# METHODS FOR CALCULATING 5-YEAR AGE GROUP POPULATION ESTIMATES BY SEX FOR SUBNATIONAL AREAS

Robert M. Leddy, Jr.
Geographic Studies Branch
Population Division
U.S. Census Bureau

September 26, 2013

The Geographic Studies Branch of the Population Division of the U.S. Census Bureau prepares 5-year age group population estimates by sex for subnational areas of countries. The estimates are linked as attribute data to digital boundary maps of these areas in a geographic information system (GIS). The target date for a set of subnational estimates for a country is July 1 of a specific year. The aggregates of these estimates are consistent with the U.S. Census Bureau's *International Data Base (IDB)* midyear 5-year age and sex group projections for that country, available at http://www.census.gov/population/international/data/idb/informationGateway.php.

For a particular country, one of the following five methods is used for calculating these subnational estimates, including estimates for the open-ended age and sex groups (usually 80 years and over):

- Constant-Share
- Iterative Proportional Fitting
- Shift-Share and Iterative Proportional Fitting combined
- Logistic Growth Rate and Iterative Proportional Fitting combined
- Shift-Share/Logistic Growth Rate Averages and Iterative Proportional Fitting combined

In applying these methods, estimates are calculated for the areas in the lowest of the first three subnational administrative levels for a country for which population data are available, whether the first-order divisions (provinces or states), second-order divisions (districts or counties), or third-order divisions (communes or townships). Estimates for the lower-order subnational areas are aggregated to the higher-order subnational areas. The calculations are based on most recent available census numbers or, absent a recent census, official estimates.

The tables with the subnational data that are calculated using any of the five methods have the following layout. Rows comprise records for the countries and their subnational divisions. Columns comprise the fields for the census total populations and for the age group population estimates for both sexes and for males and females.

## METHODS

### Constant-Share

With the *Constant-Share* method (George et al., 2004: p. 570), the age and sex group population estimates for a country's subnational areas for July 1 of a target year are calculated by prorating their census age and sex group populations to the IDB national projected age and sex group populations for that year. This is basic proportioning. Using the Constant-Share method, it is assumed (1) that the populations of a country's subnational areas are growing at the same rate as the country, and (2) that each subnational area has the same age and sex group population distribution as the country.

The estimates are calculated for each male and female 5-year and open-ended age group using the *progressive rounding* method. This method ensures that the rounded numbers yielded from the calculations add exactly to the IDB controls (unless the census numbers for a country are intentionally rounded). Thus for calculating any male or female age group estimate for the $n^{th}$ subnational area in the country for the target date, the formula is:

$$\mathbf{AG_{SAi} = [(\sum_{i=1}^{n} PC_{SAi})*(AG_{IDB}/PC_C)] - (\sum_{i=1}^{n-1} AG_i)} \qquad (1)$$

Where:

$AG_{SAi}$ = estimate for the age group (male or female) for the $n^{th}$ subnational area
$PC_{SAi}$ = census population for a subnational area (both sexes combined)
$AG_{IDB}$ = IDB projected national population for that age group for the target date
$PC_C$ = total census population for the country
$AG_i$ = estimate for a subnational area for that age group

and where the value for the expression $\mathbf{[(\sum_{i=1}^{n} PC_{SAi})*(AG_{IDB}/PC_C)]}$ is rounded to the nearest whole number.

The subnational male and female age groups are proportioned to the *total* population for both sexes for the country ($PC_{SAi}$ in Equation 1) rather than to the totals by each sex separately. This is because calculations based on the individual sexes can yield skewed age group population distributions by sex with respect to the distribution for the country.

The estimate for each age group with both sexes combined is calculated by adding the male and female estimates for that age group.

**Iterative Proportional Fitting**

In the *Iterative Proportional Fitting* method, the age and sex group population distributions in the subnational areas are taken into account (Judson and Popoff, 2004: p. 712; Arriaga, Johnson, and Jamison, 1994: pp. 43-44). The 5-year age and sex group populations for the subnational areas from the country's census form the base for calculating the age and sex group population estimates for July 1 of the target year. The numbers are prorated, i.e., raked, by a series of iterations to two sets of controls for that target date:

(1) IDB midyear 5-year age and sex group population projections for the country (same as with the Constant-Share)
(2) total population estimates for the subnational areas

The latter set, the total population estimates for the subnational areas, is calculated by prorating the census totals for the subnational areas to the IDB midyear projected national population. The reason for using totals for both sexes rather than totals for the males and females separately is the same as that for the Constant-Share method: so as not to yield estimates with a skewed age distribution by sex with respect to that for the country.

