DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 44, ARTICLE 45, PAGES 1085–1114
PUBLISHED 1 JUNE 2021
http://www.demographic-research.org/Volumes/Vol44/45/
DOI: 10.4054/DemRes.2021.44.45

Research Article

D-splines: Estimating rate schedules using high-dimensional splines with empirical demographic penalties

Carl Schmertmann

Contents

Demographic Research: Volume 44, Article 45
Research Article

D-splines: Estimating rate schedules using high-dimensional splines with empirical demographic penalties

Carl Schmertmann[1]

Abstract

BACKGROUND
High-dimensional parametric models with penalized likelihood functions strike a good balance between bias and variance for estimating continuous age schedules from large samples. The penalized spline (P-spline) approach is particularly useful for these purposes, but it in small samples it can often produce implausible age schedule estimates.

OBJECTIVE
I propose and evaluate a new type of P-spline model for estimating demographic rate schedules. These estimators, which I call D-splines, regularize and smooth high-dimensional splines by using demographic patterns rather than generic mathematical rules.

METHODS
I compare P-spline estimates of age-specific mortality rates to three alternative D-spline estimators, over a large number of simulated small populations with known rates. The penalties for the D-spline estimators are derived from patterns in the Human Mortality Database.

RESULTS
For mortality estimates in small populations, D-spline estimators generally have lower errors than standard P-splines.

CONCLUSIONS
Using penalties based on demographic information about patterns and variability in rate schedules improves P-spline estimators for small populations.

CONTRIBUTION
This paper expands demographers' toolkit by developing a new category of P-spline estimators that are more reliable for estimating mortality in small populations.

[1] Florida State University and the Center for Demography and Population Health, Florida, United States. Email: schmertmann@fsu.edu.

1. Introduction

1.1 Spline smoothing

A spline is a function $s(x)$ over an interval $x \in [L, H]$ that is constructed from piecewise polynomials and constrained to be smooth in some particular way. In many applications the polynomial pieces are cubic functions, and smoothness requires continuous levels $s(x)$, derivatives $s'(x)$, and second derivatives $s''(x)$ at all x values between L and H. Splines are valuable empirical tools for interpolation and regression problems because they have the flexibility of high-order polynomial functions, without some of the attendant risks of extreme and implausible fitted values between sample observations or outside of the sample.

There are many ways to parametrize spline functions, but the most useful for the purposes of this paper is the *cubic B-spline* system (Curry and Schoenberg 1947; de Boor 1976, 2001). With this parametrization the spline may be written as a weighted sum of *basis functions*

$$s(x) = \theta_1 B_1(x) + \theta_2 B_2(x) + \cdots + \theta_K B_K(x), \tag{1}$$

where $\theta \in \mathbb{R}^K$ and each $B_j(x)$ is a piecewise cubic function that is non-negative for all x and positive over only part of $[L, H]$. The number of parameters K is the *dimension* of the spline, which is often much larger than the number of parameters in a typical demographic model.

Figure 1 shows an example. Panel A contains $K = 36$ B-spline functions $B_j(x)$.[2] When weighted with any $\theta_1 \ldots \theta_{36}$ as in Equation (1), their sum will be a spline with piecewise cubic sections over intervals $[0, 3], [3, 6], \ldots [96, 99]$ and continuous levels, first derivatives, and second derivatives. Panel B of Figure 1 shows single-year age-specific mortality rates for Portuguese females over 1970–1979 from the Human Mortality Database (HMD) (University of California, Berkeley, and Max Planck Institute for Demographic Research, Germany 2014), and Panel C illustrates the best least-squares fit $s^*(x) = B\theta^*$ to the Portuguese log mortality rates using the specification in Equation (1). The fit in Panel C is constructed from 33 separate cubic polynomials – the first of which is used to determine the spline's value over $x \in [0, 3]$, the second for values over $x \in [3, 6]$, and so forth. Panel D of Figure 1 zooms in to show the first four of these piecewise cubic sections – i.e., those that determine the spline's value over ages 0 to 12.

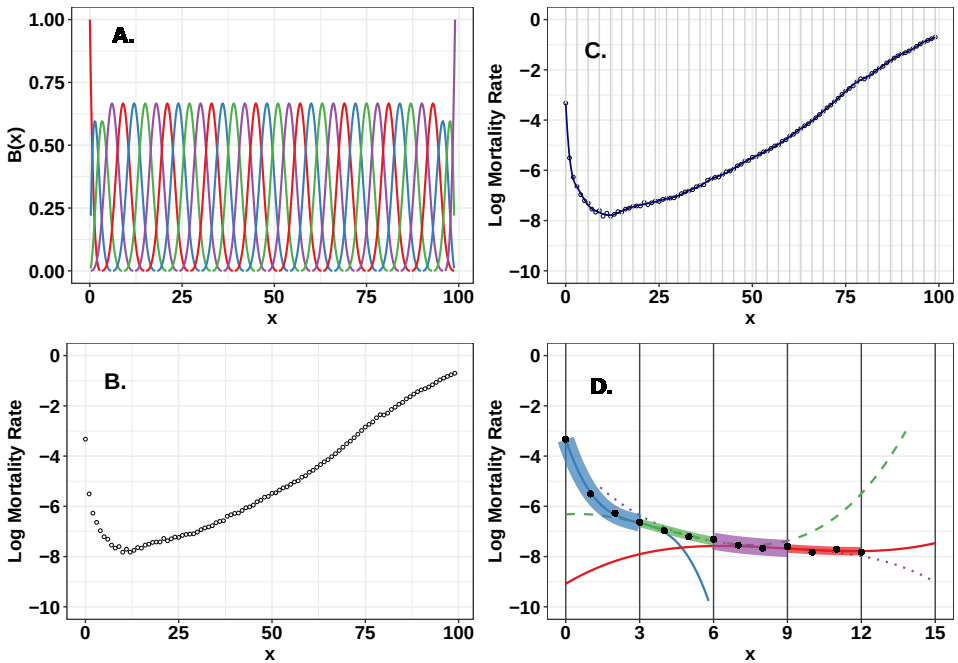
From Panel D of Figure 1 it might seem that assembling a smooth function from

[2] This particular set of cubic B-spline basis functions uses 32 interior knots at $x = 3, 6, \ldots, 96$ and boundary knots at $x = 0, 99$. It can be constructed in *R* with the command *splines::bs(0:99, knots=seq(from=3,to=96,by=3), degree=3,intercept=TRUE)*. I use this basis throughout the paper – *except* in the comparative experiments reported in Section 4, where I use slightly different basis functions in order to maintain complete consistency with P-spline estimators in the *MortalitySmooth* R package.

pieces of polynomial curves is complex, but using a B-spline basis makes fitting and interpreting a spline curve fairly simple. Because each basis function has a value of zero over most of the range of x, any one coefficient θ_j affects the spline values over only a limited range. These localized effects help to make numerical estimation stable and interpretable.

The extreme flexibility of spline functions permits excellent approximations to complex empirical functions, like mortality age schedules, that have no simple mathematical form. Such flexibility is valuable in cases like Figure 1C, for large national data sets with low sampling variability. It can become a liability in small-population settings, however, as discussed in Section 2.

Figure 1: A high-dimensional spline as an approximation to a national mortality schedule



Note: Panel A: Cubic B-spline basis functions $B_1(x) \ldots B_{36}(x)$. Panel B: Portugal 1970–1979 female log mortality rates. Panel C: Spline approximation with knots at $x = 0, 3, 6, \ldots 99$. Panel D: Log mortality rates and first four cubic polynomials used in the spline in Panel C, zoomed in for ages $x \in [0, 12]$.

1.2 Splines in applied demographic research

More than forty years ago, McNeill, Trussell, and Turner (1977) introduced splines to a demographic audience, by demonstrating how to interpolate detailed age-specific fertility rates from age-grouped data. They showed how fitting a piecewise quintic spline to cumulative fertility at ages 15, 20, ... 50 and then differentiating produces a set of single-year rates that exactly match the five-year input data, are smooth over age, and have desirable end-point properties at ages 15 and 50. They also illustrated the advantages of such a procedure over fitting a single high-order polynomial to the same data.

Fitting smooth rate schedules to vital rate data is still the major use of splines in demographic research. Very much in the spirit of McNeill, Trussell, and Turner (1977), for example, the methods protocols for both the Human Mortality Database (Wilmoth et al. 2007) and the Human Fertility Database (Jasilioniene et al. 2012) use spline functions to disaggregate and smooth data and fitted rate schedules by age.

Spline functions have also been proposed as flexible parametric models for fertility schedules, starting with investigations by Hoem et al. (1981). Schmertmann (2003) developed a quadratic spline system in which three ages (of menarche, maximum fertility, and half of maximum fertility) uniquely determine a continuous age-specific schedule. Using methods similar to those in this paper, Schmertmann (2014) proposed interpolating age-specific fertility rates from grouped data by using high-dimensional cubic splines that compromise between matching input data and matching empirical age patterns in the Human Fertility Database and other sources.

