Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR TECHNICAL REPORT 2016-002
NOVEMBER 2016

R programs for decomposing changes in life expectancy variance within and between country groups

Vladimir M. Shkolnikov (shkolnikov@demogr.mpg.de)
Dmitri A. Jdanov (jdanov@demogr.mpg.de)
Sergey Timonin (timonin@hse.ru)

For additional material see www.demogr.mpg.de/tr/

R programs for decomposing changes in life expectancy variance within and between country groups

Vladimir M. Shkolnikov (shkolnikov@demogr.mpg.de)
Dmitri A. Jdanov (jdanov@demogr.mpg.de)
Sergey Timonin (timonin@hse.ru)

Abstract. In this technical report, we present an application of the stepwise replacement algorithm for the decomposition of a change in an aggregated index by age with an additional split by mortality and population composition effects, and for the decomposition by age and country group. The method is detailed in Timonin et al. (2016). As the index to be decomposed, we use the variance in life expectancy at birth across countries. However, the scripts can be easily modified for the decomposition of other aggregate demographic measures calculated for a set of countries. In addition to the variance across the whole set of countries, we also calculate and decompose variances between and within groups of countries.

1. Background

The application of decomposition methods in the analysis of changes in demographic indicators is a popular demographic technique. The algorithm of stepwise replacement is one of the most universal tools of decomposition, and can be used for almost any demographic analysis. However, its realization may vary significantly depending on the task. The main goal of decomposition by age is to present changes in function *F*; e.g., between times *t* and *T*, as a sum of age-specific components:

$$\Delta = F(T) - F(t) = \sum_x \Delta_x \qquad (1)$$

where *x* denotes the age and Δ^x represents the contribution of a single age group to the total change. This decomposition can be extended to other dimensions; e.g., cause of death or population groups (Shkolnikov et al. 2001). The universal stepwise replacement algorithm for decomposition by age and cause of death realized as an Excel worksheet with VBA elements was published as an MPIDR Technical Report (Andreev and Shkolnikov, 2012). In this technical report, we presented the realization of the stepwise replacement algorithm for the decomposition of the variance in life expectancy across several countries by age with a split by mortality and population-composition effects, and for the decomposition by age and country (or country group) for a set of countries. The method is detailed in Timonin et al. (2016). We used the population-weighted variance in life expectancy at birth, but the scripts can be easily modified for the composition of other demographic functions. In addition to the total variance, which measures variation due to all potential factors, we also calculate and decompose the between- and within-group variances that measure variation in the factors used for defining groups, and variation caused by all factors acting within the single group, respectively. In the next section, we briefly

describe the method introduced in Timonin et al. (2016). In the sections that follow we provide information on the usage of the scripts.

2. Description of the algorithm

Below we provide just a brief description of the algorithm determined by the R scripts. We tried to include only the details that are necessary to provide the reader with a general understanding of the processes used in calculating and interpreting the results. A more detailed description of the algorithm can be found in Timonin et al. (2016).

2.1. Calculation of target indicators

Consider a set of country-specific (i=1, 2,…, n) mortality data split by age x. These data are placed in matrix of death rates $\mathbf{M}(t) = \left[m_{x,i}(t)\right]$ and by the matrix of population exposures $\mathbf{P}(t) = \left[p_{x,i}(t)\right]$ for time point t. The population-weighted average life expectancy at birth for the entire set of countries is calculated as a weighted sum of country specific life expectancies $e_0^i(t)$:

$$\overline{e_0(t)} = \sum_i \pi_i(t) e_0^i(t), \tag{2}$$

where $\pi_i(t)$ are the population weights

$$\pi_i(t) = \frac{\sum_x p_{x,i}(t)}{\sum_i \sum_x p_{x,i}(t)}. \tag{3}$$

The population-weighted cross-country variance and the standard deviation are calculated as follows:

$$Var(t) = \sum_i \pi_i(t)\left[e_0^i(t) - \overline{e_0(t)}\right]^2 \tag{4}$$

and

$$StD\ (t) = \sqrt{Var(t)}\,. \tag{5}$$

Let us assume that the whole set of countries is divided into *K* groups using certain criteria (for example, established market economies, central and eastern Europe, former Soviet republics). The total cross-country variance can be split into variances within and between country groups:

$$Var(t) = Var^{bg}(t) + Var^{wg}(t). \tag{6}$$

The between-group variance is calculated in manner similar to the approach used in (4), but using mean values for groups of countries instead of single countries:

$$Var^{bg}(t) = \sum_{g=1}^{K} w_g(t)\left[\overline{e_0^g(t)} - \overline{e_0(t)}\right]^2, \tag{7}$$

where g denotes the country group, w_g is the population weight of group g in the whole set of countries; i.e., the sum of all country-weights for countries belonging to group g:

$$w_g = \sum_{j\,\in\, group\ g} \pi_j\ , \tag{8}$$

and $\overline{e_0^g(t)}$ is the weighted average life expectancy in group g:

$$\overline{e_0^g(t)} = \sum_{j\,\in group\ g} \pi_j^g e_0^j(t), \qquad \pi_j^g = \frac{\pi_j}{w_g}. \tag{9}$$

