Distr.: General
06 November 2019

English only

Economic Commission for Europe

Conference of European Statisticians

Work Session on Demographic Projections

Belgrade, 25–27 November 2019
Item No. 4 of the provisional agenda
Methodology

Relational spline-model for interpolating demographic data and population projections

Note by Dalkhat M. Ediev[1], ediev@iiasa.ac.at

Summary

Deficiencies of demographic data pose a serious obstacle to demographic analysis and projection. Those deficiencies include age accumulation, age exaggeration, small population size and aggregated data among others. Often, those deficiencies allow obtaining reliable demographic data and building demographic projections in aggregate (abridged) tabulations only. Yet, the depth and usefulness of demographic analysis may benefit from increasing the age details of demographic tabulations. Although well-established interpolation techniques exist, they are typically limited in producing age profiles overly smoothed and often inconsistent with the original abridged tabulations. To address this issue, we suggest a relational model utilizing the traditional spline interpolation with transformed argument of the spline where the argument transformation relates the reconstructed age profile to a standard (reference) profile. The model is illustrated on empirical and projected (UN WPP 2019) population and mortality data. The method proposed may find applications in demographic reconstructions, data preparations for projections (especially in regional projections) as well as in interpolating projection results. The model and its extensions may also be useful in bridging the emerging techniques of 'Big data' that produce rich, yet non-representative, patterns with conventional unbiased, yet non-detailed, data.

[1] North-Caucasian State Academy, Institute for Applied Mathematics and Information Technologies, Russia. Wittgenstein Centre for Demography and Global Human Capital (IIASA, VID/ÖAW, WU), International Institute for Applied Systems Analysis, Austria. Lomonosov Moscow State University, Higher School of Modern Social Sciences, Demography Chair, Russia.
The research leading to these results has received funding from the Russian Foundation for Basic Research Grant 18-01-00289 "Mathematical models and methods of correcting the distortions of the age structure and mortality rates of elderly population" and Grant 18-010-01169 "Demographic changes and economic growth".

I. Introduction

1. Deficiencies of demographic data pose a serious obstacle to demographic analysis and projection. Those deficiencies include age accumulation, age exaggeration, small population size and aggregated data among others. Often, those deficiencies allow obtaining reliable demographic data and building demographic projections in aggregate (abridged) tabulations only. UN, in particular, conducts demographic projections and publishes estimates for world countries in five-year long age intervals [1]. Yet, the depth and usefulness of demographic analysis may benefit from increasing the age details of demographic tabulations. Analysis of population ageing, of cohort-period interactions, of specific age groups or cohorts may demand a detailed age information. Although well-established interpolation techniques exist (e.g. splines [1–3]), they are typically limited in producing age profiles overly smoothed and often inconsistent with the original abridged tabulations (an inconsistency feature undesirable if not prohibited in official publications). To address this issue, we suggest a relational model utilizing the traditional spline interpolation with transformed argument of the spline where the argument transformation relates the reconstructed age profile to a standard (reference) profile. The model is illustrated on empirical and projected (UN WPP 2019) population and mortality data. The method proposed may find applications in demographic reconstructions, data preparations for projections (especially in regional projections) as well as in interpolating projection results. The model and its extensions may also be useful in bridging the emerging techniques of 'Big data' that produce rich, yet non-representative, patterns with conventional unbiased, yet non-detailed, data.

II. The relational spline-model: Interpolating aggregated population stock/flow data

2. Assume, we have at hands an abridged age profile of a population indicator of interest P_y at aggregated age intervals $y = y_1, y_2, y_3, \dots$ for which there exists a detailed reference (standard) pattern P_x^* in single years of age $x = 0,1,2, \dots$. The indicator of interest may be either a population stock or a flow measured in numbers of people (population, deaths, births, migrations, marriages, etc.)[2]. The reference need not be fully consistent with the original aggregated numbers P_y, although the better the correspondence between the two, the more accurate will the method results be. Its (reference's) overall age pattern may deviate substantially from the age pattern of the modelled indicator. Yet, P_x^* should represent essential *short-term* variations in the profile to be reconstructed (P_x as we denote it). In interpolating population age composition, for example, one may use as a reference:

- either the age profile of a population close in its reproduction history to the population of interest as may be judged, e.g., a national population or population of a similar region with better data (such as, region of larger population size when it comes to studying small-area populations);
- or population composition obtained by a simplified forward/backward population projection from a year when the age details of the studied population are better known;
- population projected from the births' sequence in the past years (an important option for countries where census data are undermined by age heaping, age exaggeration etc., but time series of full or partial counts of births exist);
- or even the time series of births in the past not corrected for (unknown) survival and migrations.

