Main draft

Option 1: Forecasting vital rates from demographic summary measures

Option 2: Demographic forecasting from summary measures

Carlo G. Camarda*[1] and José Manuel Aburto[2]

[1] *Institut national d'études démographiques (INED)*
[2] *Leverhulme Centre for Demographic Science, Department of Sociology and Nuffield College at University of Oxford; CPop at University of Southern Denmark*

Abstract

In population and actuarial sciences, time-trends of summary measures (such as life expectancy or the average number of children per woman) are easy to interpret and predict. Most summary measures are nonlinear functions of the vital rates, the key variable we usually want to estimate and forecast. Furthermore smooth outcomes of future age-specific vital rates are desirable. Therefore, optimization with nonlinear constraints in a smoothing setting is necessary. We propose a methodology that combines Sequential Quadratic Programming and a P-spline approach, allowing to forecast age-specific vital rates when future values of demographic summary measures are provided. We provide an application of the model on Italian mortality and Spanish fertility data.

Keywords: Vital rates forecast; Smoothing; Constrained nonlinear optimization; Summary measures.

*Corresponding author: carlo-giovanni.camarda@ined.fr
Address: 9 cours des Humanités, 93322 Aubervilliers - France

1 Introduction

Demographic forecasting is a key research element in population dynamics and in demography. Reliable estimates of future vital rates are important because they are essential for social and economic planning. Mortality and fertility are often predicted by modelling and extrapolating rates over age and time (Bohk-Ewald et al., 2018; Booth, 2006). However, the high dimensionality, especially when using single years of age, is a major challenge for age-specific forecasting. Forecasting summary measures such as life expectancy at birth (e.g. Pascariu et al. (2018)) or the average number of children that a group of women will have over their reproductive lifespans (e.g. Alkema et al. (2011)) has the advantage of dealing with a single (or few) time series, but then the challenge is to derive reliable age-specific estimates from these forecasts.

In the past, a target summary measure (e.g. life expectancy at birth) was set and model life or fertility tables would provide the target age-specific vital rates (Booth, 2006). More recently, Ševčíková et al. (2016) derived full age specific vital rates consistent with the United Nations probabilistically projected values of life expectancy at birth and period total fertility rate (TFR) (Gerland et al., 2014). To estimate age-specific death rates, they developed a methodology based on back engineering the Lee-Carter forecasting model (Lee and Carter, 1992). Subsequently, future age patterns of fertility, consistent with the projected TFR, were derived interpolating between a starting age pattern of fertility and a target model pattern. In more recent work, Pascariu et al. (2020) leveraged the high correlation between life expectancy and death rates across ages to estimate a full mortality profile from values of life expectancy at birth. These methodologies aim at deriving age-specific vital rates from one summary indicator. A key question is whether results from these models provide reliable estimates that are consistent with complementary indicators such as lifespan variation or the variation in age-specific fertility rates.

Previous evidence in mortality research suggests that standard methodologies based on forecasting age-specific death rates and rates of mortality improvements provide accurate values for life expectancy but not necessarily for measures of dispersion in ages at death (Bohk-Ewald et al., 2017). This implies that the shape of the force of mortality may not be consistent with past trends. To overcome these limitations, Camarda (2019) incorporates demographic knowledge into the forecasting model. Our model is akin to this perspective and to indirect demographic techniques. In fertility, information about other summary measures besides TFR such as the variance of reproductive outcomes is often neglected (Hruschka and Burger, 2016). A notable exception in fertility research was made by Thompson et al. (1989), who forecast the total fertility rate, the mean age at childbearing, and the standard deviation of the age at childbearing and retrieved age-specific fertility rates relying on the gamma distribution. Aiming at filling this gap in development of forecasting methodologies, we propose a non-parametric model to derive future mortality and fertility age-patterns complying with projected summary measures of central tendency (e.g. life expectancy and TFR) and dispersion (e.g. lifespan variation and variance in age at childbearing).

