Forecasting vital rates from demographic summary measures

Carlo G. Camarda[1]

[1] Institut national d'études démographiques (INED), Paris, France

E-mail for correspondence: carlo-giovanni.camarda@ined.fr

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.

1 Introduction

Future mortality and fertility levels can be predicted either by modelling and extrapolating rates over age and time, or by forecasting summary measures, later converted into age-specific rates. The latter approach takes advantage of the prior knowledge that demographers and actuaries have on possible future values of measures such as life expectancy at birth and total fertility rate. Among others, this methodology has been lately adopted by the United Nations (Ševčíková et al., 2016). In this paper, we propose a model to derive future mortality and fertility age-patterns complying with projected summary measures. 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. Since these measures are

This paper was published as a part of the proceedings of the 34th International Workshop on Statistical Modelling (IWSM), Guimarães, Portugal, 7–12 July 2019.

commonly nonlinear functions of the estimated penalized coefficients, Lagrangian multipliers cannot be directly implemented. We hence opted for a Sequential Quadratic Programming (SQP) procedure (Nocedal & and Wright, 2006) to perform the associated constrained nonlinear optimization. We illustrate our approach with two data sets. We forecast mortality of Italian females, based on future life expectancy predicted by UN World Population Prospects (2017) and a future trend of a lifespan disparity measure obtained by time-series analysis. We also forecast Spanish fertility constrained to future values of total fertility rates, mean and variance of age at childbearing, derived by time-series analysis.

2 Model on Italian mortality data

For ease of presentation, we formulate the model on mortality data. We suppose that we have deaths, and exposures to risk, arranged in two matrices, $\boldsymbol{Y} = (y_{ij})$ and $\boldsymbol{E} = (e_{ij})$, each $m \times n_1$, whose rows and columns are classified by age at death, $\boldsymbol{a}$, $m \times 1$, and year of death, $\boldsymbol{t}_1$, $n_1 \times 1$, respectively. We assume that the number of deaths y_{ij} at age i in year j is Poisson distributed with mean $\mu_{ij}\, e_{ij}$. Forecasting aims to reconstruct trends in μ_{ij} for n_2 future years, $\boldsymbol{y}_2, n_2 \times 1$.
It is common practice to summarize mortality age-patterns by computing measures such as life expectancy at birth (e_0) and lifespan disparity measures. Time-trends of these summary measures are often regular and well-understood. Forecasting these time-series is therefore an easier task. Figure **??** (top-left panel) presents observed e_0 for Italian females from 1960 to 2016 along with the medium variant up to 2050 as computed by the UN. A second constraint is given by future values of $e^\dagger$, a lifespan disparity measure defined as the average years of life lost in a population attributable to death (Vaupel & Canudas Romo, 2003). Future values of this measure are obtained by conventional time-series models and portrayed in the top-right panel of Figure **??**. 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 $\boldsymbol{\alpha}$

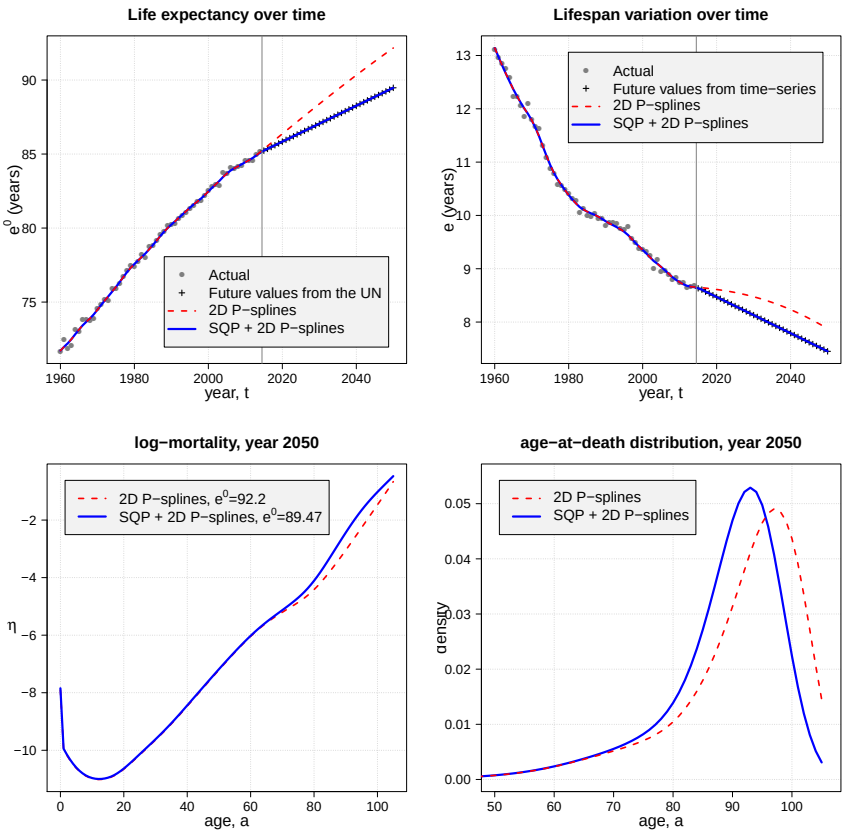

FIGURE 1. Top panels: Actual, estimated and forecast life expectancy at birth and lifespan disparity measure by United Nations and time-series, 2D P-splines and the SQP+2D P-splines. Bottom panels: Mortality in 2050 described by log–hazards and associated densities (ages 50+) by 2D P-splines and the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

can be estimated by a penalised version of the iteratively reweighted least squares algorithm:

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

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 **??** (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

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

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 (**??**) with respect to $\boldsymbol{\alpha}_j$ for $j = n_1 + 1, \ldots n_1 + n_2$:

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

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 (**??**), and matrices $\boldsymbol{H}^0 = \left[\mathbf{0}_{k_a n_1 \times n_2} : \boldsymbol{N}^0\right]^T$ and $\boldsymbol{H}^\dagger = \left[\mathbf{0}_{k_a n_1 \times n_2} : \boldsymbol{N}^\dagger\right]^T$. 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 **??**, 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 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:

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

We forecast trends of these measures by time-series analysis. We then smooth and constrain future fertility age-patterns to comply forecast values of (**??**) as in (**??**). Summary measures as well as fertility rates in 2050 are presented in Figure **??**. 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.

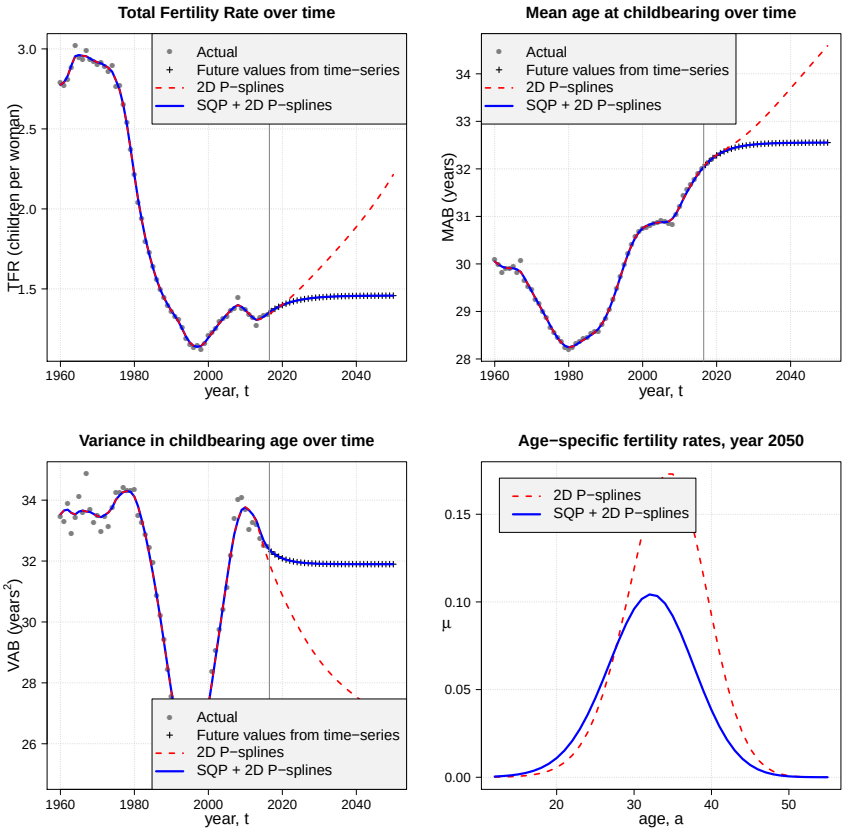

FIGURE 2. Top and left-bottom panels: Actual, estimated and forecast Total Fertility Rate, Mean and Variance in childbearing age by time-series analysis, 2D P-splines and the SQP+2D P-splines. Right-bottom panel: Age-specific fertility rate in 2050 by 2D P-splines and the SQP+2D P-splines. Spain, ages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

4 Concluding remarks

In this paper, we combine smoothing models (P-splines) and optimization with nonlinear constraints (Sequential Quadratic Programming) to forecast vital rates when future values of demographic summary measures are provided.
We envisage further applications. Forecast of vital rates for partially completed cohorts is often relevant in population studies. For instance, final fertility history of a given cohort may be hypothesized though age-pattern is not yet observed and its estimation will be necessary. We also plan to adopt our approach to reconstruct demographic scenarios which are conventionally based on summary measures.
From a methodological perspective, future work will be realized to incorporate uncertainty and to objectively select the amount of smoothness in future mortality and fertility age-patterns.

References

Currie, I. D. et al. (2004). Smoothing and Forecasting Mortality Rates. *Statistical Modelling*, **4**, 279-298.

Nocedal, J. & Wright, S. J. (2006). *Numerical Optimization*. Springer.

Ševčíková, H. et al. (2016). Age-Specific Mortality and Fertility Rates for Probabilistic Population Projections. In R. Schoen (Ed.), *Dynamic demographic analysis*, 285 – 310. Springer.

United Nations, Population Division (2017). *World Population Prospects: The 2017 Revision, Volume II*. ST/ESA/SER.A/400.

Vaupel & Canudas Romo (2003). Decomposing change in life expectancy: A bouquet of formulas in honor of Nathan Keyfitz's 90th birthday. *Demography*, **40**, 201-216.