At the start, using the Iterative Proportional Fitting method requires a two-step procedure.

*Step 1:* The subnational area census populations for each male and female 5-year or open-ended age group are raked (by column) to the *first* set of controls: the corresponding IDB age group projections by sex for the country. This yields the first set of age group population estimates for each subnational area. Thus for an age group for a subnational area in a country that has *n* number of subnational areas:

$$\mathbf{AG_{SA1} = PC_{AG} * AG_{IDB} / (\sum_{i=1}^{n} AG_{Ci})} \qquad (2)$$

Where:

$AG_{SA1}$ = first estimate for a specific age group (male or female) for the subnational area
$PC_{AG}$ = census population for that age group for the subnational area
$AG_{IDB}$ = IDB midyear projected national population for that age group
$AG_{Ci}$ = census population for that age group for each of the *n* subnational areas

*Step 2:* The resulting first male and female age group estimates ($AG_{SA1}$ in Equation 2) are raked by subnational area (i.e., by row) to the *second* set of controls: the total estimates for the subnational areas, which aggregate to the IDB midyear projected total population for the country. This set yields a second series of age group estimates for each subnational area. The estimate for each age group among *x* number of male and female age groups for a subnational area is calculated as follows:

$$\mathbf{AG_{SA2} = AG_{SA1} * PE_{SA} / (\sum_{i=1}^{x} AG_{SA1i})} \qquad (3)$$

Where:

$AG_{SA2}$ = second estimate for a specific age group (male or female) for the subnational area
$AG_{SA1}$ = first estimate for the age group for the subnational area calculated in Step 1 (Equation 2)
$PE_{SA}$ = total estimated population for the subnational area for the target date
$AG_{SA1i}$ = estimate for each of *x* number of male and female age groups for the subnational area calculated in Step 1 (Equation 2)[1]

The completion of Step 2 ends the first iteration. The *second* iteration is begun by returning to Step 1 (raking again by columns). The variable $AG_{SA2}$ for each male and female age group is substituted for $PC_{AG}$ in Equation 2 and multiplied by a new ratio: the IDB midyear projected national population for that age group divided by the aggregate of the subnational areas' second set of estimates for that age group resulting from Equation 3. That is, the following equation is used to yield a third set of age group estimates, $AG_{SA3}$, for each subnational area:

$$\mathbf{AG_{SA3} = AG_{SA2} * AG_{IDB} / (\sum_{i=1}^{n} AG_{SA2i})} \qquad (4)$$

[1] If the data set for a subnational area contains 5-year age groups for ages between 0 and 79 and open-ended age groups for 80 years and over, estimates would be calculated for 34 groups -- 17 male, 17 female.

This iteration is completed by returning to Step 2 (to rake again by rows), substituting $AG_{SA3}$ for $AG_{SA1}$ in Equation 3 to yield a fourth set of estimates. Thus for the age group $AG_{SA4}$ for a subnational area:

$$\mathbf{AG_{SA4} = AG_{SA3} * PE_{SA} / (\sum_{i=1}^{x} AG_{SA3i})} \qquad (5)$$

The iterations are performed again as the aggregates of each new series of resulting male and female age group estimates converge on both sets of controls, i.e., the IDB midyear national 5-year age and sex group population projections *and* the total population estimates for the subnational areas.

The numbers resulting from the last iteration are subjected to the progressive rounding procedure (four iterations of the two-step procedure suffice). In the form of Equation 1, this is using the IDB midyear national 5-year age and sex group projection controls for the variable $AG_{IDB}$ and the subnational area total population controls ($PE_{SA}$ in Equations 3 and 5) for the variable $PC_{SAi}$. The 5-year and open-ended age and sex group estimates that result for each subnational area are the *final* population estimates. Age group population estimates for both sexes are obtained by adding their male and female counterparts.

## Shift-Share and Iterative Proportional Fitting Combined

It was noted in the foregoing section that the subnational population estimates used as controls for calculating age and sex group estimates are numbers prorated from a census; i.e., they are based on the country's growth rate. With one or more trend methods, the controls can be calculated based on the population growth rates of the individual subnational areas relative to the country. In simple terms, if the population of a province grew at a faster rate than that of the country between the country's two most recent censuses, the province can be expected to continue to grow faster than the country afterwards.