The TOPALS model for mortality and fertility projection and forecasting by de Beer (2012) is based on linear splines (i.e., sequences of connected straight lines, which are a particularly simple type of piecewise polynomial). TOPALS combines fixed fertility or mortality schedules with changing linear spline offsets to produce forecasts of future rate schedules. TOPALS has also been used as a regression model to fit small-area mortality schedules to sparse data in both frequentist and Bayesian models (Gonzaga and Schmertmann 2016, 2018; Rau and Schmertmann 2020).

Several efforts to produce coherent time series for mortality rates have used B-splines as estimators. These include the "B3" model for under-five mortality that has been adopted by the United Nations Inter-Agency Group for Mortality Estimation (Alkema and New 2014), and Alexander and Alkema's (2018) model for neonatal mortality trajectories.

Most importantly for the analysis in this paper, splines – and in particular the penalized B-splines described later in Section 2.3.1 – are a central component of recent methods for describing age patterns of mortality rates. These include the models estimated by the *MortalitySmooth* package in *R* (Camarda 2012), models that build mortality schedules from multiple overlapping splines (Camarda, Eilers, and Gampe 2016), the mortality projection models of de Jong and Tickle (2006) and Hilton et al. (2019) that generalize the a_x

and b_x vectors of the Lee–Carter (1992) forecasting approach with spline functions, and the presmoothing of rates before singular value decompositions in the forecasting model of Hyndman and Ullah (2007).

1.3 Purpose and organization of this paper

Here I evaluate a new type of penalized spline model for estimating demographic rate schedules, with a particular interest in applications to sparse data from small areas or small subpopulations. Using mortality data as an example throughout, I first describe the *P-spline* approach (Eilers and Marx 1996) used in many of the studies cited above. I then propose a variant of this approach that I call *D-splines*. D-splines regularize and smooth high-dimensional splines by using demographic knowledge derived from large demographic databases, rather than generic arithmetical rules.

The rest of the paper reports comparative results from tests of several alternative D-spline estimators on simulated small-area mortality data. These results are promising: D-spline estimators appear to have low errors and to produce schedules that reliably reflect known properties of human mortality schedules.

2. Likelihood and penalties

2.1 Notation: Spline function for a demographic schedule

A generic spline function for a demographic rate schedule over A single-year ages $x = 0 \ldots (A-1)$ is

$$s = \boldsymbol{B}\theta = \begin{bmatrix} B_1(0) & \cdots & B_K(0) \\ \vdots & \ddots & \vdots \\ B_1(A-1) & \cdots & B_K(A-1) \end{bmatrix} \theta = \begin{bmatrix} b_0'\,\theta \\ b_1'\,\theta \\ \vdots \\ b_{A-1}'\,\theta \end{bmatrix},$$

where $\boldsymbol{B}$ is a $A \times K$ matrix in which each of the K columns is a B-spline basis function (de Boor 2001) evaluated at ages $x \in 0 \ldots (A-1)$ and knots are closely-spaced, $\theta \in \mathbb{R}^K$ is a vector of parameters, and b_x' is the $1 \times K$ row vector of spline constants that affect the schedule's value at age x.

In principal the schedule $s = \boldsymbol{B}\theta$ could represent any kind of age-specific rates.

Throughout the rest of this paper I assume that s is a mortality schedule for $A = 100$ ages $x = 0 \ldots 99$, and I use the basis functions shown in Figure 1A. In all examples there are thus $K = 36$ B-spline basis functions constructed using interior knots at 32 ages $3, 6, \ldots, 96$, and $\boldsymbol{B}$ is a 100×36 matrix of known constants.

2.2 Data and likelihood

Suppose that we observe single-year age-specific death counts $(D_0 \ldots D_{A-1})$ for a small population, that person-years of exposure by age $(N_0 \ldots N_{A-1})$ are known, and that the schedule $s = \boldsymbol{B}\theta$ represents age-specific log mortality rates. If deaths at each age have independent Poisson distributions, $D_x \sim Poisson(N_x \exp(s_x))$, then the log likelihood for the parameter vector θ is

$$\begin{aligned} L(\theta) &= c - \sum_x N_x \exp(b_x'\theta) + \sum_x D_x\,(b_x'\theta) \\ &= c - \sum_x \hat{D}_x(\theta) + \sum_x D_x\,(b_x'\theta), \end{aligned} \tag{2}$$

where c is a scaling constant and $\hat{D}_x = N_x \exp(b_x'\theta)$ is the expected number of deaths at age x.

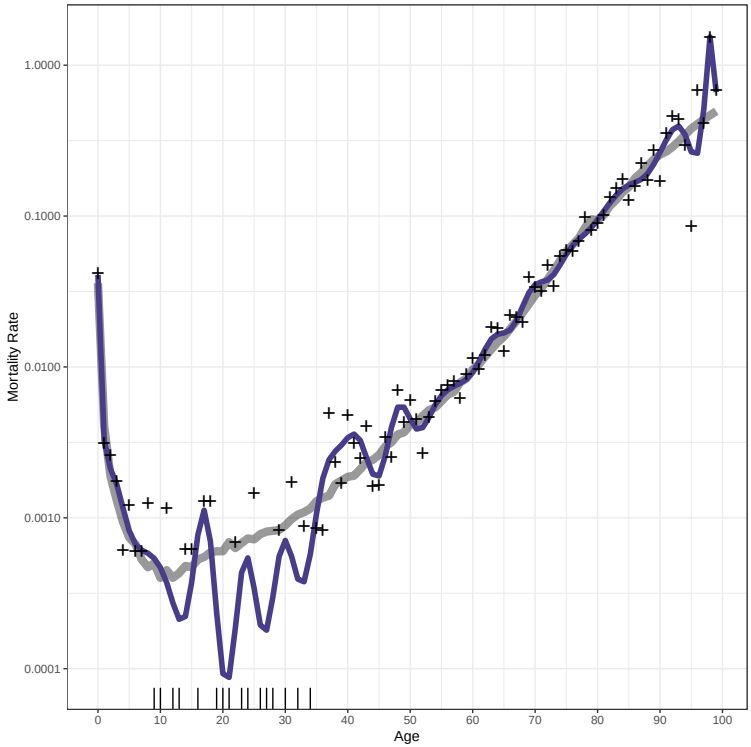
2.3 Penalized log likelihood

As an estimator, a high-dimensional spline function $s = \boldsymbol{B}\theta$ with many parameters is often quite vulnerable to sampling noise. For example, Figure 2 displays the maximum likelihood fit for $\theta \in \mathbb{R}^{36}$ using the 100×36 $\boldsymbol{B}$ matrix described above, with randomly generated age-specific data for female exposure and deaths in a population of 100,000 females that has the age structure and mortality rates of Portugal 1970–1979 from Figure 1B. In this simulated dataset, intended to represent a moderately large "small population" over the course of one year, there were 1016 deaths to 100,000 women.

It is evident from Figure 2 that a spline with 36 parameters overfits mortality rates, in the sense that the true schedule does not have erratic fluctuations over small age ranges like those in the fitted model. The roller-coaster pattern in estimated mortality rates over ages 10–50 is the most obvious example: adjusting the function up and down over this local age range matches the sample data well because there are small clusters of ages with zero deaths, but the spline produces inaccurate estimates of the mortality risks at those ages. This illustrates a classic bias-variance tradeoff: a high-dimensional spline function is flexible enough to represent small-scale features in the true rate schedule accurately

(low bias), but that same flexibility means that in a small-sample setting it may overinterpret coincidental features of sample data (high variance). A spline model correctly estimates the broad pattern of sharply decreasing mortality at young child ages followed by increases in adolescence and adulthood, but when compared to high-quality mortality schedules derived from large populations the spline has implausibly high curvature and non-monotonicity over some age ranges.

Figure 2: **Unpenalized spline $s = B\theta$ with 36 degrees of freedom: maximum likelihood fit for a simulated small-area sample with known mortality rates. True mortality rates, in grey, are HMD rates for Portuguese females 1970–1979**



Note: Points correspond to simulated death/exposure ratios in a population of 100,000 females with Portugal's 1970–1979 age distribution. Tick marks along the horizontal axis represent 16 single-year ages with no deaths. The erratic curve is the unpenalized spline fit that maximizes the sample likelihood.

2.3.1 P-spline penalties on parameters

A P-spline approach to estimation Eilers and Marx (1996) addresses the bias-variance problems evident in Figure 2 by including a penalty term in the log likelihood and then maximizing the penalized function.[3] For the purposes of this paper, the key feature of these penalties is that they apply to the vector of B-spline coefficients $\theta \in \mathbb{R}^K$ rather than to the estimated demographic schedule $s = \boldsymbol{B}\theta$.