The within-group variance is calculated as the weighted sum of the variances within all country groups:

$$Var^{wg} = \sum_{g} w_g \sum_{j\,\in\, group\ g} \pi_j^g(t)\left[e_0^j(t) - \overline{e_0^g(t)}\right]^2\ . \tag{10}$$

2.2. Decomposition

Two types of decomposition are realized in this technical report. The first type further splits age-specific components determined by equation (1) into mortality and population composition parts (M- and P-effects):

$$\Delta^x = \Delta_M^x + \Delta_P^x\ . \tag{11}$$

The first term in (11) expresses changes in the index function produced by changes in the country-specific mortality rates at age *x*. The second term is the change in the index function

produced by changes in the population composition (the population distribution by country) at age *x*.

The second decomposition allows us to estimate the effect of the country groups on changes in the age-specific component. As in the first decomposition, here each age component Δ^x from equation (1) should be presented as the sum of the group- and the age-specific components:

$$\Delta^x = \sum_g \Delta_g^x . \qquad (12)$$

Both decompositions were realized as the implementation of the stepwise replacement algorithm (Andreev et al. 2002). A detailed formal description of the realization is provided in Timonin et al. (2016).

3 Requirements

We tested scripts using R version 3.1. The calculation of the group effects requires the use of the package *combinat,* which can be installed from the standard R repository (CRAN). There are no additional requirements for standard R installation.

4 Usage

The zip file included in the technical report contains three R scripts and the DATA folder. This folder contains the input data used in the examples. R scripts contain the following functions:

fun.r – function for the calculation of the target indicator (standard deviation). The function can be modified for the computation of average (across countries) life expectancy or another quantity.

decomp.r – decomposition by age and the P- and M- effects within each age

decomp_groups.r – decomposition by age and the country groups within each age

Input data for the scripts includes country- and age-specific death rates and population estimates by five-year age groups (0, 1-4, 5-9, 10-14, 15-19, …, 90+). The last age group can be 85+, 90+, or 95+.

4.1 R functions

The following functions are available after loading all three scripts into R workspace (e.g., using source command):

decomp() - decomposition with P- and M-effects

Usage: *decomp(pop_file = "pop.csv", mx_file = "mx.csv", groups_file = "groups.csv", ages=c(0, 15, 65, 111), years = c(1984, 1994), sex = "Male")*

Arguments:
pop_file – file with population estimates (the format of the file is described below).
mx_file – file with age-specific death rates.
groups_file – file with the definition of the country groups.
ages – vector that defines the age groups for the age-specific effects. By default the (*ages=c(0, 15, 65, 111)*) decomposition is performed by the age groups 0-14, 15-64, 65+.
years – vector of two elements; defines the time points for the decomposition.
sex – string variable with the value "Male" or "Female." The variable is used to select data from the data files and to calculate life expectancy (Coale-Demeny formula).

Example: *decomp(years = c(1990, 2000))*

Returns the data frame in the following format:
columns – overall standard deviation, standard deviation between groups, standard deviation within groups.
rows – standard deviation in the first year, standard deviation in the second year, age-specific components of the M-effect, age-specific components of the P-effect.

Example of the output:

```
                total       between       within
1984   4.0039804318  3.7581186646  1.381449746
1994   6.0927220550  5.8518043771  1.696363042
0.mx   0.0226475997  0.0369566601 -0.039677709
15.mx  1.5983875088  1.5785750588  0.311714115
65.mx  0.5037477957  0.5123529960  0.054038282
0.px  -0.0072616225 -0.0044308105 -0.009684624
15.px -0.0296756419 -0.0301259603 -0.003252078
65.px  0.0008959834  0.0003577685  0.001775310
```

In addition, the data frame is stored in file *res.csv*

decomp.g() - decomposition with country group effects.

Usage: *decomp(pop_file = "pop.csv", mx_file = "mx.csv", groups_file = "groups.csv", ages=c(0, 15, 65, 111), years = c(1984, 1994), sex = "Male")*

Arguments: see *decomp()*.