One such method, the *Shift-Share* method, can be used to calculate a population estimate for a subnational area for a target date based on the linear rate of the growth or decline of its share of the country's population between two censuses (George et al., 2004: pp. 570-571). If a province had 10 percent of a country's population at a 2000 census and 11 percent of that country's population at a 2010 census, it will be projected to have 12 percent of the country's population in 2020 and 13 percent in 2030. In using this method, the boundaries of the subnational areas would have had to remain constant during the time spanning both censuses and the target date.

With the Shift-Share method, the average annual rate of change of the ratio of the population of each subnational area to that of the country between the country's latest two censuses is calculated. This intercensal rate of change is extrapolated to the target year to yield a new ratio. With the projected population for the country from the IDB for that target midyear in the denominator, a population estimate for each subnational area for the midyear can be calculated.

The above steps for calculating an estimate for a subnational area using the Shift-Share method are combined in the following equation:

$$\mathbf{SA_P = \{[((SA_2/NT_2) - (SA_1/NT_1)) / (t_2 - t_1)] * (t_P - t_2) + (SA_2/NT_2)\} * NT_P} \qquad (6)$$

Where:

$SA_P$ = subnational area population for projection (target) date
$SA_2$ = subnational area population at latest census
$SA_1$ = subnational area population at census before latest
$NT_P$= national total population for projection (target) date
$NT_2$ = national total population at latest census
$NT_1$= national total population at census before latest
$t_P$ = date of projected population (target date)
$t_2$ = date of latest census
$t_1$ = date of census before latest

The target dates are always midyear dates (July 1). The national total population used for each target date ($NT_P$) is the midyear projected total from the IDB for the country.

The resulting set of population estimates for the subnational areas from these calculations, $SA_P$ in Equation 6, becomes the second set of controls for the Iterative Proportional Fitting method in lieu of estimates calculated by only prorating census numbers to the IDB projected national total. These controls ($SA_P$) are substituted for the values of $PE_{SA}$ in Equation 3 to calculate the age and sex group population estimates for the subnational areas using the Iterative Proportional Fitting method.

It must be noted that for subnational areas that experienced exceptionally high intercensal rates of growth with respect to the country, the Shift-Share method could yield exponentially (and unrealistically) high estimates for the target date. The opposite is true for exceptionally high rates of decline. Accelerated excessive population loss may be predicted; in fact negative estimates can result. When using this method for such cases, the rates would have to be constrained.[2]

### Logistic Growth Rate and Iterative Proportional Fitting Combined

An alternative to the Shift-Share method is the *Logistic Growth Rate* method. The Logistic Growth Rate method is based on the predication that population growth rates follow an S-curve progression of initially being slow, accelerating, peaking, and finally tapering, a path that is consistent with Malthusian theory (George et al., 2004: p. 568). Requiring the setting of lower and upper limits as asymptotes, the method can be particularly effective for constraining high rates of growth and decline. The Geographic Studies Branch uses a version of this method based on Arriaga's "modified logistic function" (Arriaga, Johnson, and Jamison, 1994: pp. 303-304) for calculating subnational area total population estimates for a target date. As with the Shift-Share method, the subnational boundaries would have had to remain constant during the time spanning the two censuses and the target date.

The calculations are performed in three steps. In the first, an average annual intercensal logistic growth rate is calculated based on the ratios of the populations of each subnational area to that of the country at the two latest censuses. This calculation is as follows:

$$\mathbf{LGR = \ln\{[(U - PR_1) / (PR_1 - L)] / [(U - PR_2) / (PR_2 - L)]\} / (t_2 - t_1)} \qquad (7)$$

Where:

LGR = average annual logistic growth rate between the country's two latest censuses
$PR_2$ = population ratio: subnational area to country, latest census
$PR_1$ = population ratio: subnational area to country, census before latest
$t_2$ = date of latest census
$t_1$ = date of census before latest
L = lower asymptote
U = upper asymptote

[2] For example, the rates of annual population growth in these cases might not be permitted to rise above 5 percent or fall below -1 percent.

The variables $PR_1$ and $PR_2$, representing the ratios of the subnational area populations to the country populations at the two most recent censuses, can be defined by using variables from Equation 6. That is, $\mathbf{PR_1= SA_1/ NT_1}$ and $\mathbf{PR_2= SA_2/NT_2}$.