Unlike comparable approaches, we assume only smoothness of future vital rates, which is achieved by a two-dimensional P-spline approach as in Currie et al. (2004), and we allow constraints to multiple series of summary measures. One of the main challenges is that summary measures are commonly non-linear functions of the estimated penalized coefficients such that standard optimization techniques may not be effective (e.g. Lagrangian multipliers). To overcome this limitation, we introduce an algorithm based on Sequential Quadratic Programming (SQP) (Nocedal and Wright, 2006). SQP is suitable since it is one of the most

successful methods for solving non-linearly constrained optimization problems. We illustrate our approach with two datasets: 1) We forecast age-specific mortality of Italian females consistent with future life expectancy predicted by United Nations (2019) and a future trend of a lifespan variation measure obtained by time-series analysis; and 2) We forecast age-specific fertility for Spain constrained to future values of total fertility rates, mean and variance of age at childbearing, derived by time-series analysis.

2 Data

We use data on death counts and population estimates from the World Population Prospects for Italian females from 1960-1965 to the latest period 2010-2015 available (United Nations, 2019). To test our model we use the projected life expectancy to 2050. For the fertility application, we use data from the World Population Prospects for the total fertility rate (TFR) (United Nations, 2019), and age-specific births and population from the Human Fertility Database for Spanish females (The Human Fertility Database, 2019).

3 Method

3.1 Model on Italian mortality data

The model is formulated with the mortality example to ease presentation, but the same process applies with fertility data. Supposed that deaths and exposures to risk are arranged in two matrices, $\boldsymbol{Y} = (y_{ij})$ and $\boldsymbol{E} = (e_{ij})$, both with dimension $m \times n_1$, whose rows and columns are classified by age $\boldsymbol{a}$, $m \times 1$, and period (year) $\boldsymbol{t}_1$, $n_1 \times 1$, respectively. As with previous work, we assume that the number of deaths y_{ij} at age i in year j is Poisson distributed with mean $\mu_{ij}\, e_{ij}$ (Camarda, 2019). In this case μ_{ij} is the force of mortality often approximated by age-specific death rates. The aim of our model is to construct reliable trends in μ_{ij} for n_2 future years, $\boldsymbol{y}_2, n_2 \times 1$, based on projected summary measures, such as life expectancy at birth (e_0) and lifespan variation ($e^\dagger$).

Figure 1 (top-left panel) presents the observed values of e_0 for Italian females from 1960 to 2016 along with the UN's projected medium variant up to 2050. The top-right panel shows the observed values of $e^\dagger$ from 1960 to 2016. Values from 2017 to 2050 for $e^\dagger$ were obtained with conventional time-series models. Future mortality patterns, both by age and over time, must adhere to these predicted trends.

We arrange data as a column vector, that is, $\boldsymbol{y} = \mathtt{vec}(\boldsymbol{Y})$ and $\boldsymbol{e} = \mathtt{vec}(\boldsymbol{E})$ and we model our Poisson death counts as follows: $\ln(E(\boldsymbol{y})) = \ln(\boldsymbol{e}) + \boldsymbol{\eta} = \ln(\boldsymbol{e}) + \boldsymbol{B}\,\boldsymbol{\alpha}$, where $\boldsymbol{B}$ is the regression matrix over the two dimensions: $\boldsymbol{B} = \boldsymbol{I}_{n_1} \otimes \boldsymbol{B}_a$, with $\boldsymbol{B}_a \in \mathbb{R}^{m \times k_a}$. Over time, we employ an identity matrix of dimension n_1 because we will incorporate a constraint for each year. Over age, $\boldsymbol{B}_a$ includes a specialized coefficient for dealing with mortality at age 0. In order to forecast, data and bases are augmented as follows:

$$\breve{\boldsymbol{E}} = [\boldsymbol{E} : \boldsymbol{E}_2]\,, \qquad \breve{\boldsymbol{Y}} = [\boldsymbol{Y} : \boldsymbol{Y}_2]\,, \qquad \breve{\boldsymbol{B}} = \boldsymbol{I}_{n_1+n_2} \otimes \boldsymbol{B}_a\,, \tag{1}$$