[2] Similar ideas are also applicable to rates, although a rate might better be interpolated by a more nuanced procedure described further down in the paper in the context of mortality interpolation.

3. We start calculations in the model by, first, abridging the reference profile in the same way as the data to be interpolated:

$$P_y^* = \sum_{x \in y} P_x^*, y = y_1, y_2, y_3, \ldots, \quad (1)$$

Based on (abridged) observed and reference schedules, we construct cumulated subtotals starting from the minimum age:

$$cumP_y^* = P_{y1}^* + P_{y2}^* + P_{y3}^* + \cdots + P_y^*, \quad (2)$$

and

$$cumP_y = P_{y1} + P_{y2} + P_{y3} + \cdots + P_y. \quad (3)$$

4. The transformed (cumulated) schedules are used to build a monotone interpolation spline function $S()$ that, essentially, serves as a relational model between schedules (2) and (3):

$$cumP_y = S\left(cumP_y^*\right), y = y_1, y_2, y_3, \ldots. \quad (4)$$

The spline function, in turn, is used to produce interpolated cumulated subtotals:

$$cumP_x = S(cumP_x^*),\ x = 0,1,2, \ldots, \quad (5)$$

where cumulated subtotals $cumD_x^*$ for the standard may be obtained thanks to its available age details. Once the cumulated interpolations (5) are ready, they are used to produce interpolations of the original data:

$$P_x = cumP_{x+1} - cumP_x, x = 0,1,2, \ldots. \quad (6)$$

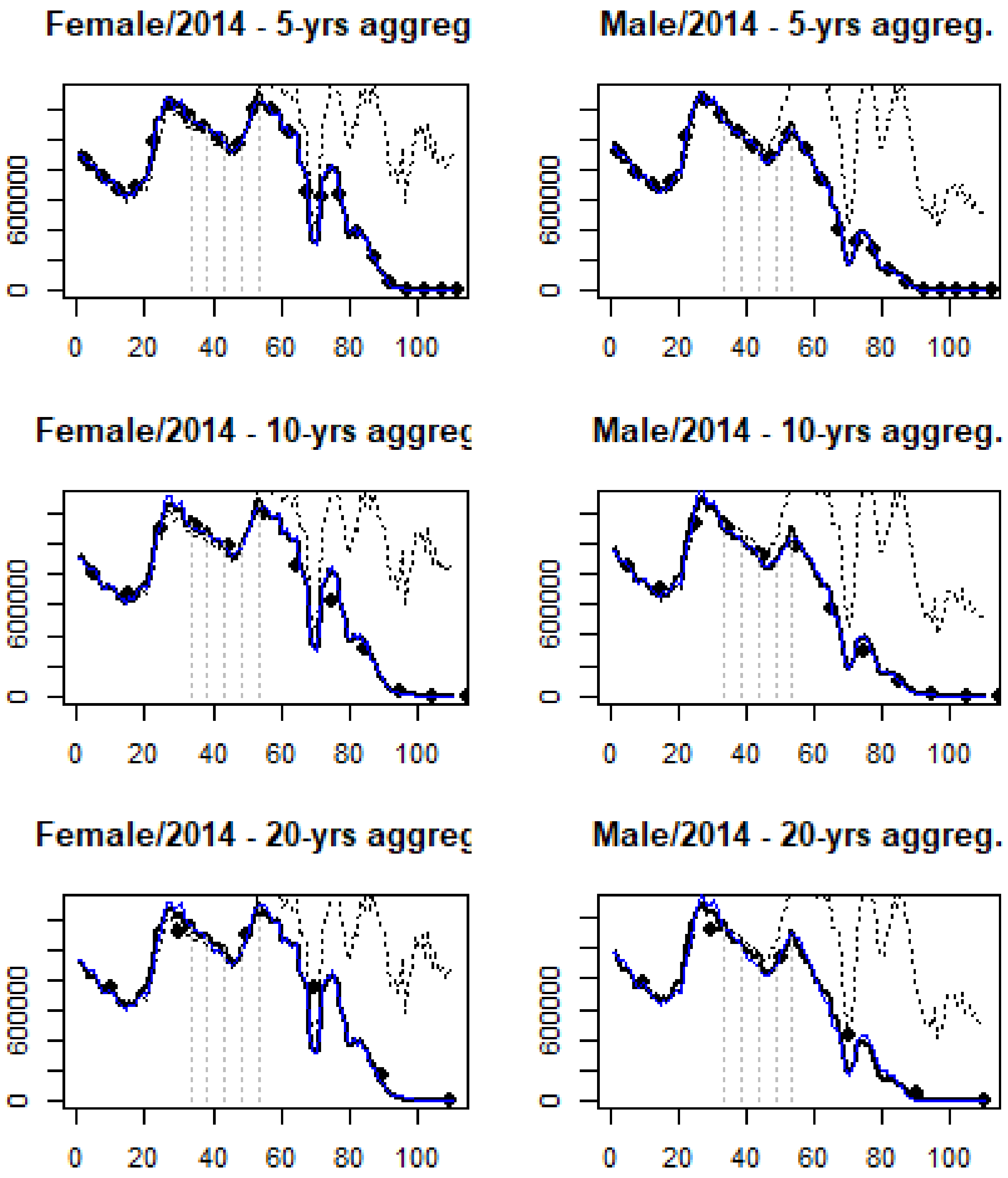
5. The constructed profile (6) will, on the one hand, be fully consistent with the initial data in that, being abridged, it will produce the original profile P_y exactly. On the other hand, within the age intervals y_1, y_2, y_3, ..., it will follow age details of the standard population. Usage of the splines will also assure continuity and smoothness of the reconstructed profile at junction points between the y-intervals.