In a P-spline approach we maximize a penalized log likelihood function

$$f(\theta \,|\, \lambda) \;=\; L\,(\theta) - \frac{\lambda}{2} \cdot \theta' \boldsymbol{\Delta}_p' \boldsymbol{\Delta}_p \theta, \tag{3}$$

where λ is a scalar penalty multiplier and $\boldsymbol{\Delta}_p$ is a $(K-p) \times K$ matrix of constants such that $\boldsymbol{\Delta}_p\theta$ is the vector of p^{th} differences in spline coefficients (Eilers and Marx 1996). If $p = 1$, for example, then $\boldsymbol{\Delta}_1\theta = (\theta_2 - \theta_1, \theta_3 - \theta_2, \ldots, \theta_K - \theta_{K-1})'$ represents the vector of differences between successive B-spline coefficients and the penalty is proportional to the sum of the squared differences $\sum_{j=2}^{K} (\theta_j - \theta_{j-1})^2$.

With equally-spaced knots for the spline basis functions in the columns of $\boldsymbol{B}$ the penalized log likelihood function (Equation 3) has higher values for demographic functions that fit the data *and* which approximate p^{th}-order polynomials. Maximizing the penalized function therefore requires a tradeoff between fitting the data and simplifying the fitted curve, where "simplifying" means choosing a smoother curve that looks more like a p^{th}-order polynomial. P-splines represent an elegant approach to the bias-variance tradeoffs for spline models illustrated in Figure 2. Importantly, they are the foundation of the *MortalitySmooth* package in *R* Camarda (2012).

A critical decision in the P-spline approach is the choice of the penalty multiplier $\lambda > 0$, which determines the tradeoff between fit and smoothness: higher values of λ increase the relative importance of smoothness when calculating the spline coefficients $\theta^*(\lambda)$ that maximize the penalized log likelihood. Following recommendations from Currie, Durban, and Eilers (2004), the default settings in many applications (and in the *MortalitySmooth* package) choose λ by searching sequentially over a discrete grid of possible values, finding the $\theta^*(\lambda)$ that maximizes Equation (3) for each candidate λ, and then selecting the λ value that minimizes the *Bayesian Information Criterion* (BIC, Schwarz 1978). For Poisson-distributed deaths in our example,

$$BIC(\lambda) \;=\; 2 \sum_{x=0}^{A-1} D_x \ln\left(\frac{D_x}{\hat{D}_x\left[\theta^*(\lambda)\right]}\right) \;+ \mathrm{df}(\lambda) \cdot \ln(A),$$

[3] This process is known in statistics and machine learning as *regularization* (Neumaier, 1998).

where df(λ) is a measure of the effective degrees of freedom of the fitted curve (see Eilers and Marx 1996; and this paper's Appendix). BIC is a scalar index that defines a specific tradeoff between model fit and parsimony: a simpler, smoother model (with higher λ and fewer degrees of freedom) produces a better BIC score only if the model's fit to observed mortality data does not worsen too much.

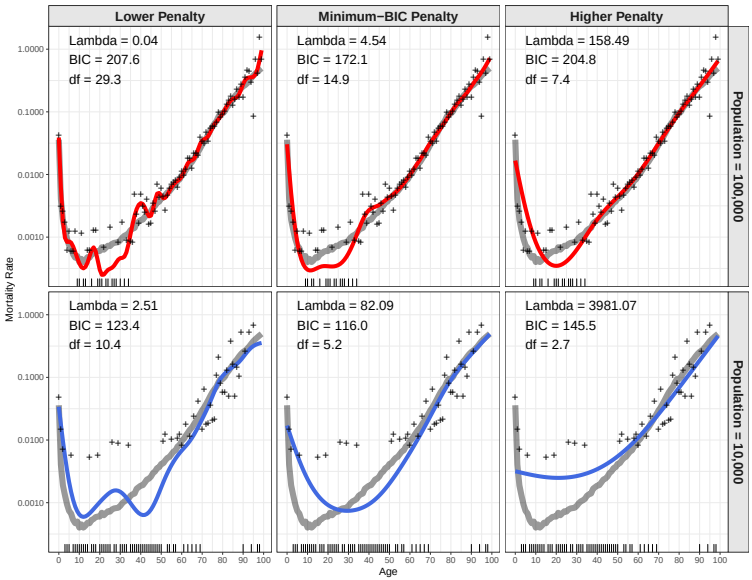
Figure 3 illustrates how penalized P-splines improve estimated mortality schedules for small populations. The top three panels contain the same data as in Figure 2, for a simulated population of n = 100,000 Portuguese females. Bottom panels represent a simulated dataset for a smaller population of n = 10,000 with the same age structure and mortality rates. In Figure 3 the middle panel in each row contains the estimated fit at the minimum-BIC value of the penalty multiplier λ, while the left and right columns correspond to lower and higher than optimal values of λ, respectively.

The preferred P-spline fit in the top row, center column of Figure 3 represents a significant improvement over the unpenalized maximum likelihood fit to the same data in Figure 2: fitted rates vary much more smoothly, are more nearly linear over older adult ages, and are monotonically increasing over older ages. These are common features in mortality schedules estimated from large populations, and the P-spline model finds good tradeoffs between fitting the small-sample data and smoothing the schedule.

However, there is also evidence in Figure 3 that P-spline fits for mortality schedules in small populations could be improved. In particular, the typical pattern in age-specific mortality schedules over ages 0–25, with a very steep decline in mortality risk over ages 0–10, followed by a more gradual rise over 10–25, is sometimes not well approximated by local cubic functions with a global penalty λ that affects smoothness at all ages.

For populations in the size range covered in Figure 3 (i.e., tens or hundreds of thousands of residents of one sex), minimum-BIC smoothing by P-splines tends to produce estimated functions are too symmetric over child and young adult ages – in the sense that the estimated decline in mortality risks over the youngest ages is too similar to the increase over late childhood and early adulthood. The leftmost panels in Figure 3 suggest, for example, that fitting the rapid decline in mortality risks at the youngest ages often requires a lower-than-optimal global penalty λ. With this lower penalty, the rest of the estimated schedule has implausible rises and falls, has a trough in adolescent and young adult ages that is too deep and wide, or both. In other words, reducing bias to improve fit at the youngest ages may cause undesirable variance and bias at other ages. Going in the other direction, increasing the global λ penalty (rightmost panels) in order to lower variance can cause undesirable bias at almost all ages. These problems of balancing bias and variance are more severe in smaller samples (bottom panels).

Figure 3: P-spline fits for two simulated small-area samples, using three alternative penalties



Note: Population size is 100,000 in the top panels (which contain data identical to Figure 2), and 10,000 in bottom panels (same structure but 1/10th as many females in the simulated small area). Middle column illustrates the penalized fits that minimize the Bayesian Information Criterion. Left and right columns show fits with alternative λ values that lead to approximately twice and half as many effective degrees of freedom, respectively, as the BIC-minimizing value in the center column.

2.3.2 An alternative: D-spline penalties on the fitted schedule

The fundamental difficulty illustrated in Figure 3 is that demographic schedules do not always have mathematically convenient shapes. Even a very flexible, well-tuned parametric model can have trouble matching the nuances of real age patterns.

The P-spline approach to function smoothing has proven to be very valuable, but it is not specifically designed for fitting demographic rate schedules. In particular, it relies on polynomial functions as a kind of gold standard for functional shapes. That reliance makes good sense for a generic curve-fitting tool, but it may not be optimal for fitting specific types of curves for which demographers already have specialized models and considerable prior knowledge.

Here I investigate an alternative approach that I call *D-splines*. D-splines also use penalized, high-dimensional spline functions, but penalties are based on deviations from *demographic*, rather than *arithmetic*, standards. The essential idea is to penalize features

of the fitted schedule $s = \boldsymbol{B}\theta$ directly, using demographic prior knowledge. D-splines are variants of standard P-splines. They are similar to the shape penalties discussed by Eilers (2017: Section 3.3), with the important difference (described more thoroughly below) that penalties are precalibrated from outside data, rather than chosen by hand or selected with a data-dependent index such as BIC.

D-splines use a large collection of demographic rates to define squared-error penalties for schedules that deviate from typical age patterns. Schedules have low penalties if they are similar to those in a high-quality empirical database, and variation within the database helps to calibrate the definition of "similar." In this sense the D-spline penalties developed here for mortality are very similar to the cohort shape penalties developed from the Human Fertility Database at the Max Planck Institute for Demographic Research and Vienna Institute of Demography by Schmertmann et al. (2014) for forecasting fertility schedules. The penalties investigated here are especially similar to the spline-based procedure for fertility rate interpolation proposed in Schmertmann (2014).