where $\boldsymbol{E}_2$ and $\boldsymbol{Y}_2$ are filled with arbitrary future values. If we define a weight matrix $\boldsymbol{V} = \mathrm{diag}(\mathtt{vec}(\boldsymbol{1}_{m \times n_1} : \boldsymbol{0}_{m \times n_2}))$, the coefficients vector $\boldsymbol{\alpha}$ can be estimated by a penalised version of the iteratively reweighted least squares algorithm:

$$(\breve{\boldsymbol{B}}'\boldsymbol{V}\tilde{\boldsymbol{W}}\breve{\boldsymbol{B}} + \boldsymbol{P})\tilde{\boldsymbol{\alpha}} = \breve{\boldsymbol{B}}'\boldsymbol{V}\tilde{\boldsymbol{W}}\tilde{\boldsymbol{z}}\,, \tag{2}$$

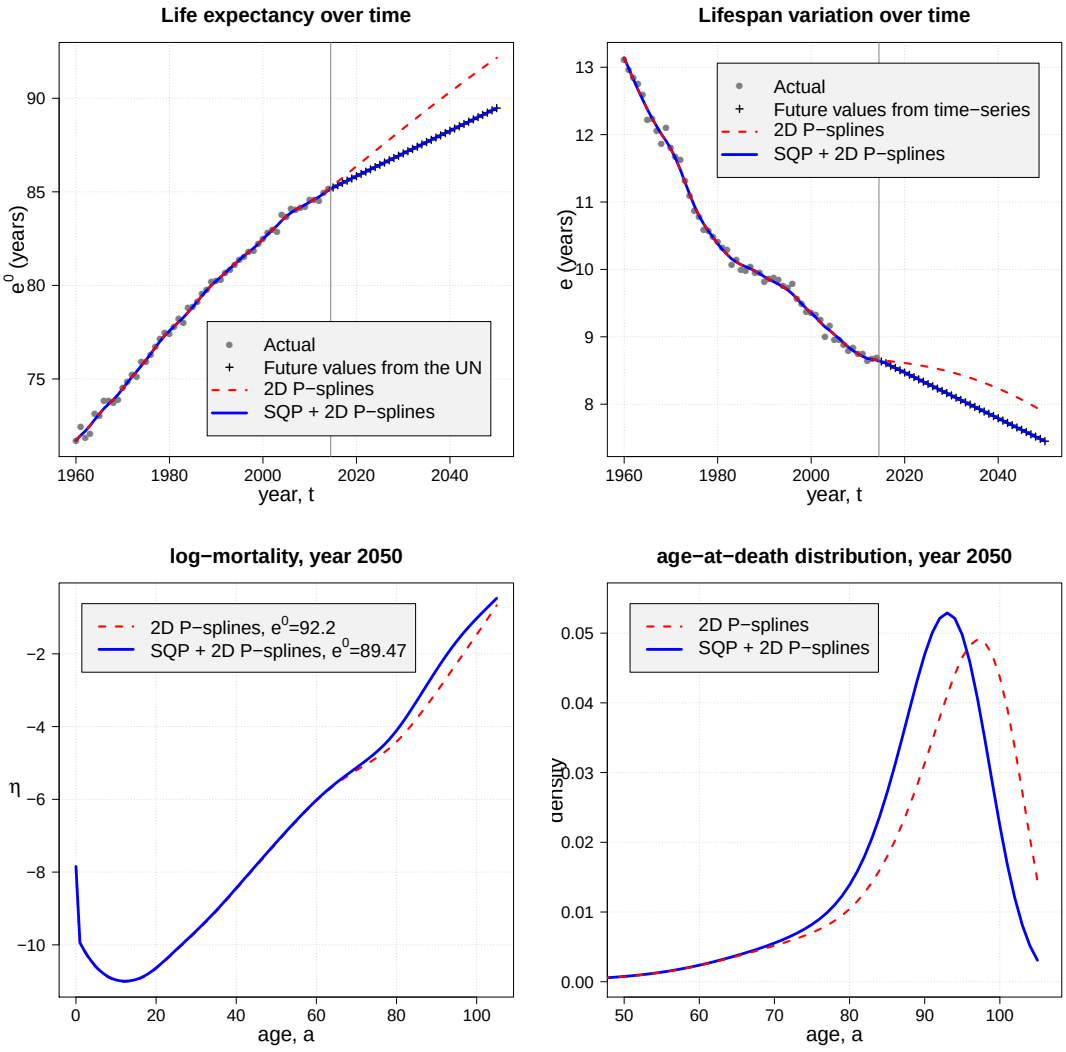

Figure 1: Top panels: Actual, estimated and forecast life expectancy at birth and lifespan disparity measure by United Nations and time-series, 2D P-splines and the SQP+2D P-splines. Bottom panels: Mortality in 2050 described by log-hazards and associated densities (ages 50+) by 2D P-splines and the SQP+2D P-splines. Italian females, ages 0-105, years 1960-2014, forecast up to 2050.

where a difference penalty $\boldsymbol{P}$ enforces smoothness behaviour of mortality both over age and time. Outcomes from this approach in terms of life expectancy and $e^{\dagger}$ are depicted with a dashed line in Figure 1 (top panels), and departures from the UN and time-series projected values are evident.

Both life expectancy and average years of life lost are nonlinear function of the coefficients vector $\boldsymbol{\alpha}$. For a year j and associated k_a coefficients $\boldsymbol{\alpha}_j$, we denote mortality by $\boldsymbol{\mu}_j = \exp(\boldsymbol{B}_a\boldsymbol{\alpha}_j)$. We can write our summary measures as follows