Example: *decomp(years = c(1990, 2000))*

Returns the data frame in the following format:
columns – overall standard deviation, standard deviation between groups, standard deviation within groups.

rows – standard deviation in the first year, standard deviation in the second year, age-specific components by group.

Example of the output:

```
                 total      between       within

1984     4.003980432  3.758118665  1.381449746
1994     6.092722055  5.851804377  1.696363042
0.EME    0.110411884  0.129335679 -0.039567957
0.CEE   -0.018000027 -0.017069720 -0.005711019
0.FSU   -0.077025880 -0.079740109 -0.004083357
15.EME   0.227005027  0.214237531  0.075110106
15.CEE   0.007242944  0.003805876  0.011785460
15.FSU   1.334463896  1.330405692  0.221566472
65.EME   0.337969130  0.345408323  0.031191808
65.CEE  -0.010550345 -0.017169358  0.018652314
65.FSU   0.177224994  0.184471799  0.005969470
```

In addition, the data frame is stored in file *res.csv.*

4.2 Format of the input data

Three files are used as input data for the calculations. These files contain age-specific death rates, population estimates, and the definition of the country groups. All of the files are comma-separated plain text files, with the first line in each file being a header. The file names are passed to the decomposition functions as arguments. The default file names are mx.csv, pop.csv, and groups.csv. All of the data files could include more data than is necessary for the calculations; e.g., data for other years or data for the other gender. Below we describe the variables in each data file. The current version of the functions for the calculation of life expectancy at birth (ex.per.abr) works with five-year age groups. For this reason, the aggregated five-year age groups with the first groups “below 1” and “1-4” should be used for the age-specific death rates and the population estimates.

File with age-specific death rates (default name mx.csv):

Variables:
Country – country name (string)
Year – year
Sex – Male or Female (string)
The next columns contain mortality rates by age group for a specific country, year, and sex. The number of age groups is not specified, but it should be the same for each of the countries in the dataset. The first column with age-specific rates should include data for the age group “below 1,” the second column should include data for the age group 1-4, and all of the following columns (except the last) should display data for five-year age groups. The last column contains data for an open age interval.

Example:

```
Country, Year,   Sex,p0,    p1,    p5,    p10,   ……,p90,   p95
AUS,     1970,  Male,0.0212,0.0010,0.0004,0.0003,……,0.2975,0.4334
AUS,     1971,  Male,0.0202,0.0009,0.0004,0.0004,……,0.2978,0.3805
……
USA,     2009,Female,0.0058,0.0002,0.0001,0.0001,……,0.1645,0.2881
```

Population estimates (default name pop.csv):

Variables:
Country – country name (string)
Year – year
Sex – Male or Female (string)
The next columns contain mortality rates by age group for a specific country, year, and sex. The number of age groups is not specified, but should be the same for the whole dataset. The first column with age-specific rates should include data for the age group "below 1," the second column should contain data for the age group 1-4, and all of the following columns (except the last) should include data for five-year groups. The last column should display data for an open age interval.

Example:

```
Country, Year,   Sex,p0,     p1,     p5,    ,…… ,p90,   p95
AUS,     1970,  Male,127698, 484993, 632281, …… , 3847 ,584
AUS,     1971,  Male,132740, 502793, 635373, …… , 3969, 628
……
USA,     2009,Female,1970155,7931658,9868478,…… ,989573,315609
```

Definition of the country groups:

group - group name (string); e.g., FSU
Country – country name (string)

Example:

```
group, Country
EME,   AUS
EME,   AUT
EME,   BEL
CEE,   BGR
FSU,   BLR
EME,   CAN
……
EME,   USA
```

5 Examples

As a supplement to this technical report, we provide the dataset used for the calculations in the paper by Timonin et al. (2016). The data are placed in the sub-folder DATA of the supplementary zip file. The sub-folder contains the following files: mx.csv and pop.csv with the mortality rates and the population estimates for all of the 37 countries used in the analysis for all of the available years between 1970 and 2010, and groups.csv with the definition of the country groups. We assigned each of the countries to one of three groups: established market economies (EME), central and eastern Europe (CEE), and the former Soviet Union (FSU). The mortality rates and the population estimates were taken from the Human Mortality Database (HMD, available at www.mortality.org).

In two examples below, we assume that the data files from the supplementary data (files from the DATA folder) are placed in the working directory. If this is not the case, the user should include the full path to the files in the file names and pass it to the function as respective arguments. We also assume that the package *combinat* is already installed, and that the R scripts from this technical report are loaded (sourced) into the R workspace.