I propose several alternative D-spline penalties in Section 3, all constructed as follows

- define a *residual* vector $\varepsilon \in \mathbb{R}^G$ whose elements should all be close to zero for "good" schedules $s = \boldsymbol{B}\theta$
- calculate empirical residuals $\varepsilon_1 \ldots \varepsilon_{222}$ across the set of all 222 1x10 (single-year age by ten-year period) observed life tables in the Human Mortality Database for periods beginning in 1970 or later at the University of California and Max Planck Institute for Demographic Research (2014).
- use the $G \times G$ empirical covariance matrix $\hat{\boldsymbol{V}}_{HMD} = \frac{1}{222}\sum_i (\varepsilon_i \varepsilon_i')$ as an estimate for the expected covariance of residuals
- replace the P-spline penalty on parameters θ with a "D-spline" penalty based on the residuals for fitted functions $\boldsymbol{B}\theta$

The result is a penalized likelihood function

$$f(\theta) = L(\theta) - \frac{1}{2}\,\varepsilon'(\theta)\,\left[\hat{\boldsymbol{V}}_{HMD}^{-1}\right]\,\varepsilon(\theta) \tag{4}$$

for which the maximization tradeoff is "fit versus fidelity to patterns in demographic schedules," rather than "fit versus local smoothness."[4] The absence of the unknown smoothing constant λ is an important benefit: unlike P-splines, in the D-spline ap-

[4]The $\frac{1}{2}$ constant in the penalized log likelihood arises from an assumption that residuals are approximately normally distributed over demographic schedules. Informal examination of empirical HMD residuals for the penalties used in this paper suggests that normality is quite reasonable: residual distributions are close to symmetric, with very thin tails.

proach there is no λ to be estimated by grid search in an outer loop. Instead the penalty term/tolerance for deviations is calibrated empirically to mimic residual patterns found in high-quality demographic data. The researcher only has to estimate the model once. Another potentially important difference stems from penalizing the fitted function value $s = \boldsymbol{B}\theta$ (as in D-splines) rather than the parameters θ (as in P-splines). Unlike P-splines, D-splines do not require a specific class of basis functions.[5]

In all the specific examples that follow, residual vectors are linear functions of the spline coefficients θ, with the form

$$\varepsilon(\theta) \;=\; \boldsymbol{AB}\theta - c \;\in \mathbb{R}^G, \tag{5}$$

where $\boldsymbol{A} \in \mathbb{R}^{G\times G}$ and $c \in \mathbb{R}^G$ are predefined constants, as described in next section. With residuals as in Equation (5), the D-spline penalized likelihood is

$$f(\theta) = L(\theta) - \frac{1}{2}\,(\boldsymbol{AB}\theta - c)'\,\left[\hat{\boldsymbol{V}}_{HMD}^{-1}\right]\,(\boldsymbol{AB}\theta - c)\,, \tag{6}$$

and the θ vector the maximizes the function can be found by an iterative Newton-Raphson algorithm. Details for that algorithm, and for calculation of effective degrees of freedom, are in the Appendix.

3. Experimental D-spline penalties for mortality schedules

3.1 Slope penalties: D–1

Log mortality schedules for human populations have characteristic shapes that can be described in terms of slopes (or equivalently, first-differences over single years of age). One possible approach to characterizing "good" spline schedules is therefore to measure the difference between the slopes in a proposed spline function $s = \boldsymbol{B}\theta$ and the average slopes in HMD schedules at the same ages.

For example, the average value of $(\ln \mu_1 - \ln \mu_0)$ across 222 1x10 HMD schedules is –2.432, with a standard deviation of 0.292. This suggests that steep negative slopes are likely between age 0 and age 1 in "good" schedules, but that we should be fairly tolerant about the exact slope value because of the large standard deviation. In contrast, between integer ages 75 and 76 the average value and standard deviation of $(\ln \mu_{76} - \ln \mu_{75})$ across HMD schedules are +0.117 and 0.024, respectively, which suggests that between

[5] More precisely, D-splines will produce the same fits for any set of basis functions that have the same column space.

these ages almost all "good" schedules have slopes within a narrow range of small positive values.

Applying this approach simultaneously to all $G = 99$ first-differences for intervals starting with ages 0...98, the HMD slopes have mean vector m_1 and covariance matrix $\boldsymbol{V}_1$. Defining $\boldsymbol{\Delta}_1$ as the standard 99×100 first-differencing matrix, this produces a penalized log likelihood for the spline schedule $\boldsymbol{B}\theta$:

$$f_1(\theta) = L(\theta) - \frac{1}{2} \left(\boldsymbol{\Delta}_1 \boldsymbol{B}\theta - m_1\right)' \boldsymbol{V}_1^{-1} \left(\boldsymbol{\Delta}_1 \boldsymbol{B}\theta - m_1\right). \tag{7}$$

The value θ_1^* that maximizes this function produces a fitted schedule $\boldsymbol{B}\theta_1^*$ for mortality, denoted from here on as the *D–1 estimator*.

3.2 Curvature penalties: D–2

A slightly less demanding criterion than that in Section 3.1 might penalize curvature (second differences) in the fitted spline schedule that failed to match HMD empirical patterns. Defining $\boldsymbol{\Delta}_2$ as the standard 98×100 second-differencing matrix and defining m_2 and $\boldsymbol{V}_2$ as above (but now using the $G = 98$ second differences in each HMD schedule) yields an analogous penalized log likelihood with different constants:

$$f_2(\theta) = L(\theta) - \frac{1}{2} \left(\boldsymbol{\Delta}_2 \boldsymbol{B}\theta - m_2\right)' \boldsymbol{V}_2^{-1} \left(\boldsymbol{\Delta}_2 \boldsymbol{B}\theta - m_2\right), \tag{8}$$

a different optimal value θ_2^* , and a different D-spline fit $\boldsymbol{B}\theta_2^*$. Call this estimator *D–2*.

3.3 Lee-Carter penalties: D–LC

One can also define "good" spline schedules according to their fidelity to existing models. In this approach residuals might represent features of a schedule that cannot be represented within a specified model family. For example, in the most commonly used mortality modelling framework, Lee and Carter (1992), schedules are modeled as

$$\ln \mu_x = a_x + k \cdot b_x,$$

where a and b are $A \times 1$ vectors of predetermined constants that represent a baseline schedule and typical deviations from that schedule, as estimated from a singular value decomposition on reference data.

In the Lee-Carter model the scalar parameter k determines the level of deviation from the baseline, so that (with $A = 100$ ages) the vector

$$\begin{bmatrix} \ln \mu_0 - a_0 \\ \vdots \\ \ln \mu_{99} - a_{99} \end{bmatrix} = k \cdot \begin{bmatrix} b_0 \\ \vdots \\ b_{99} \end{bmatrix}$$

must lie in the column space of vector $b \in \mathbb{R}^{100}$. This suggest another way to define D-spline penalties using the HMD. Specifically, we can estimate the Lee-Carter a and b vectors from the HMD and then define the $G = 100$ residuals for any schedule $\{\ln \mu\} \in \mathbb{R}^{100}$ as the part of $\{\ln \mu\} - a$ that lies outside of the column space of b. In matrix notation this vector of residuals is

$$\varepsilon \;=\; \left[\boldsymbol{I} - b(b'b)^{-1}b'\right]\,\left[\{\ln \mu\} - a\right] \;=\; \boldsymbol{M}_b\left[\{\ln \mu\} - a\right].$$

After calculating the Lee-Carter residuals' covariance ($\boldsymbol{V}_{LC}$) across HMD schedules, the corresponding penalized log likelihood for D-spline estimation is

$$f_{LC}(\theta) = L(\theta) - \frac{1}{2}\left(\boldsymbol{M}_b\left(\boldsymbol{B}\theta - a\right)\right)' \boldsymbol{V}_{LC}^{-1}\left(\boldsymbol{M}_b\left(\boldsymbol{B}\theta - a\right)\right). \tag{9}$$

Call θ^*_{LC} that maximizes Equation (9) the *D–LC estimator*.

It is important to notice that maximizing Equation (9) does not exactly reproduce a Lee-Carter fit. Instead, it requires that any deviations from Lee-Carter schedule shapes in the fitted spline $s = \boldsymbol{B}\theta$ must be compensated by better fits to observed mortality. In other words, Equation (9) rewards, but does not insist on, Lee-Carter-like schedules when attempting to fit data.