$$\begin{aligned} e^0(\boldsymbol{\alpha}_j) &= \mathbf{1}_m' \exp[\boldsymbol{C}\,\boldsymbol{\mu}_j] + 0.5 \\ e^{\dagger}(\boldsymbol{\alpha}_j) &= -\exp[\boldsymbol{C}\,\boldsymbol{\mu}_j]'\,\boldsymbol{C}\boldsymbol{\mu}_j \end{aligned} \tag{3}$$

where $\boldsymbol{C}$ is a $(m \times m)$ lower triangular matrix filled only with -1.

Constrained nonlinear optimization is therefore necessary and a SQP approach is implemented. Let denote with $\boldsymbol{N}^0$ and $\boldsymbol{N}^{\dagger}$ the $(k_a n_2 \times n_2)$ matrices with block-diagonal structures

containing derivatives of (3) with respect to $\boldsymbol{\alpha}_j$ for $j = n_1 + 1, \dots n_1 + n_2$:

$$\begin{aligned} \frac{\partial e^0(\boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}_j} &= \mathbf{1}'_m \mathtt{diag}[\exp(\boldsymbol{C}\boldsymbol{\mu}_j)]\, \boldsymbol{C}\, \mathtt{diag}(\boldsymbol{\mu}_j)\boldsymbol{B}_a \\ \frac{\partial e^\dagger(\boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}_j} &= -\boldsymbol{B}'_a \left\{\boldsymbol{C}'[\boldsymbol{C}\boldsymbol{\mu}_j \circ \exp(\boldsymbol{C}\boldsymbol{\mu}_j)] \circ \boldsymbol{\mu}_j\right\} + \\ &\quad -\boldsymbol{B}'_a \left\{[\boldsymbol{C}' \exp(\boldsymbol{C}\boldsymbol{\mu}_j)] \circ \boldsymbol{\mu}_j\right\} , \end{aligned} \tag{4}$$

where $\circ$ represents element-wise multiplication. Target life expectancy and lifespan disparity for future years are given by n_2-vectors $\boldsymbol{e}^0_{\mathrm{T}}$ and $\boldsymbol{e}^\dagger_{\mathrm{T}}$.