6. The method is illustrated in Figure 1, where Russian data for 2014 [1] are, first abridged into 5-, 10-, and 20-years-long age intervals and then interpolated by the method proposed where the reference age profile is replaced by births' sequence in 110 years preceding 2014 (the earlier births' numbers, before 1959, are roughly estimated from 1959 population age composition by back-survival). Births in the past represent only a rough picture of current population age composition, because they do not account for migrations and differential survival that shaped the current population. Nonetheless, even such a remotely relevant reference combined with spline interpolation produces rather accurate interpolations. Interpolations are nearly perfect when based on data aggregated into 5-years-long age intervals. They are quite reasonable even when the initial data is aggregated into 20-years-long age groups.

7. Notably, the interpolated population structures in Figure 1 are free of signs of age heaping at Census 2010 (as may be judged from visible signs of heaping at ages 30, 35, 40, 45, and 50 in Census 2010) which points to usefulness of the proposed method in cleaning population data from age heaping effects.

Figure 1

8. Original data (thick black lines), abridged interpolation inputs (points), the reference schedule of births in the respective birth cohort (thin broken lines) and interpolation results (blue lines): Russian women (panes to the left), men (panes to the right) in 2014; with three levels of abridging: into five-years-long intervals (panes in the first row), ten-years-long intervals (the middle panes), and twenty-years-long age intervals (the panes in the bottom). The vertical broken lines represent ages 30,35,40,45, and 50 in the year of Census (2010) where the original data shows signs of age heaping that does not appear in interpolation results.



III. Interpolating aggregated demographic rates: with applications to mortality

9. Next, we illustrate usage of the proposed relational spline-model in interpolating demographic rates taking the death rates as example, although the method may also be used with other types of rates too.

10. Rates in aggregated age groups are, typically, available for countries with deficient demographic data. Same applies to populations of small size, where detailed mortality profiles would show too much of random fluctuations. Usual way to assess detailed mortality profile in such cases would rely on model life tables, mortality models, or relational models. Those methods, however, would not produce detailed demographic statistics consistent with the available (aggregated) official statistics in the sense that, being aggregated, they do not necessarily coincide with the original abridged data. Assuming the available aggregated mortality statistics is reliable, fitting the detailed estimates to it would presumably improve statistical accuracy of the estimates. For official statistical publications, internal consistency would, indeed, be a desired feature in itself. To achieve that goal, we apply the above-presented spline-based relational model.

11. Let M_y be the available death rates in aggregated age intervals $y = y_1, y_2, y_3, \ldots$, while M_x are the envisaged interpolated death rates in single years of age $x = 0,1,2, \ldots$. We assume that the detailed (possibly, interpolated) population estimates $P_x, x = 0,1,2, \ldots$ are already available for the population of interest (alternatively, one may use a standard, e.g., life table, population instead, as will be illustrated further down in the section). Assume further, M_x^* to be a reference mortality schedule with full age details. That profile may be selected (as we do here), for example, from a set of model life tables or, perhaps, from a national life table when regional population is studied. Here, we use extended set of UN model life tables [2,3] to determine the reference mortality profile. To this end, we find the best-fit reference by aggregating the model life tables:

$$M_y^* = \frac{\sum_{x \in y} P_x M_x^*}{\sum_{x \in y} P_x}, y = y_1, y_2, y_3, \ldots, \tag{7}$$

and finding the model life table that has the lowest sum of squared relative deviations from M_y.