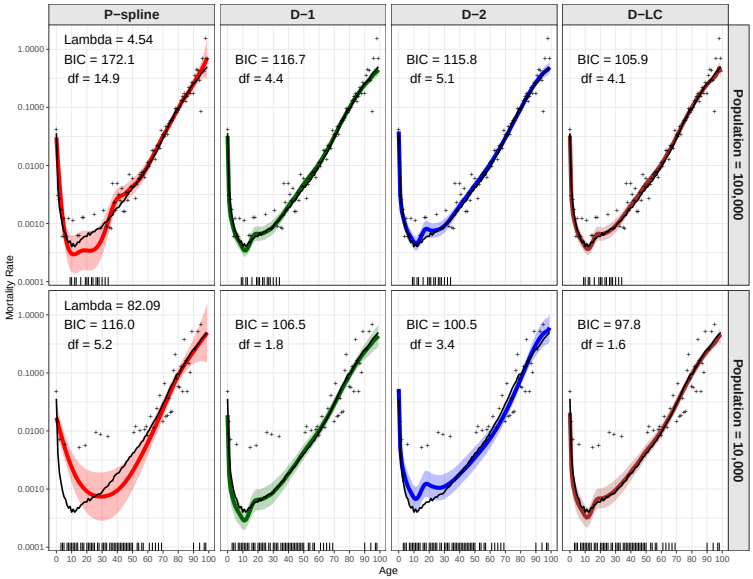
3.4 A preliminary comparison

Figure 4 shows the spline functions that maximize the four alternative criteria when using the simulated small-area data used earlier in Figures 2 and 3. It also illustrates the estimated uncertainty of those estimates as bands representing 95% confidence intervals for the spline function value at each age. The Appendix contains details about estimation and approximation of uncertainty.

For these particular data sets, all three D-spline methods appear to fit the true schedule of rates better than the minimum-BIC fit using standard P-splines. They also appear to have lower uncertainty. The improvement in fit over P-splines is especially pronounced

when the population at risk is smaller, as in the bottom panels. It also appears that the D–1 and D–LC estimators fit the overall shape of the mortality schedule better than the D-spline fit with curvature penalties (D–2). Figure 4 contains only two simulated mortality datasets, however. In order to examine whether the patterns in this figure generalize, in the next section I perform a large experiment with HMD data over many thousands of samples.

Figure 4: Alternative fits to simulated small-population mortality data from Figure 3. P-spline fits are minimum-BIC curves



Note: D–1 and D–2 fits penalize first-difference (slope) and second-difference (curvature) residuals for the schedule of log mortality rates, respectively. D–LC fit penalizes deviations from a Lee-Carter model. Tick marks along the horizontal axes represent single-year ages with no deaths. Thin dark lines represent true HMD rates from which data is simulated, and are identical in each panel. Thicker lines represent estimated schedules for each method. Shaded bands represent 95% confidence intervals for the spline estimates, calculated using the covariance formulas derived in Appendix A.5.

4. Comparative study

4.1 D-spline constants and test samples

I use 222 single-year age by ten-year period ("1×10") female mortality schedules in the HMD as the foundation for small-population experiments. These schedules come from 49 different countries, over decades spanning the 1970s to the 2010s. For each of the proposed D-spline criteria and each of the 49 countries in the selected HMD data, I calculated the $G \times 1$ mean residual vector over all schedules that did *not* belong to that country, and the $G \times G$ covariance matrix $\boldsymbol{V}$ of the residuals across the same schedules. These calculations provide country-specific constants for the D-spline objective functions, which in all cases are estimated data from *other* countries. For example, no data from Portugal influenced the D-spline constants used in Figure 4. Similarly, Australian data is excluded when calculating constants for Australia; Austrian data is excluded for Austria; and so on for all 49 countries in the selected HMD dataset.

I then used the HMD 1×10 exposure data associated with each mortality schedule to create simulated (deaths,exposure) samples, as follows. For each schedule $i = 1 \ldots 222$ I rescaled observed age-specific exposures (N_{ix}) to represent a small population with the same age structure: $N^*_{ix} = P^* \cdot \frac{N_{ix}}{\sum_x N_{ix}}$, where the small population P^* is either 10,000 or 100,000. For each of the 222×2=444 simulated small populations, I drew 100 independent samples of simulated deaths at ages $x = 0 \ldots 99$ using the true log mortality rates from the corresponding HMD schedule:

$$D^{sim}_{ix} \sim Poisson(N^*_{ix} \cdot \mu_{ix}) \qquad x \in 0 \ldots 99 \quad \text{[repeated 100 times]}.$$

4.2 Experimental design

For each of the 222×2×100 = 44,400 (HMD schedule, population size, simulation) combinations, I estimated the optimal schedule $\boldsymbol{B}\theta$ separately for each the three D-spline penalty functions[6]. D-spline calculations used a standard Newton-Raphson algorithm as in Equation (A-1). For each (HMD schedule, population size, simulation) combination I also estimated a P-spline function $\boldsymbol{B}\theta$ with the same basis $\boldsymbol{B}$, using the *Mort1Dsmooth* function from the *MortalitySmooth* package in *R* Camarda (2012), with all settings other than the interior knots set at default values.

The experiment therefore comprises 44,400×4 methods = 176,600 estimated

[6] As noted earlier, in order to enhance comparability with the P-spline estimator, reported D-spline results for the Monte Carlo experiments used the $\boldsymbol{B}$ matrix as implemented in the *MortalitySmooth* package (Camarda 2012). The specific command to generate the 100×36 matrix is *MortSmooth_bbase(x = 0:99, xl = –0.99, xr = 99.99, ndx = 33, deg = 3)*. Results are not sensitive to this choice: errors for D-spline estimators are virtually identical if the basis matrix is replaced with the D-spline default defined earlier and illustrated in Figure 1A.

mortality schedules, each with a known true schedule from the HMD.[7] For each schedule estimate I examined three error measures:

1. errors in age-specific log mortality rates (estimated $\ln \hat{\mu}_x$ − true $\ln \mu_x$)
2. errors in life expectancy at birth: $\hat{e}_0$ − true e_0
3. errors in working-age mortality per 1000: 1000 $({}_{45}\hat{q}_{20}$ − true${}_{45}q_{20})$

The next section reports tabular and graphical summaries of these error measures over fitting methods and sample sizes.

4.3 Evaluation of fitting errors

4.3.1 Overall shape: errors in estimating age-specific mortality rates

Table 1 reports summary data comparing minimum-BIC P-spline fits with the three alternative D-spline estimators. As in the figures presented earlier, the top rows contain information for simulated small populations of 100,000 women, bottom rows for smaller populations of 10,000.

Summaries over many simulated populations and true mortality schedules in Table 1 confirm the intuition from the single example in Figure 4. Namely, in small populations D-spline estimators tend to have much smaller fitting errors than standard P-splines. The mean absolute error of estimated mortality rates is more than two times larger for P-splines than for any of the D-spline variants in the tests with samples of 100,000 woman-years; it is also notably higher in samples of 10,000 woman-years. In the smaller samples represented in the bottom block in Table 1, there is a notable negative bias in P-spline estimated rates, and mean absolute errors are several times larger for P-splines than for the D–1 and D–LC estimators.

For any one of the 222×100 = 22,200 samples of a given sample size there are four alternative estimated schedules, one for each method. The rows labeled *Lowest BIC* and *Lowest Mean Abs Error* in Table 2 show the percentage of those 22,200 samples in which each estimator was the best-performing. For example, in samples of size 100,000 the standard P-spline estimator had the lowest BIC of the four alternatives in 0.9% of samples and the lowest mean absolute error in 0.2%. In contrast, for those same samples the D–LC estimator had the lowest BIC in 54% and the lowest mean absolute error in 38%. Thus in direct within-sample comparisons P-spline estimation performs much more poorly than D-spline alternatives – especially the LC variant.

[7] The D–2 estimator did not converge in 33 of 22,200 samples of size 10,000, and in 31 of the 22,200 samples of size N = 100,000. Calculations and plots in this section omit those 64 cases.