Solution of the associated system of equations at the step $\nu + 1$ is given by

$$\begin{bmatrix} \boldsymbol{\alpha}_{\nu+1} \\ \boldsymbol{\omega}_{\nu+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{L}_\nu & : & \boldsymbol{H}^0_\nu & : & \boldsymbol{H}^\dagger_\nu \\ \boldsymbol{H}^{0T}_\nu & : & \mathbf{0}_{n_2\times n_2} & : & \mathbf{0}_{n_2\times n_2} \\ \boldsymbol{H}^{\dagger T}_\nu & : & \mathbf{0}_{n_2\times n_2} & : & \mathbf{0}_{n_2\times n_2} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{r}_\nu - \boldsymbol{L}_\nu\boldsymbol{\alpha}_\nu \\ \boldsymbol{e}^0_{\mathrm{T}} - \boldsymbol{e}^0(\boldsymbol{\alpha}_\nu) \\ \boldsymbol{e}^\dagger_{\mathrm{T}} - \boldsymbol{e}^\dagger(\boldsymbol{\alpha}_\nu) \end{bmatrix} , \tag{5}$$

where $\boldsymbol{L}$ and $\boldsymbol{r}$ are left- and right-hand-side of the system in (2), and matrices $\boldsymbol{H}^0 = [\mathbf{0}_{k_a n_1 \times n_2} : \boldsymbol{N}^0]'$ and $\boldsymbol{H}^\dagger = \left[\mathbf{0}_{k_a n_1 \times n_2} : \boldsymbol{N}^\dagger\right]'$. Vector of $\boldsymbol{\omega}$ denotes the current solution of the associated Lagrangian multipliers for both set of constraints.

Future values for e^0 and $e^\dagger$ forecast by the proposed method are exactly equal to the UN and time-series values (Figure 1, top panels). The bottom panels show the forecast mortality age-pattern in 2050: the shape obtained by the suggested approach is not a simple linear function of the plain P-splines outcome, and differences are evident by looking at the associated age-at-death distributions.

3.2 Spanish Fertility Data

We forecast Spanish fertility using three commonly-used summary measures: Total Fertility Rate describing average number of children per women in a given year, and mean and variance of childbearing age which measure fertility shape over age. In formulas:

$$\begin{aligned} TFR(\boldsymbol{\alpha}_j) &= \mathbf{1}'_m\, \boldsymbol{\mu}_j \\ MAB(\boldsymbol{\alpha}_j) &= \boldsymbol{\mu}'_j\, (\boldsymbol{a} + 0.5)\, /\, TFR(\boldsymbol{\alpha}_j) \\ VAB(\boldsymbol{\alpha}_j) &= \boldsymbol{\mu}'_j\, (\boldsymbol{a} + 0.5)^2\, /\, TFR(\boldsymbol{\alpha}_j) - MAB(\boldsymbol{\alpha}_j)^2 . \end{aligned} \tag{6}$$

We forecast trends of these measures by time-series analysis. We then smooth and constrain future fertility age-patterns to comply forecast values of (6) as in (5). Summary measures as well as fertility rates in 2050 are presented in Figure 2. Differences between proposed approach and plain 2D P-splines are clear. Whereas P-splines blindly extrapolate previous trends mainly accounting for the last observed years, the proposed approach enforces future age-patterns to adhere combinations of summary measures, guiding future fertility toward demographic meaningful trends.

4 Discussion

Appendix A

The Gini coefficient is an indicator of relative variation. It was originally proposed in Economics to measure income or wealth inequality and has been adopted in demography and