For the second step, a preliminary population estimate for each subnational area is calculated. The average annual logistic growth rate LGR is extrapolated to the target date using the time span between the latest census and the target date. If the lower asymptote (L) is 0 and the upper asymptote (U) is 1, Equation 7 becomes:

$$\mathbf{SA_{P0} = NT_P / \{1 + [(1/PR_2) - 1] / \exp[LGR *(t_P - t_2)]\}} \qquad (8)$$

Where:

$SA_{P0}$ = preliminary total population estimate for a subnational area for the target date
LGR = average annual logistic growth rate between the country's 2 latest censuses (from Equation 7) extrapolated to the target date
$PR_2$ = population ratio: subnational area to country, latest census (from Equation 7)
$t_2$ = date of latest census
$t_P$ = date of projected population (target date)
$NT_P$= national total for the projection (target) date (from the IDB)

The third step is necessary if the aggregate of the preliminary total population estimates for the subnational areas for the target year ($\sum SA_{P0}$) does not equal the national projected total ($NT_P$).

That is, the final population estimates for the subnational areas for the target year, $SA_P$, are calculated by prorating the preliminary population values $SA_{P0}$ to the national total for the target date, $NT_P$. Here the progressive rounding method in the form of Equation 1 is applied. Like the Shift-Share method, the values for $SA_P$ become the second set of controls for calculating the subnational age and sex group estimates using the Iterative Proportional Fitting method. That is, $SA_P$ is substituted for the value of $PE_{SA}$ in Equation 3.

### Shift-Share/Logistic Growth Rate Averages and Iterative Proportional Fitting Combined

Another option is to average the subnational population estimates calculated from the Shift-Share and Logistic Growth Rate methods, and with the results apply the Iterative Proportional Fitting method. Averaging the values calculated from more than one trend method may improve the accuracy of the results by offsetting, at least to some degree, extremes that may arise from using the methods individually (George et al., 2004: p. 593).

For each subnational area for the target year, this is taking the average of the Shift-Share $SA_P$ from Equation 6 and the Logistic Growth Rate $SA_P$ calculated by prorating the value of $SA_{P0}$ from Equation 8 to the national total population $NT_P$. The average is the value used for $PE_{SA}$ in Equation 3, that area's control population for calculating its age and sex group population estimates with the Iterative Proportional Fitting method.

## GAPS IN CENSUS DATA; CATACLYSMIC EVENTS

### Gaps and Discrepancies in Census Data

The following are examples of how population bases for calculating subnational age and sex group population estimates may be prepared when there are gaps in the census data.

***Subnational Bases.*** As a rule, populations for the subnational areas from the latest censuses are used as the base numbers. However, latest censuses for some countries may provide populations for only the higher-order subnational areas, while populations for the lower-order subnational areas may only be found in previous censuses. If the boundaries of the lower-order subnational areas remained constant between the censuses, the populations for those areas from the previous census may be prorated to the next higher-order subnational area populations from the latest census.

There also are (rare) instances in which the census populations of the next lower-order administrative divisions of a subnational area do not aggregate to the reported census population of that area. To offset these discrepancies, either the populations of the divisions are prorated to the subnational area's census population, or the aggregate is used in place of the subnational area's census population.

***Age Groups by Sex.*** The census for a country may have 5-year age and sex group populations for the higher-order subnational areas, but only total male and female populations for the lower-order subnational areas. For calculating subnational age and sex group population estimates for that country using the Iterative Proportional Fitting method or that method combined with any of the trend methods (Shift-Share, Logistic Growth Rate, or their averages), the age and sex group distributions from the higher-order subnational areas may be adapted for the lower-order subnational areas in them at the time of the census.

***Broad Age Groups.*** A census may have provided, e.g., populations for the 5-year age groups of 65-69 and 70-74 for a country's first-order administrative divisions (provinces), but populations only for the 10-year age group of 65-74 for its second-order administrative divisions (districts). In this case, the 65-74 age group populations for the districts may be divided into the 65-69 and 70-74 groups by interpolation, based on the proportions of the two 5-year age group numbers for their provinces. Thus:

For the 65-69 group:

$$\mathbf{P_2(65\text{-}69) = P_2(65\text{-}74) * P_1(65\text{-}69) / P_1(65\text{-}74)} \qquad (9)$$

Where:

$P_2$ = age group population for a district
$P_1$ = age group population for the province that contains that district

The value for **$P_2$(65-69)** is rounded to the nearest whole number.

For the 70-74 group, it is by subtraction:

$$\mathbf{P_2(70\text{-}74) = P_2(65\text{-}74) - P_2(65\text{-}69)} \qquad (10)$$

If the country's census provides 10-year age groups, 5-year age groups can be interpolated into them based on the IDB 5-year age group proportions for that country for that census year.