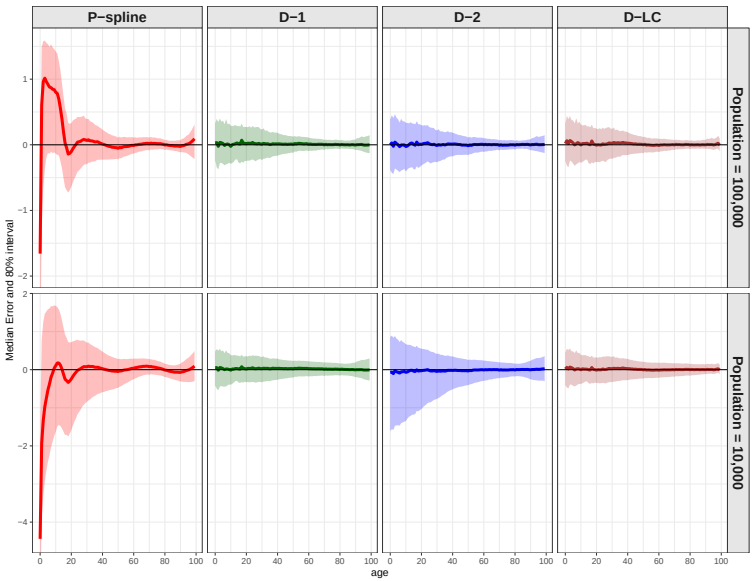
Table 1: Summaries of fitting errors by sample size and estimation method. For each of 222 HMD mortality schedules, each method was applied 100 times to independent samples representing 100,000 or 10,000 woman-years

Population	Error Measure	P-spline	D–1	D–2	D–LC
100,000	Mean Abs Err	0.25	0.11	0.12	0.10
	Mean Error	0.07	0.01	0.00	0.01
	Median df	7.2	3.5	4.4	3.3
	Median BIC	143.7	113.6	116.5	112.4
10,000	Mean Abs Err	0.48	0.17	0.31	0.16
	Mean Error	–0.13	0.02	–0.08	0.01
	Median df	2.5	1.6	3.2	1.4
	Median BIC	88.4	79.9	85.8	79.5
Percent of 222 × 100 = 22,200 samples of a given size in which estimation method has...					
100,000	Lowest BIC	0.9%	42%	3%	54%
	Lowest Mean Abs Err	0.2%	30%	32%	38%
10,000	Lowest BIC	8%	37%	0.4%	55%
	Lowest Mean Abs Err	0.4%	29%	15%	56%

Note: Error at age x for sample s is $\left(\ln \hat{\mu}_{xs} - \ln \mu_x^{true}\right)$. Results in the tables are for all ages 0 . . . 99 combined. BIC = Bayesian Information Criterion, df = effective degrees of freedom, calculated as described in Appendix.

As suggested by the preliminary examples in Figure 4, P-spline fitting errors are particularly large at the youngest ages. In many test samples, especially with smaller populations, the minimum-BIC P-spline curve has less curvature than the true schedule (as in the bottom left panel of Figure 4), and consequently produces estimated rates that tend to be too low for infants and very young children, and then too high for older children and young adults. Figure 5 confirms this pattern, by showing the mean difference (estimated log rate-true log rate) by age for each method and sample size. In small populations standard P-spline estimators clearly have difficulty fitting mortality schedules accurately at ages below 20. As discussed earlier, that difficulty arises from the asymmetric, not-particularly-cubic shape of human mortality schedules at those young ages. D-spline variants, which penalize deviations from the shape of known mortality schedules rather than from global polynomials, perform better.

Figure 5: **Median errors and 10–90% intervals by age, population size, and fitting method. Each plotted point is the average error $\ln \hat{\mu}_x - \ln \mu_x^{true}$ over 22,200 estimates (100 independent samples from each of 222 HMD mortality schedules)**



Note: Shaded bands show the intervals from the 10th to 90th percentile of errors. Top panels correspond to populations with 100,000 women, bottom panels to 10,000 women. Vertical scale differs in top and bottom panels.

The evidence from the simulation study suggests a clear ranking of estimators in terms of fitting age-specific rates. D–LC performs best and standard P-splines perform worst. Among the other two variants, D–1 estimators outperform D–2, mainly because of lower bias at very young ages. D–1 and D–LC estimators do not differ much on the performance metrics in Table 1, but for all summary error measures the Lee-Carter variant is slightly better.

4.3.2 Overall level: e_0

Perhaps the most important metric for success of a small-population method is accuracy in estimating life expectancy at birth. This is the most commonly reported mortality summary, so a good method should have low errors for e_0. Because e_0 is a summary measure of the overall level of mortality, it is possible that a method that does not fit individual age-specific rates well will still produce good estimates of life expectancy.

This would happen if overestimates of mortality rates in some age groups tended to be offset by underestimates in other age groups.

Table 2 reports errors for e_0, and provides evidence that standard P-splines perform much better for life expectancy than for age-specific mortality risks. Table 2 shows that all four estimators have average errors of about one-third of a year in populations of size 100,000 and about one year in populations of size 10,000. Bias is low for all estimators except for P-spline estimators in populations of 10,000 (for which bias = +0.22 years). The main source of error at both population sizes is sampling variance. Life expectancy estimates for the P-spline and D–2 variants are only slightly more variable than those for the other two methods.

There is not a clear "winner" among the four methods on this metric, but overall D–1 slightly outperforms other estimators in estimating e_0, as indicated by the higher probability that it has the lowest mean error. Although the differences are much smaller than those for age-specific rates in Table 1, it is interesting to note that for e_0 the bias-variance tradeoffs seem to be similar: less flexible methods (i.e., those with lower effective degrees of freedom) do a bit better at estimating e_0 from small samples.

Table 2: Errors in estimated life expectancy at birth (e_0)

Population	Error Measure	P-spline	D–1	D–2	D–LC
100,000	Mean Abs Err	0.36	0.34	0.35	0.33
	Mean Error	0.04	–0.01	0.00	0.00
10,000	Mean Abs Err	1.10	0.96	1.11	0.94
	Mean Error	0.22	0.02	0.05	0.08
Percent of 222 × 100 = 22,200 samples of a given size in which estimation method has...					
100,000	Lowest Mean Abs Err	26%	31%	22%	22%
10,000	Lowest Mean Abs Err	22%	33%	19%	26%

4.3.3 Working-age mortality: $1000 \times {}_{45}q_{20}$

As a final comparison I examine estimated working-age mortality over the small populations in the test samples, measured as deaths per 1000 over the age range from 20–65, i.e., $1000\,{}_{45}q_{20}$. Table 3 shows estimation errors for this index across samples. Like life expectancy, this mortality measure covers a range of ages. As a consequence, it is possible that bias and variance in estimated rates at single-year ages may tend to "wash out" in the summary measure. That is the case with estimated working-age mortality. As with e_0, errors are much more similar across methods than they are for age-specific rates.

For working-age mortality, D–LC estimators again have the lowest mean absolute

error. P-spline estimators are slightly more likely than the other methods produce the lowest errors in samples of size 100,000 and D–LC is the most likely to produce the smallest error in samples of 10,000. But, as they were for life expectancy, these differences in performance are very small. The main conclusion from comparing estimation errors for this mortality index is that all four methods perform reasonably well: biases are small for all methods, and there are not large differences in the chances of a method being the "winner" in one of the 22,200 samples of a given size. However, as with other metrics, there is an advantage (here very slight) for the "stiffer" methods with fewer effective degrees of freedom.

Table 3: Errors in estimated deaths per 1000 over working ages ($1000 \times_{45} q_{20}$)

Population	Error Measure	P-spline	D–1	D–2	D–LC
100,000	Mean Abs Err	6.8	6.4	6.4	6.3
	Mean Error	–0.7	+0.4	+0.4	–0.6
	Mean Abs Pct Error	5.8	5.3	5.6	5.2
10,000	Mean Abs Err	20.6	14.0	18.3	13.9
	Mean Error	4.2	1.5	+0.0	–2.0
	Mean Abs Pct Error	17.7	11.7	15.8	11.2
Percent of 222 × 100 = 22,200 samples of a given size in which estimation method has...					
100,000	Lowest Mean Abs Err	28%	27%	24%	22%
10,000	Lowest Mean Abs Err	19%	29%	21%	31%

5. Discussion

The simulation study in Section 4 demonstrates that demographically penalized splines (D-splines) outperform standard P-splines for fitting mortality schedules in small populations. The advantages of D-splines are especially notable at young ages and in smaller populations. Although D-spline estimation requires calculating a large number of constants in advance, easy access to large demographic datasets makes this a simple task.[8] After constants have been calculated once, D-spline estimators are "precalibrated" and do not require a separate process to select an optimal global penalty constant λ.

[8] In the examples in this paper D-spline penalties use 100 means and 5050 unique variances and covariances, which can be calculated on a standard laptop computer in less than a few seconds. Inverting a 100×100 covariance matrix in which many elements are very small could in principle lead to serious numerical instabilities. In practice such instabilities do not appear to be a problem. In the cross-validated simulation exercise 147 such matrices (one matrix per method for each of 49 countries) were inverted using *R*'s generalized inverse function (*ginv* in the *MASS* library) without any notable numerical problems in the fitted D-spline schedules.

Demographic analyses have many objectives. Although D-splines produce notably better fits for age-specific mortality rates, their advantages are less important for estimating indices like e_0 or ${}_{45}q_{20}$ that span wide age groups. For these indices the standard P-spline approach and D-spline estimators all tend to perform well, because all do a good job at estimating the *average* mortality rate over an age interval, even when do not get the individual age-specific rates right.