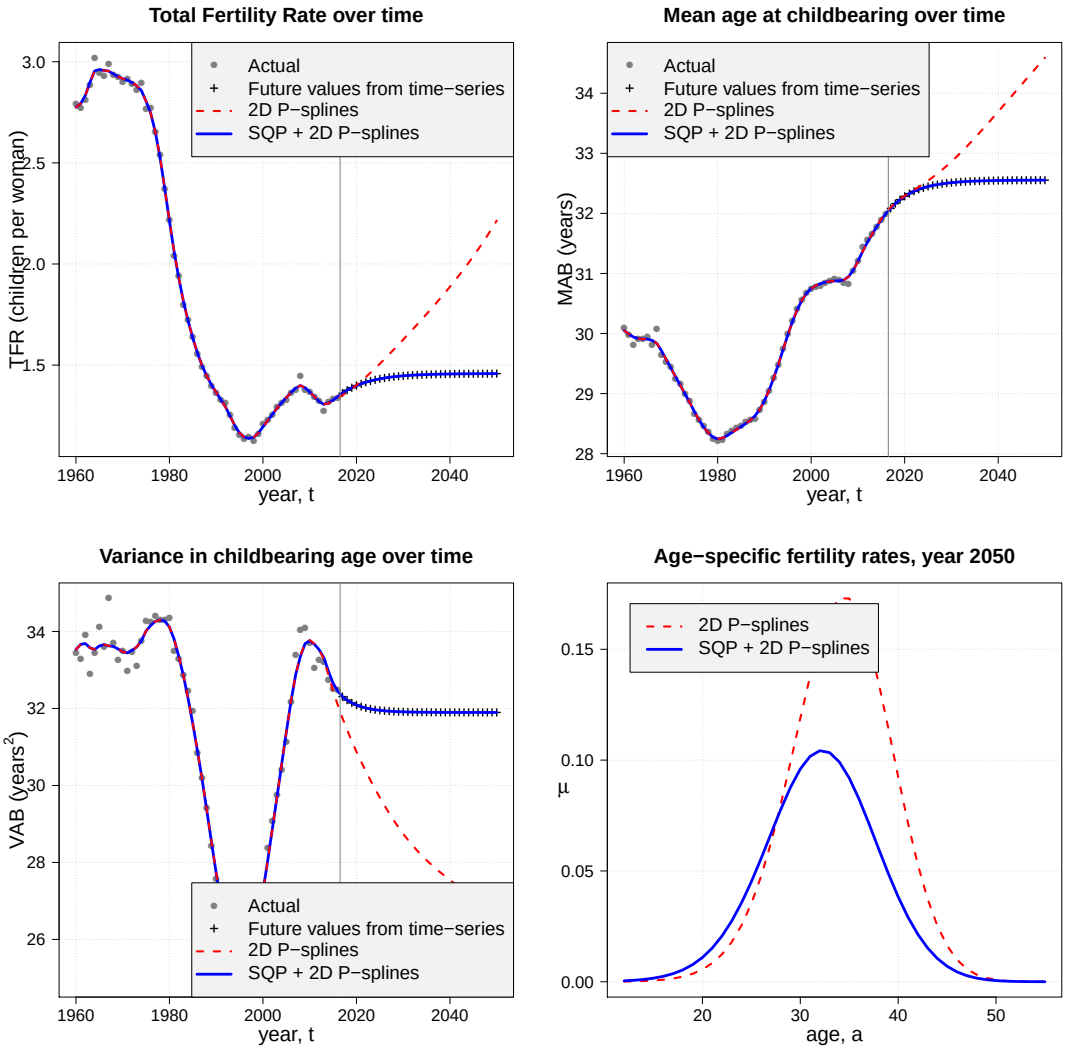

Figure 2: Top and left-bottom panels: Actual, estimated and forecast Total Fertility Rate, Mean and Variance in childbearing age by time-series analysis, 2D P-splines and the SQP+2D P-splines. Right-bottom panel: Age-specific fertility rate in 2050 by 2D P-splines and the SQP+2D P-splines. Spain, ages 12-55, years 1960-2016, forecast up to 2050.