12. Once the reference schedule is determined, we produce empirical and hypothetical reference distributions of deaths in the population of interest:

$$D_y = P_y M_y, D_y^* = P_y M_y^*. \tag{8}$$

Those are used to produce accumulated subtotals:

$$cumD_y^* = D_{y1}^* + D_{y2}^* + D_{y3}^* + \cdots + D_y^*,$$

$$cumD_y = D_{y1} + D_{y2} + D_{y3} + \cdots + D_y. \tag{9}$$

The cumulated subtotals are related to each other through a monotone spline function:

$$cumD_y = S(cumD_y^*), y = y_1, y_2, y_3, \ldots, \tag{10}$$

that is used to produce interpolated cumulated subtotals:

$$cumD_x = S(cumD_x^*),\ x = 0,1,2, \ldots \tag{11}$$

The interpolated subtotals, in turn, are used to produce interpolated deaths:

$$D_x = cumD_{x+1} - cumD_x, x = 0,1,2, \ldots \tag{12}$$

that, finally, lead to interpolated death rates:

$$M_x = \frac{D_x}{P_x}, x = 0,1,2, \ldots \tag{13}$$

13. As a numerical illustration to the described method, we consider death rates in five years-long age intervals for men in Burkina Faso in 1985-1990, for which population we have also produced age-heaping-cleaned smoothed population age profile (by applying an age-heaping model [4] not shown here). The aggregated death rates come from abridged life tables of UN WPP 2019 report [5]. Interpolation results are presented in Figure 2. In the other example of mortality interpolation (Figure 3), we present results for male death rates in Egypt in 2017 [6] based on reference life table population age composition instead of the actual population age composition. This example illustrates that the method proposed is successfully capturing both the empirical levels of (abridged) death rates and the reference mortality profile while keeping the interpolated profile smooth even when the actual population age composition is replaced by proper standard profile.

Figure 2

Abridged (dots) and interpolated deeath rates for Burkina Faso, males, 1985-1990. Broken line: the reference UN model life table.