The D-spline approach clearly depends on observed regularities – and observed variability – in demographic schedules. Simulation tests in this paper are cross-validated, in the sense that data from a country is not used to build the D-spline constants used when estimating that country's mortality schedule. However, cross-validation occurs within "HMD-type countries" only, which leave unanswered questions about generalizability to mortality schedules from other times, other places, and other epidemiological environments.

We can safely conclude from the comparative study presented here that demographically-penalized spline estimators outperform standard penalized spline approaches in small samples from populations with mortality risks similar to those in contemporary rich countries with high-quality data. For many purposes – such as small-area estimates for contemporary US counties, German *Kreise*, or Korean townships – results show that flexible estimators with demographic penalties derived from the HMD are likely to produce much better estimates than standard P-splines. For other purposes – such as estimates for historical or disadvantaged populations – the jury is still out, but the D-spline framework provides a structure in which demographers can evaluate the costs and benefits of alternative penalties derived from other sources.

References

Alexander, M. and Alkema, L. (2018). Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research* 38(15): 335–372. doi:10.4054/DemRes.2018.38.15.

Alkema, L. and New, J.R. (2014). Global estimation of child mortality using a Bayesian B-spline bias reduction model. *The Annals of Applied Statistics* 8(4): 2122–2149. doi:10.1214/14-AOAS768.

Amemiya, T. (1985). *Advanced econometrics*. Cambridge: Harvard University Press.

Camarda, C. (2012). MortalitySmooth: An R package for smoothing poisson counts with P-splines. *Journal of Statistical Software* 50(1): 1–24. doi:10.18637/jss.v050.i01.

Camarda, C.G., Eilers, P.H., and Gampe, J. (2016). Sums of smooth exponentials to decompose complex series of counts. *Statistical Modelling* 16(4): 279–296. doi:10.1177/1471082X16641796.

Currie, I.D., Durban, M., and Eilers, P.H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling* 4(4): 279–298. doi:10.1191/1471082X04st080oa.

Curry, H. and Schoenberg, I. (1947). On Pólya frequency functions IV: The spline functions and their limits. *Bulletin of the American Mathematical Society* 53(11): 1114.

de Beer, J. (2012). Smoothing and projecting age-specific probabilities of death by TOPALS. *Demographic Research* 27(20): 543–592. doi:10.4054/DemRes.2012.27.20.

de Boor, C. (2001). *A practical guide to splines*. Applied Mathematical Sciences. New York: Springer.

de Boor, C. (1976). *Splines as linear combinations of B-splines. A survey*. New York: Academic Press.

de Jong, P. and Tickle, L. (2006). Extending Lee–Carter mortality forecasting. *Mathematical Population Studies* 13(1): 1–18. doi:10.1080/08898480500452109.

Eilers, P.H.C. (2017). Uncommon penalties for common problems. *Journal of Chemometrics* 31(4): e2878. doi:10.1002/cem.2878.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* 11(2): 89–102. doi:10.1214/ss/1038425655.

Eilers, P.H.C. and Marx, B.D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics* 2(6): 637–653. doi:10.1002/wics.125.

Gonzaga, M.R. and Schmertmann, C.P. (2016). Estimating age-and sex-specific mortality rates for small areas with TOPALS regression: An application to Brazil in 2010.

Revista Brasileira de Estudos de População 33(3): 629–652. doi:10.20947/S0102-30982016c0009.

Greene, W.H. (1997). *Econometric analysis*. Upper Saddle River: Prentice Hall.

Greville, T. (1964). Numerical procedures for interpolation by spline functions. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis* 1(1): 53–68.

Hilton, J., Dodd, E., Forster, J.J., and Smith, P.W.F. (2019). Projecting UK mortality by using Bayesian generalized additive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 68(1): 29–49. doi:10.1111/rssc.12299.

Hoem, J.M., Madien, D., Nielsen, J.L., Ohlsen, E.M., Hansen, H.O., and Rennermalm, B. (1981). Experiments in modelling recent Danish fertility curves. *Demography* 18(2): 231–244. doi:10.2307/2061095.

Human Fertility Database (n.d.). The human fertility database [electronic resource]. Rostock: Max Planck Institute for Demographic Research and Vienna: Vienna Institute of Demography. http://www.humanfertility.org/cgi-bin/main.php.

Human Mortality Database (2014). The human mortality database [electronic resource]. Berkeley: University of California and Rostock: Max Planck Institute for Demographic Research. http://www.mortality.org.

Hyndman, R.J. and Ullah, M.S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis* 51(10): 4942–4956.

Jasilioniene, A., Jdanov, D.A., Sobotka, T., Andreev, E.M., Zeman, K., Shkolnikov, V.M., Goldstein, J.R., Philipov, D., and Rodriguez, G. (2012). Methods protocol for the human fertility database. Rostock: Max Planck Institute for Demographic Research.

Lee, R.D. and Carter, L.R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association* 87(419): 659–671. doi:10.1080/01621459.1992.10475265.

McNeill, D., Trussell, T., and Turner, J. (1977). Spline interpolation of demographic data. *Demography* 14(2): 245–252.

Neumaier, A. (1998). Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review* 40(3): 636–666. doi:10.1137/S0036144597321909.

Rau, R. and Schmertmann, C.P. (2020). District-level life expectancy in Germany. *Deutsches Ärzteblatt International* 117(29–30): 493–499. doi:10.3238/arztebl.2020.0493.

Rizzi, S., Gampe, J., and Eilers, P.H.C. (2015). Efficient estimation of smooth distributions from coarsely grouped data. *American Journal of Epidemiology* 182(2): 138–147. doi:10.1093/aje/kwv020.

Schmertmann, C.P. (2003). A system of model fertility schedules with graphically intuitive parameters. *Demographic Research* 9(5): 81–110. doi:10.4054/DemRes.2003.9.5.

Schmertmann, C.P. (2014). Calibrated spline estimation of detailed fertility schedules from abridged data. *Revista Brasileira de Estudos de População* 31: 291–307. doi:10.1590/S0102-30982014000200004.

Schmertmann, C.P. and Gonzaga, M.R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography* 55(4): 1363–1388. doi:10.1007/s13524-018-0695-2.

Schmertmann, C.P., Zagheni, E., Goldstein, J.R., and Myrskylä, M. (2014). Bayesian forecasting of cohort fertility. *Journal of the American Statistical Association* 109(506): 500–513. doi:10.1080/01621459.2014.881738.

Schoenberg, I.J. (1973). *Cardinal spline interpolation*. Philadelphia: SIAM.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6(2): 461–464.

Wilmoth, J.R., Andreev, K., Jdanov, D., Glei, D.A., Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., and Vachon, P. (2007). Methods protocol for the human mortality database [electronic resource]. Berkeley: University of California and Rostock: Max Planck Institute for Demographic Research. http://mortality.org.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* 93(441): 120–131. doi:10.1080/01621459.1998.10474094.

Appendix A: Notes on estimating D-splines

A-1 Newton-Raphson

In a penalized likelihood maximization problem, the objective function has the form

$$f(\theta) \;=\; L(\theta) - P(\theta),$$

where $\theta \in \mathbb{R}^K$ is a vector of parameters and L and P are the likelihood and penalty functions, respectively. In order to maximize we need to set the derivative of $L(\theta) - P(\theta)$ equal to a zero vector in $\mathbb{R}^K$. In other words

$$L_\theta(\theta) = P_\theta(\theta),$$

This is a nonlinear system of K equations in K parameters θ, which can be approximated by the linear system in the neighborhood of an arbitrary θ_t as

$$L_\theta(\theta_t) \,+\, \boldsymbol{L_{\theta\theta'}(\theta_t)}\,(\theta - \theta_t) \;=\; P_\theta(\theta_t) \,+\, \boldsymbol{P_{\theta\theta'}(\theta_t)}\,(\theta - \theta_t),$$

where $\boldsymbol{L_{\theta\theta'}}$ and $\boldsymbol{P_{\theta\theta'}}$ are $K \times K$ Hessian matrices of second derivatives.

The optimal θ for this local linearized system solves

$$\begin{aligned}
[\boldsymbol{L_{\theta\theta'}}(\theta_t) - \boldsymbol{P_{\theta\theta'}}(\theta_t)]\,(\theta - \theta_t) \;&=\; -[\,L_\theta(\theta_t) - P_\theta(\theta_t)\,] \\
[\boldsymbol{L_{\theta\theta'}}(\theta_t) - \boldsymbol{P_{\theta\theta'}}(\theta_t)]\,\theta \;&=\; [\boldsymbol{L_{\theta\theta'}}(\theta_t) - \boldsymbol{P_{\theta\theta'}}(\theta_t)]\,\theta_t - [\,L_\theta(\theta_t) - P_\theta(\theta_t)\,] \\
\boldsymbol{H}_t\,\theta \;&=\; \boldsymbol{H}_t\,\theta_t - g_t,
\end{aligned} \tag{A-1}$$

where $\boldsymbol{H}_t \in \mathbb{R}^{K\times K}$ and $g_t \in \mathbb{R}^K$ are the Hessian matrix and the gradient vector of $f(\theta)$, evaluated at θ_t. Newton-Raphson iteratively solves for θ, updates $\boldsymbol{H}$ and g using that new θ, re-solves for θ, and so on to convergence (Amemiya 1985: 137–139).