survival analysis to measure lifespan variation (Bonetti et al., 2009; Gigliarano et al., 2017; Hanada, 1983; Shkolnikov et al., 2003) There exist several alternative and equivalent ways to define the Gini coefficient (Yitzhaki and Schechtman, 2013). For our purposes and the remainder of this article, we will use the following formulation:

$$G(\boldsymbol{\alpha}_j) = \mathbf{1}_m^T - \frac{\mathbf{1}_m^T \exp[2\,\boldsymbol{C}\,\boldsymbol{\mu}_j]}{e^0(\boldsymbol{\alpha}_j)}, \tag{7}$$

which is equivalent to $G = 1 - \frac{\int_0^\infty \ell(x)^2\,dx}{\int_0^\infty \ell(x)\,dx}$ (Hanada, 1983; Michetti and Dall'Aglio, 1957). Where $\int_0^\infty \ell(x)^2\,dx$ is the resulting life expectancy at birth of doubling the hazard at all ages. The Gini coefficient takes values between 0 and 1. A coefficient equal to 0 corresponds to the case of perfect equality in ages at death. The Gini index increases as lifespans become more spread and unequal in the population, reaching a value of 1 in the case of perfect inequality.

References

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F., Buettner, T., and Heilig, G. K. (2011). Probabilistic projections of the total fertility rate for all countries. *Demography*, 48(3):815–839.

Bohk-Ewald, C., Ebeling, M., and Rau, R. (2017). Lifespan disparity as an additional indicator for evaluating mortality forecasts. *Demography*, 54(4):1559–1577.

Bohk-Ewald, C., Li, P., and Myrskylä, M. (2018). Forecast accuracy hardly improves with method complexity when completing cohort fertility. *Proceedings of the National Academy of Sciences*, 115(37):9187–9192.

Bonetti, M., Gigliarano, C., and Muliere, P. (2009). The gini concentration test for survival data. *Lifetime Data Analysis*, 15(4):493–518.

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, 22(3):547–581.

Camarda, C. G. (2019). Smooth constrained mortality forecasting. *Demographic Research*, 41:1091–1130.

Currie, I. D., Durban, M., and Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical modelling*, 4(4):279–298.

Gerland, P., Raftery, A. E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B. K., Chunn, J., Lalic, N., Bay, G., Buettner, T., Heilig, G. K., and Wilmoth, J. (2014). World population stabilization unlikely this century. *Science*, 346(6206):234–237.

Gigliarano, C., Basellini, U., and Bonetti, M. (2017). Longevity and concentration in survival times: The log-scale-location family of failure time models. *Lifetime Data Analysis*, 23(2):254–274.

Hanada, K. (1983). A formula of gini's concentration ratio and its application to life tables. *Journal of Japanese Statistical Society*, 13(2):95–98.

Hruschka, D. J. and Burger, O. (2016). How does variance in fertility change over the demographic transition? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1692):20150155.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671.

Michetti, B. and Dall'Aglio, G. (1957). La differenza semplice media. *Statistica*, 7(2):159–255.

Nocedal, J. and Wright, S. J. (2006). Sequential quadratic programming. In *Numerical optimization*, pages 529–562. Springer.

Pascariu, M. D., Basellini, U., Aburto, J. M., and Canudas-Romo, V. (2020). The linear link: Deriving age-specific death rates from life expectancy. *Risks*, 8(4):109.

Pascariu, M. D., Canudas-Romo, V., and Vaupel, J. W. (2018). The double-gap life expectancy forecasting model. *Insurance: Mathematics and Economics*, 78:339–350.

Ševčíková, H., Li, N., Kantorová, V., Gerland, P., and Raftery, A. E. (2016). Age-specific mortality and fertility rates for probabilistic population projections. In *Dynamic demographic analysis*, pages 285–310. Springer.

Shkolnikov, V. M., Andreev, E. E., and Begun, A. Z. (2003). Gini coefficient as a life table function: Computation from discrete data, decomposition of differences and empirical examples. *Demographic Research*, 8(11):305–358.

The Human Fertility Database (2019). Max Planck Institute for Demographic Research and Vienna Institute of Demography. URL https://www.humanfertility.org.

Thompson, P. A., Bell, W. R., Long, J. F., and Miller, R. B. (1989). Multivariate time series projections of parameterized age-specific fertility rates. *Journal of the American Statistical Association*, 84(407):689–699.

United Nations (2019). *World population prospects: the 2019 revision*. United Nations.

Yitzhaki, S. and Schechtman, E. (2013). *The Gini Methodology*. Springer New York.