Example 1. Age decomposition with M- and P-effects

Type the following command to decompose the change in the standard deviation between 1970 and 1984 by the age groups 0-14, 15-64, and 65+ and the P- and M-effects:

```
> r = decomp(years=c(1970,1984))
```

The result will be stored in in the data frame *r*

```
> r
              total       between        within
1970  2.0444352467  1.5320831868  1.3536752146
1984  4.0039804318  3.7581186646  1.3814497461
mx0   0.2633302326  0.3426022141 -0.0710841949
mx15  1.0947343367  1.2653177360 -0.0526488881
mx65  0.5947340864  0.6127633280  0.1470671885
px0   0.0121581054  0.0121456133  0.0040987290
px15 -0.0059501281 -0.0067586682 -0.0006536309
px65  0.0005385522 -0.0000347454  0.0009953280
```

The first two lines present the overall standard deviation and the standard deviation between and within the groups for the years 1970 and 1984. The next three lines are the age-specific M-effects by the age groups 0-14, 15-64, and 65+ for the overall standard deviation; the between-group standard deviation; and the within-group standard deviation

```
> r[3:5,]
          total   between      within
mx0  0.2633302 0.3426022 -0.07108419
mx15 1.0947343 1.2653177 -0.05264889
mx65 0.5947341 0.6127633  0.14706719
```

The last three lines present the age-specific P effects for the same age groups

```
> r[6:8,]
              total       between        within
px0   0.0121581054  0.0121456133  0.0040987290
px15 -0.0059501281 -0.0067586682 -0.0006536309
px65  0.0005385522 -0.0000347454  0.0009953280
```

The columns correspond to the overall standard deviation, the between-group standard deviation, and the within-group standard deviation, respectively. For example, the third element in the fifth row (*r[5,3]*) is the contribution of mortality change at age 65+ to the shift in the within-group standard deviation between 1984 and 1970.

Obviously, the sum of the column of all age-specific components is equal to the difference in the standard deviation:

```
> colSums(r[3:8,])
     total    between     within
1.95954519 2.22603548 0.02777453
> r[2,]-r[1,]
     total  between     within
1.959545 2.226035 0.02777453
```

Example 2. Age decomposition with country group effects

Type the following command to decompose the difference in the standard deviation between 1984 and 1970 by age groups and groups of country:

```
> r = decomp.g(years=c(1970,1984))
```

The result will be stored in in the data frame *r*

```
> r
               total      between       within
1970    2.044435247  1.532083187  1.353675215
1984    4.003980432  3.758118665  1.381449746
0.EME   0.396665294  0.495406211 -0.068244833
0.CEE  -0.041237067 -0.039118479 -0.012243650
0.FSU  -0.079939890 -0.101539904  0.013503018
15.EME  0.505763792  0.595389359 -0.045668063
15.CEE  0.046118135  0.033713724  0.035036931
15.FSU  0.536902281  0.629455985 -0.042671387
65.EME  0.489530634  0.474695949  0.173348861
65.CEE -0.002377414 -0.003763559  0.003352278
65.FSU  0.108119419  0.141796193 -0.028638623
```

The first two lines present the overall standard deviation and the standard deviation between and within the groups for the years 1970 and 1984. The next three lines show the group-specific components (EME, CEE, and FSU groups) for the age group 0-14:

```
> r[3:5,]
            total     between      within
0.EME  0.39666529  0.49540621 -0.06824483
0.CEE -0.04123707 -0.03911848 -0.01224365
0.FSU -0.07993989 -0.10153990  0.01350302
```

The next three rows are the age-specific components for the age group 15-64 by country group, and the last three lines show the effects for the age group 65+.

References

Andreev EM, Shkolnikov VM, Begun AZ. (2002) Algorithm for decomposition of differences between aggregate demographic measures and its application to life expectancies, healthy life expectancies, parity-progression ratios and total fertility rates. Demographic Research. 2002; 7(14): 499-522.

Andreev EM, Shkolnikov VM (2012) An Excel spreadsheet for the decomposition of a difference between two values of an aggregate demographic measure by stepwise replacement running from young to old ages. MPIDR *Technical Report* 2012-002 April 2012

Shkolnikov, V., Valkonen, T., Begun, A., Andreev, E. (2001). Measuring inter-group inequalities in length of life. Genus, LVII(3-4), 33-62.

Timonin, S., Shkolnikov, V.M., Jasilionis D., Grigoriev P., Jdanov, D.A., Leon, D.A.(2016) Disparities in length of life across developed countries: measuring and decomposing changes over time within and between country groups. *Population Health Metrics*.14:29. DOI 10.1186/s12963-016-0094-0