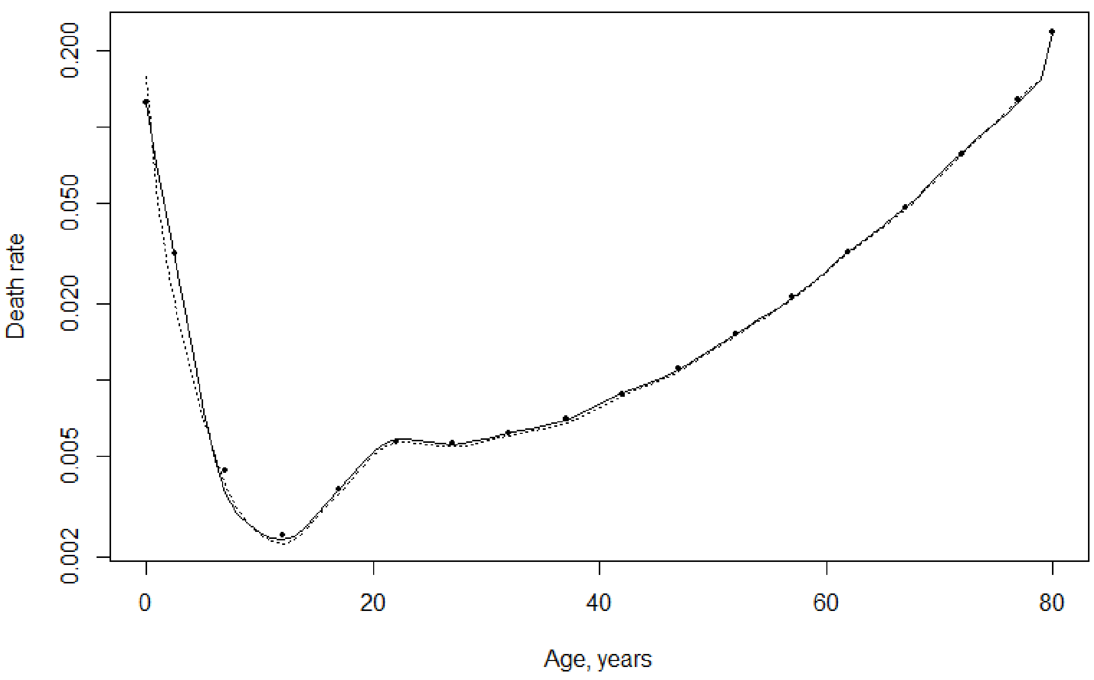
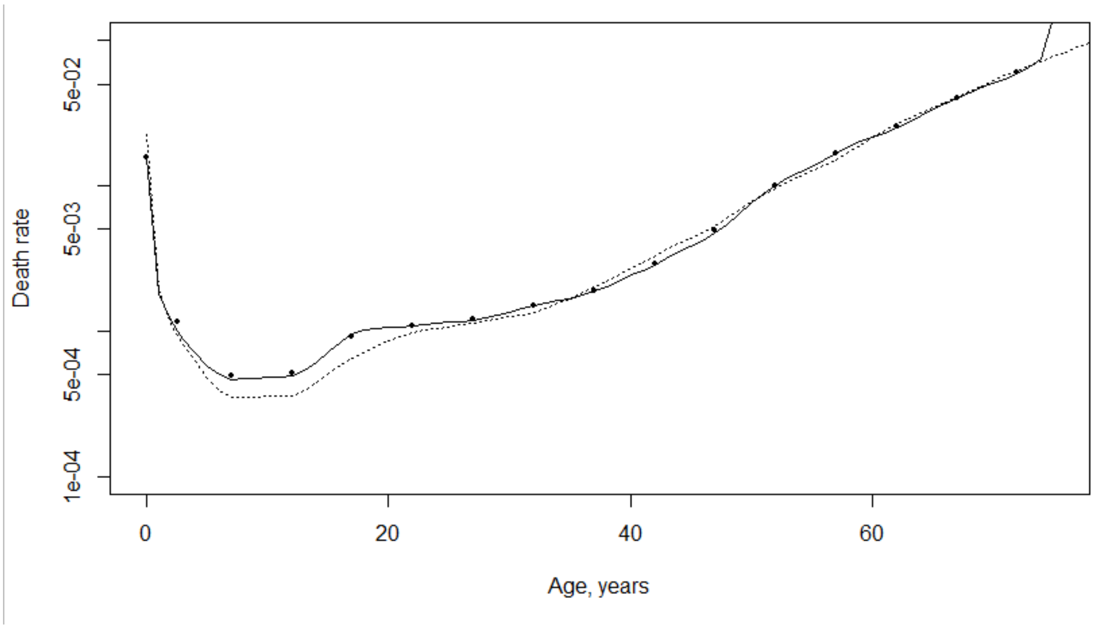


Figure 3

Abridged (dots) and interpolated deeath rates for Egypt, males, 2017. Broken line: the reference UN model life table.



IV. A full-scale interpolation exercise: extracting detailed profiles from UN WPP projections

14. In this section, we present an advanced example of application of the method to developing a detailed population projection based on abridged projection results. We use the UN World Population Prospects [5] that contains a consistent set of long-term population projections for world countries. Those projections are useful in studying long-term developments of social welfare and we aim at conducting such analysis for Russian Federation.

15. The Russian government has adopted a pension reform leading to the extension of the legal retirement ages (from 60 to 65 for men and from 55 to 60 for women by 2028) that has generated a far-reaching discussion – as well as a wide-spread opposition – in the Russian society. While demographic arguments are in the core of both the governmental proposal and the wider discussion, the discourse is complicated by lack of relevant demographic statistics. The available data sets do not allow computing a single cohort life table for Russia. In the meantime, the debate relies on guesstimates of individual survival prospects, remaining lifetimes and pension contributions/benefits. For the lack of objective statistics, discussants turn to period indicators as proxies to cohort ones and even go as far as to suggest the difference between the period life expectancy at birth and the age at retirement as a proxy for the (cohort) expectation of life at retirement. Here, we fill the gap by constructing (and extending) life tables for cohorts entering their 20s after 1959 and studying various implications of the former and newly proposed pensionable ages to those cohorts. Our aim is also to go beyond the basic demographics of pensions in Russia and discuss normative balances of payments to/from the pension system for the Russian cohorts, estimate rates of return for those payments, so that we could also study inter-generational actuarial fairness as well as comparative (dis)advantages of the funded and the pay-as-you-go versions of the system.

16. To conduct the sorts of analysis mentioned, we need a long time series of detailed population and mortality statistics. Unfortunately, such series in sufficient age detail is only available

for years after 1959, and the projections available in long term (UN WPP) are only provided in abridged from. (UN does produce interpolated numbers that, however, fall short in their detail and quality of data demands of our study.) For years 1926-1959 (with exception of the WWII years), there are estimates by Andreev, Darsky and Kharkova [7] that, however, are also in abridged from (in five-year-long age groups, ages 0 to 85+) that need interpolation (that is done by applying spline interpolations; war-period numbers are roughly replaced by linear interpolations for the population and log-linearly for the death rates).

