2)} - \dfrac{\mathrm{SC}(t_1,s)}{\mathrm{NC}(t_1)}}{(t_2 - t_1)} \right] \times (t_E - t_2) + \frac{\mathrm{SC}(t_2,s)}{\mathrm{NC}(t_2)} \right\} \times \mathrm{NP}(t_E)$$

Where:

$\mathrm{SE}(t_E,s)$ = subnational area population estimate for time t_E
$\mathrm{NP}(t_E)$ = projected national total population for time t_E (from IDB)
$\mathrm{SC}(t_C,s)$ = subnational area census population for time t_C
$\mathrm{NC}(t_C)$ = national census population for time t_C
t_E = time of the estimated population (July 1 of the IDB year)
t_2 = time of latest census
t_1 = time of census before latest

Once the $\mathrm{SE}(s,t_E)$ values for all of the subnational areas have been calculated, these values must be adjusted using the progressive rounding procedure (Step (17), p. 4) so that they add exactly to the IDB national control ($\mathrm{NP}(t_E)$ in the equation above). For subarea **s** for year $\mathbf{t_E}$, the adjusted value becomes the final population that can be used for F(t,s) in Step (7) on p. 2, where $t=t_E$. Thus:

$$F(t_E,s) = \text{round}\left[\sum_{i=1}^{s} SE(t_E,i) * NP(t_E) \ / NP_0(t_E) \right] - \sum_{i=1}^{s-1} F(t_E,i)$$

Where:

$SE(t_E,i)_i$ = population estimates calculated from the Shift-Share equation for subarea **i** for time t_E
$NP(t_E)$ = projected national total population for time t_E (from IDB)
$NP_0(t_E)$ = sum of the $SE(t_E,s)$ values from the Shift-Share equation for all subareas, i.e., $NP_0(t_E) =\sum_i[SE(t_E,i)]$

With the value for $F(t_E,s)$, the 5-year age/sex group population estimates for subarea **s** for time $\mathbf{t_E}$ can be calculated using the General Rake Model.

In using the Shift-Share method, it may be necessary to apply constraints to subareas that experienced much higher growth rates than the country or experienced significant population loss while the country as a whole grew between the censuses. This is so as not to yield unusually high or low (or negative) population estimates for the area.

Logistics Growth Rate

The Logistics Growth Rate method is based on the premise that population growth follows an S curve pattern (Siegel and Swanson, 2004, p. 568). Bounded by lower and upper limits (asymptotes), growth rates for subareas relative to their countries can be constrained when applying this method. For calculating the subnational 5-year age/sex group estimates for a country's subnational areas for a range of years, Arriaga's "modified logistics function" (Arriaga and Associates, 1994, pp. 303-304), a form of the Logistics Growth Rate method, is applied to calculate the Logistics Growth Rate (LGR) which is based on the population change of a subnational area relative to that of the country between the latest 2 censuses. For subarea **s** the equation is as follows:

$$LGR(s) = \ln\{ [(U – PR(t_1,s)) / (PR(t_1,s) - L)] / [(U – PR(t_2,s)) / (PR(t_2,s) - L)] \} / (t_2 – t_1)$$

Where:

$LGR(s)$ = average annual logistics growth rate for subarea **s** between the country's latest 2 censuses
$PR(t_C,s)$ = population ratio: subarea s to country at census for time t_C
t_2 = time of latest census
t_1 = time of census before latest
L = lower asymptote

U = upper asymptote

The variables $PR(t_1,s)$ and $PR(t_2,s)$, representing the ratios of the subarea **s** populations to the country populations at a census, can be defined by using variables from the Shift-Share equation above. That is, $PR(t_1,s) = SC(t_1, s) / NC(t_1)$ and $PR(t_2,s) = SC(t_2, s) / NC(t_2)$.

The LGR is extrapolated from the latest census to the range of years for which the subareas' 5-year age/sex group population estimates are to be calculated. To calculate a population $SE_0(t_E,s)$ for subarea s for year $\mathbf{t_E}$, using the IDB projected population for the country, and setting the lower asymptote (L) to 0 and the upper asymptote (U) to 1, the LGR equation becomes:

$$SE_0(t_E,s) \;= NP(t_E) /\{\, 1+ [\, (1 /PR(t_2,s)\,) - 1] / \exp [\, LGR(s) * (t_E - t_2)\,]\, \}$$

Where:

$SE_0(t_E,s)$ = population estimate for subarea s for the year t_E
LGR(s) = average annual logistics growth rate for subarea **s** between the country's latest 2 censuses
$PR(t_2,s)$ = population ratio: subarea s to country at latest census
t_2 = time of the latest census
t_E = time of the estimated population (July 1 of the IDB year)
$NP(t_E)$ = projected national total population for time $\mathbf{t_E}$ (from IDB)

The population estimates for the subareas [$SE_0(t_E,s)$'s] for year t_E must be adjusted with the progressive rounding procedure (Step (17), p. 4) so that the numbers aggregate to the IDB national control $NP(t_E)$. The results are the final values for calculating the subareas' 5-year age/sex groups for year with the General Rake Model. For subarea **s** for year $\mathbf{t_E}$, the adjusted value becomes the final population that can be used for F(t,s) in Step (7) on p. 2, where $t=t_E$. Thus:

$$F(t_E,s) = \text{round}\left[\, \sum_{i=1}^{s} SE_0(t_E, i) * NP(t_E) \;/\; NP_0(t_E) \right] - \sum_{i=1}^{s-1} F(t_E,i)$$

Where:

$SE_0(t_E,i)$ = population estimates calculated from the Logistics Growth Rate equation for subarea **i**
$NP(t_E)$ = projected national total population for time t_E (from IDB)
$NP_0(t_E)$=sum of the $SE_0(t_E, s)$ values from the Logistics Growth Rate equation for all subareas, i.e., $NP_0(t_E) =\sum_i [SE(t_E,i)]$

With the value for $F(t_E,s)$ the 5-year age/sex group population estimates for subarea **s** for year $\mathbf{t_E}$ can be calculated using the General Rake Model.

Averaging of Final Populations from the Shift-Share and Logistics Growth Rate Methods

For subarea **s** for year $\mathbf{t_E}$, the $F(t_E,s)$ populations calculated from the Shift-Share and Logistics Growth Rate methods can be averaged into a final population that can be used as a control for applying the General Rake Model. Obtaining an average final population from the final populations calculated from more than one trend method may improve the accuracy of the age-sex group population estimates by offsetting, at least to some degree, errors that may arise from using the methods individually (Siegel and Swanson, 2004, p. 593).

References

Arriaga, Eduardo E., and Associates. 1994. *Population Analysis with Microcomputers*. U.S. Census Bureau, International Programs Center: Washington, DC.

Siegel, Jacob and David A. Swanson. 2004. *The Methods and Materials of Demography*. Second Edition. Elsevier Academic Press: San Diego, CA.