***Open-ended Age Groups.*** If a census has the open-ended age groups of 75 years and over, those groups can be divided into the age groups 75-79 and 80 years-and-over proportionally, based on the IDB national numbers for those age groups for that census year.

***Unstated Age Groups.*** Some national censuses provide populations for groups whose ages are *unknown* or *unstated.* In preparing the census population base for calculating the subnational age and sex group population estimates for one or more target years, the *unstated* age group numbers for each subnational area are distributed proportionally to that area's 5-year and open-ended age group populations by sex.

***Urban-Rural Population Age Group Distributions by Sex****.* A national census that does not provide subnational 5-year age group populations by sex may provide an urban and rural breakdown of these population groups for the country. If the census also provides urban and rural total populations for the subnational areas, the national urban and rural age and sex group distributions may be adapted for the areas, using the areas' total populations for controls, including by sex when known. When there are no urban and rural breakdowns for the subnational area census populations, each area may be designated *urban* or *rural* based on an empirically determined "urban" population density threshold.[3]

***Trend Methods.*** A latest national census may provide populations for a country's provinces and districts, while the previous census populations only for the provinces. In this case, the province population controls for the target date are calculated using a trend method, and the district controls are calculated by prorating the district populations from the latest census to the province controls. The same is done when district boundaries between censuses were changed while the province boundaries remained intact, and equivalent populations for the new districts from the earlier census are not available.

### Cataclysmic Events

The impacts of cataclysmic events such as earthquakes or tsunamis are factored into the calculations for the subnational age and sex group population estimates for the affected areas. This is contingent on obtaining the numbers of fatalities and displacements ascertained from authoritative sources, as well as numbers reflecting the trajectories of the recoveries of those areas.

## DIGITAL MAPS AND POPULATION LINKAGE

### Linkage of Population Estimates to Digital Maps

The age and sex group population estimates for the subnational administrative areas are linked (joined) to the polygons for those areas in a digital map of the country in a GIS. As linked, these estimates must aggregate to the IDB national projected totals for that country.

[3] For example, an area may be designated *urban* if its population density exceeds a threshold of 500 people per square kilometer.

However, the correspondence between the subnational areas reported in a census and the polygons in the digital map is not always one-to-one. The map may be of a vintage that has preexisting areas that were split into new areas prior to the latest census. It becomes necessary to fit records from the census to the polygons in the map.

For a polygon that contains two or more areas, a record with the aggregates of their 5-year age and sex group population estimates is appended to the table, which then can be linked to the polygon. For a new area that was split from parts of two or more preexisting areas, it typically is not known how the new area's population is distributed in the preexisting areas. Here, the age and sex group estimates for the new area are divided by the number of the preexisting areas – the "parts" of the new area. The resulting age and sex group estimates for each part are combined with the estimates for the remainder of the preexisting area from which it was split, and a record with the combined numbers is appended for linkage to the corresponding polygon. The age and sex group estimates are in nearest whole numbers. The estimates for all but one of the parts of the new area are decremented from the area's total age and sex group estimates to yield estimates for the remaining part without the effects of uneven division.

## Global Population Map

The Geographic Studies Branch assembled digital maps of the countries of the world with their subnational divisions into one seamless global map in a GIS for its Global Population Map. The digital international boundaries of the global map were aligned to those from the "Large Scale International Boundary Lines and World Vector Shorelines" polygon file from the U.S. Department of State, Office of the Geographer and Global Issues. Subnational population estimates by age and sex calculated with the methods presented in this document were joined to the global map. The preparations of the country digital maps for and the assemblage of the Global Population Map are described by Fitzwater (2013).

# REFERENCES

Arriaga, Eduardo E., Peter D. Johnson, and Ellen Jamison. 1994. *Population Analysis with Microcomputers: Volume I*. U.S. Census Bureau, International Programs Center: Washington, DC.

Fitzwater, John Thomas. 2013. "Methods Used for the Development of a Global Population Map." U.S. Census Bureau, Population Division, Geographic Studies Branch: Washington, DC.

George, M.V., Stanley K. Smith, David A. Swanson, and Jeff Tayman. 2004. "Population Projections." Chapter 21 in *The Methods and Materials of Demography*. Second Edition. Edited by Jacob S. Siegel and David A. Swanson. San Diego, CA: Elsevier Academic Press.

Judson, D.H. and Carole L. Popoff. 2004. "Selected General Methods." Appendix C in *The Methods and Materials of Demography*. Second Edition. Edited by Jacob S. Siegel and David A. Swanson. San Diego, CA: Elsevier Academic Press.