17. We interpolate UN projections in a step-wise procedure, where we start with data for the last available year (2015) in the HMD and use it to obtain a reference population for interpolating UN projections. To this end, we, first, interpolate death rates from UN projection into single year of time and age and use them to project the last HMD population to the nearest projection year in UN projection set. (We use cohort component method with Rogers-Castro migration profile and births interpolated to single years of time from UN assumptions, details may be requested from the author.) The projected age profile of the population, then, is used as a reference profile to interpolate the UN projection results. Discrepancy between the reference and the interpolated populations allows improving the migration estimates and (iteratively) improving the reference and the interpolation. In the end, we obtain detailed population projection for the entire period between the reference year of HMD data and the UN projection year. Once that procedure is completed we move to the next UN projection year using the interpolation results of the previous step to produce reference population for the next step. An example of the interpolation result is presented in Figure 4, where we present both the interpolated age structure of Russian women in 2020, the UN’s own interpolation [5] (thin blue line) obtained by conventional spline interpolation as well as the most recent (year 2015) available age structure from HMD (broken line) shifted by five years to synchronize with the population age structure in 2020. The Russian case illustrates that the conventional interpolations while describing well the age structure at ages with gradual change in population numbers, may distort the age pattern dramatically when it comes to ages with rapid population change (cohorts in their 70s born in WWII period, for example, in Figure 4).

18. In addition to the interpolation of the available data into single years of age, we also extend all non-HMD mortality data until age 110 years by applying constrained mortality extrapolation method [8] based on the Kannisto’s logistic mortality model [9] and the *combined method* [Eq. (7), 10] utilizing both the conventional and Mitra’s [11] estimates of the remaining life expectancy in the open age interval. To extrapolate the population numbers to old age, we use the extrapolated death rates and stable population model [12].

19. Obtaining detailed long time series of population and mortality estimates enables us to conduct meaningful analysis of the long-term implications of the adopted pension reform. An example of such analysis is presented in Figure 5 where survival of Russian cohorts to the formal age at retirement is shown under the old and new retirement regimes. Figure 6 depicts development of PAYG pension system’s internal rate of return for cohorts of Russian men and women.

Figure 4

Interpolated age structure of Russian women in 2020 (the thick solid line), the UN's own interpolation [5] (the thin blue line) and the age structure as of 2015 (HMD data) shifted by five years to synchronize with the population age structure in 2020 (the broken green line).

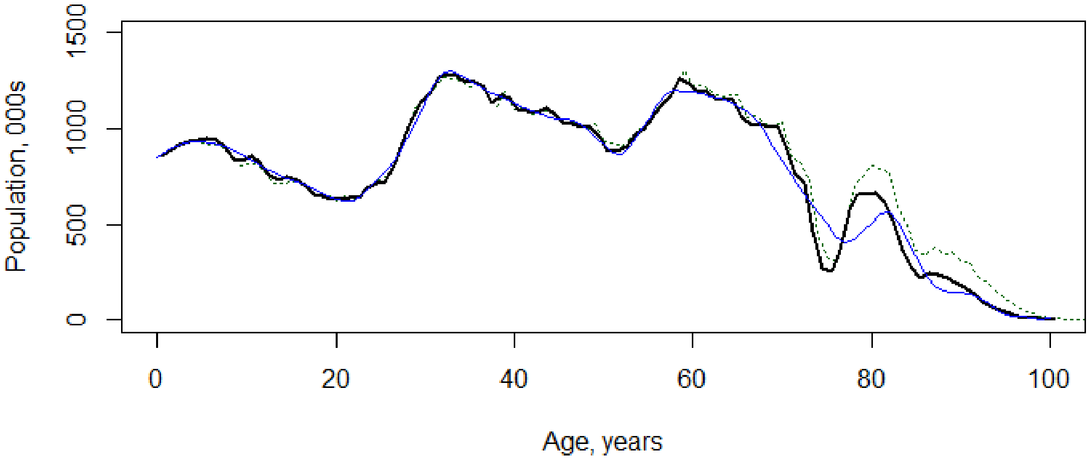


Figure 5

Survival (percent) of Russian cohorts of men (the pane to the left) and women (the pane to the left) to the old (the solid lines) and the newly introduced (the broken lines) formal ages at retirement.

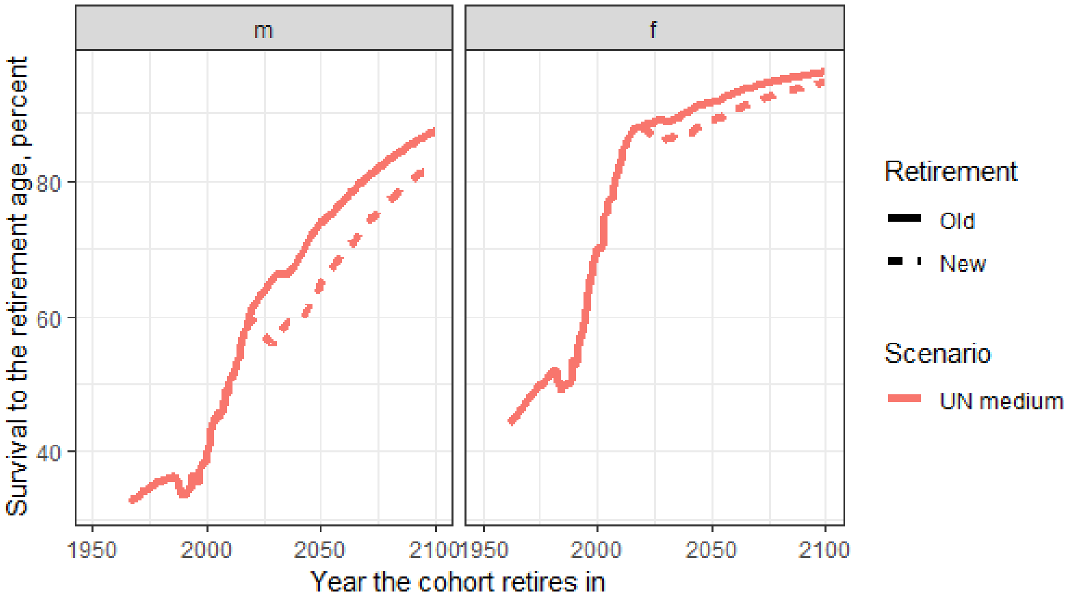
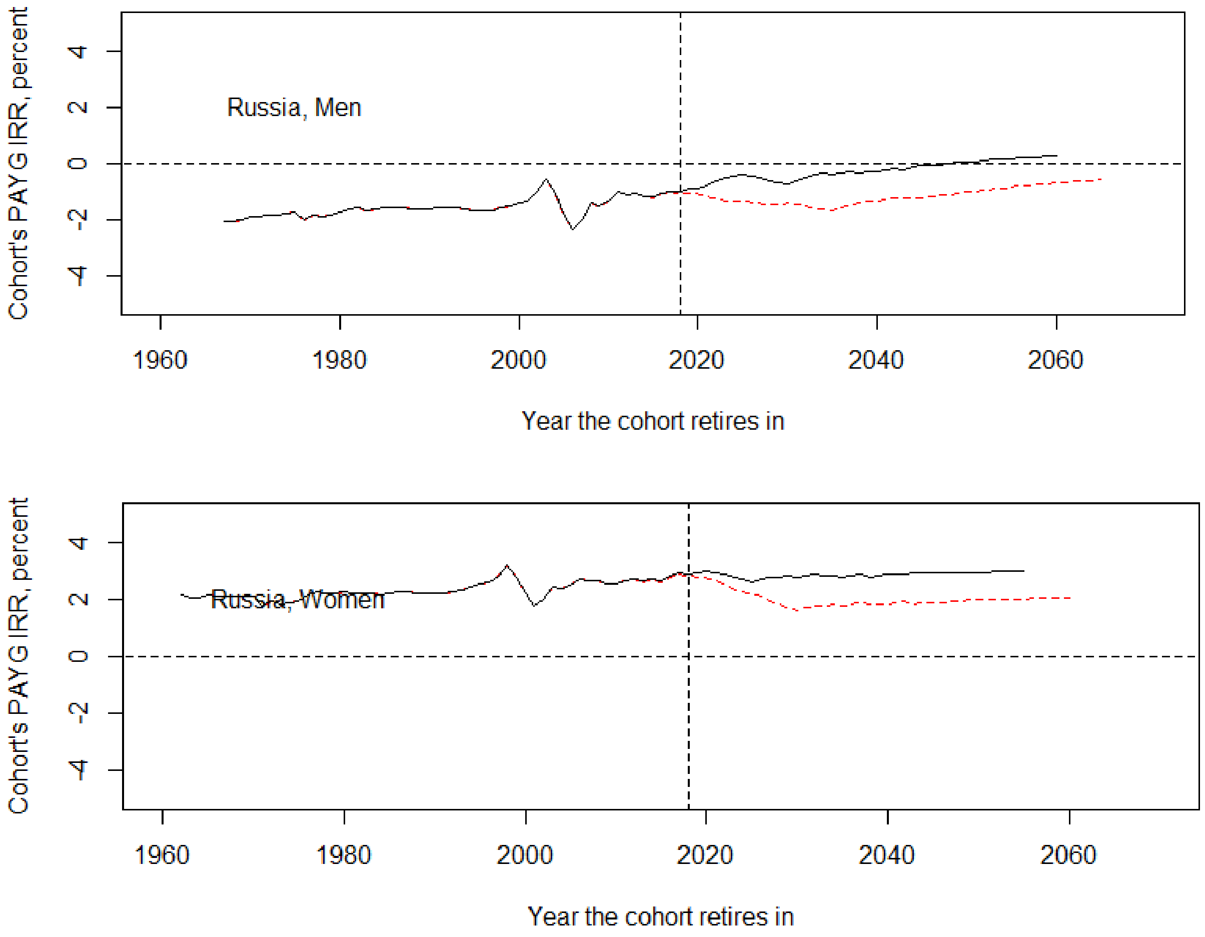


Figure 6

Internal rates of return of pay-as-you-go pension system in Russia for Russian male and female birth cohorts, percent. Horizontal axis: year when a birth cohort reaches formal age at retirement. Source: own calculations based on own interpolations of the UN WPP 2019 projection results.



V. Conclusion

20. Various application examples presented suggest efficiency of the interpolation model proposed. The model may find fruitful applications in decomposing data into single years of age/time whenever the detailed data is either not available or not reliable (due to age heaping or small population size).

21. In improving data, the method could be useful in sub-national population projections when regional inputs for the projection are not well-established or subject to deficiencies, especially for smaller areas.

22. A particularly promising area of application of the method is in extending population details available from limited and, possibly, not fully representative, sources, such as surveys or ‘Big data’ exercises to the general population data for which the data are only available with a limited detail.

23. When population projection is conducted in an abridged format, the model proposed may be used to obtain a more detailed projection patterns in a way similar to those represented in illustrative examples in the paper.

References

[1] B. University of California, Max Planck Institute for Demographic Research (Rostock), Human Mortality Database. Online database sponsored by University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany), (2019). www.mortality.org (accessed May 15, 2018).

[2] UN DESA/Population Division, United Nations, Model Life Tables for Developing Countries, Popul. Stud. (NY). 77 (1982).

[3] UN DESA/Population Division, World Population Prospects: Model Life Tables, (2017). https://esa.un.org/unpd/wpp/Download/Other/MLT/ (accessed August 20, 2017).

[4] D.M. Ediev, Demographic Losses of Deported Soviet Peoples [Demograficheskiye Poteri Deportirovannykh Narodov SSSR], AGRUS, Stavropol, 2003.

[5] United Nations Population Division, World Population Prospects 2019, (2019). https://population.un.org/wpp/ (accessed September 29, 2019).

[6] United Nations Statistics Division, Demographic and Social Statistics, (2019). https://unstats.un.org/unsd/demographic/products/dyb/dybcens.htm (accessed September 28, 2019).

[7] E. Andreev, L.E. Darsky, T.L. Kharkova, Demographic history of Russia: 1927-1959 (in russian), Informatika, Moscow, 1998.

[8] D.M. Ediev, Constrained Mortality Extrapolation to Old Age: An Empirical Assessment, Eur. J. Popul. (2017). doi:10.1007/s10680-017-9434-4.

[9] A.R. Thatcher, V. Kannisto, J.W. Vaupel, The Force of Mortality at Ages 80-120. Monographs on Population Aging, Odense University Press, Odense, Denmark, 1998. http://www.demogr.mpg.de/Papers/Books/Monograph5/ForMort.htm.

[10] D.M. Ediev, Expectation of life at old age: revisiting Horiuchi-Coale and reconciling with Mitra, Genus. 74 (2018). doi:10.1186/s41118-018-0029-7.

[11] S. Mitra, Estimating the Expectation of Life at Older Ages, Popul. Stud. (NY). 38 (1984) 313–319. doi:10.2307/2174079.

[12] N. Keyfitz, H. Caswell, Applied mathematical demography., Springer, 2005.