A-2 Poisson likelihood

The Poisson log likelihood is

$$L(\theta) \;=\; \text{constant} - \sum_x N_x \exp(s_x) + \sum_x D_x s_x,$$

where $\boldsymbol{B} \in \mathbb{R}^{A\times K}$ is the B-spline basis, $\theta \in \mathbb{R}^K$ are the spline parameters, $b_x \in \mathbb{R}^K$ is the row of $\boldsymbol{B}$ that corresponds to age x, and $s_x = b_x'\,\theta$ is the spline approximation to the log mortality rate at age x. In terms of expected deaths, the likelihood is

$$L(\theta) \;=\; \text{constant} - \sum_x \hat{D}_x + \sum_x D_x s_x.$$

The gradient of L is

$$\begin{aligned} L_\theta(\theta) \;&= -\sum_x \frac{\partial \hat{D}_x}{\partial \theta} \;+\; \sum_x D_x \frac{\partial s_x}{\partial \theta} \\ &= -\sum_x \hat{D}_x b_x + \sum_x D_x b_x \\ &= \boldsymbol{B}'\,(D - \hat{D}), \end{aligned}$$

where $D, \hat{D} \in \mathbb{R}^A$. The Hessian of L is

$$\boldsymbol{L_{\theta\theta'}}(\theta) \;=\; -\boldsymbol{B'}\,[\boldsymbol{diag}(\boldsymbol{\hat{D}})]\,\boldsymbol{B}.$$

A-3 D-spline penalty function

The penalty term is

$$P(\theta) = \;\tfrac{1}{2}\,\varepsilon'(\theta)\,\boldsymbol{V}^{-1}\,\varepsilon(\theta),$$

where the residuals $\varepsilon \in \mathbb{R}^G$ are

$$\varepsilon(\theta) \;=\; \boldsymbol{AB}\theta - c$$

for some constants $\boldsymbol{A}, \boldsymbol{V^{-1}} \in \mathbb{R}^{G \times G}$ and $c \in \mathbb{R}^G$.

The gradient of P is

$$P_\theta(\theta) \;=\; \boldsymbol{B'A'V}^{-1}\left(\boldsymbol{AB}\theta - c\right)$$

and the Hessian of P is

$$\boldsymbol{P_{\theta\theta'}}(\theta) \;=\; \boldsymbol{B'A'V}^{-1}\boldsymbol{AB}.$$

A-4 Iterative optimization

In order to maximize the penalized log likelihood we need to set the derivative of $L(\theta) - P(\theta)$ equal to a zero vector in $\mathbb{R}^K$. In this case

$$\boldsymbol{B'}\left(D - \hat{D}\right) = \boldsymbol{B'A'V}^{-1}\left(\boldsymbol{AB}\theta - c\right). \tag{A-2}$$

Newton-Rapshon iteratation repeatedly finds θ_{t+1} that solves $\boldsymbol{H}_t\theta_{t+1} = \boldsymbol{H}_t\theta_t - g_t$ via least squares regression. In this case

$$\begin{aligned}\left[-\boldsymbol{B'}\left[\boldsymbol{diag(\hat{D}_t)}\right]\boldsymbol{B} - \boldsymbol{B'A'V}^{-1}\boldsymbol{AB}\right]\theta_{t+1} \;=\; &\left[-\boldsymbol{B'}\left[\boldsymbol{diag(\hat{D}_t)}\right]\boldsymbol{B} - \boldsymbol{B'A'V}^{-1}\boldsymbol{AB}\right]\theta_t \\ &- \left[\boldsymbol{B'}\left(D - \hat{D}\right) - \boldsymbol{B'A'V}^{-1}\left(\boldsymbol{AB}\theta - c\right)\right]\end{aligned}$$

A-5 Approximate uncertainty

The Newton-Raphson algorithm uses a quadratic approximation to $f(\theta)$ at each iteration. This implies that the covariance of θ estimates is approximately the inverse of the negative Hessian, evaluated at the final estimate θ^*:

$$\Sigma_\theta \approx \left[\boldsymbol{B'}\,diag(\hat{D}^*)\boldsymbol{B} \;+\; \boldsymbol{B'A'V}^{-1}\boldsymbol{AB}\right]^{-1} \tag{A-3}$$

where $\hat{D}^* \in \mathbb{R}^A$ is the vector of predicted deaths by age, evaluated at $\theta = \theta^*$. Thus the covariance of the estimated age-specific log mortality rates from the spline function $s = \boldsymbol{B}\theta^*$ is the $A \times A$ matrix

$$\Sigma_s \approx \boldsymbol{B}\,\Sigma_\theta\boldsymbol{B'} \;=\; \boldsymbol{B}\left[\boldsymbol{B'}\,diag(\hat{D}^*)\boldsymbol{B} \;+\; \boldsymbol{B'A'V}^{-1}\boldsymbol{AB}\right]^{-1}\boldsymbol{B'} \tag{A-4}$$

A-6 Effective degrees of freedom

Following Ye (1998), we can estimate the effective degrees of freedom of a model as $\sum_i \frac{\partial \hat{y}_i}{\partial y_i}$, which in the Poisson mortality case would be $\sum_x \frac{\partial \hat{D}_x}{\partial D_x}$. In other words, the more flexible the model, the more sensitive the *predicted* number of deaths at age x is to the *observed* number of deaths at x.

Start with the maximizing condition in Equation (A-2), and think of it as telling us how to select parameters $\theta \in \mathbb{R}^K$ given observed deaths $D \in \mathbb{R}^A$. We have to pick θ to balance the two sides of the equation.

$$\boldsymbol{B}'\left(D - \hat{D}(\theta)\right) = \boldsymbol{B}'\boldsymbol{A}'\boldsymbol{V}^{-1}\left(\boldsymbol{A}\boldsymbol{B}\theta - c\right).$$

Rearranging to emphasize the new interpretation produces

$$\boldsymbol{B}'D \;=\; \boldsymbol{B}'\hat{D}(\theta) \;+\; \boldsymbol{B}'\boldsymbol{A}'\boldsymbol{V}^{-1}\left(\boldsymbol{A}\boldsymbol{B}\theta - c\right).$$

So for small changes in deaths $\Delta D \in \mathbb{R}^A$ and parameters $\Delta\theta \in \mathbb{R}^K$, maximization requires

$$\boldsymbol{B}'\,\Delta D \;\approx\; \left[\boldsymbol{B}'\boldsymbol{diag}(\hat{\boldsymbol{D}})\boldsymbol{B}\right]\,\Delta\theta \;+\; \left[\boldsymbol{B}'\boldsymbol{A}'\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{B}\right]\Delta\theta$$

or

$$\begin{aligned}\Delta\theta \;&\approx\; \left[\boldsymbol{B}'\boldsymbol{diag}(\hat{\boldsymbol{D}})\boldsymbol{B} \;+\; \boldsymbol{B}\boldsymbol{A}'\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{B}\right]^{-1}\;\boldsymbol{B}'\,\Delta D \\ &\approx\; \hat{\boldsymbol{Q}}^{-1}\boldsymbol{B}'\,\Delta D,\end{aligned}$$

where $\hat{\boldsymbol{Q}} = \left[\boldsymbol{B}'\boldsymbol{diag}(\hat{\boldsymbol{D}})\boldsymbol{B} \;+\; \boldsymbol{B}'\boldsymbol{A}'\boldsymbol{V}^{-1}\boldsymbol{A}\boldsymbol{B}\right] \in \mathbb{R}^{K\times K}$. The change in θ required to balance a small change in D therefore implies that predicted deaths $\hat{D}$ must change as

$$\begin{aligned}\Delta\hat{D} \;&\approx\; \left[\boldsymbol{diag}(\hat{\boldsymbol{D}})\,\boldsymbol{B}\right]\;\Delta\theta \\ &\approx\; \left[\boldsymbol{diag}(\hat{\boldsymbol{D}})\,\boldsymbol{B}\right]\;\hat{\boldsymbol{Q}}^{-1}\boldsymbol{B}'\,\Delta D.\end{aligned}$$

Thus the "hat matrix" is $\left[\boldsymbol{diag}(\hat{\boldsymbol{D}})\,\boldsymbol{B}\right]\;\hat{\boldsymbol{Q}}^{-1}\boldsymbol{B}'$ and the effective degrees of freedom in a D-spline model is the sum of its diagonal elements

$$
\begin{aligned}
\text{df} &= \text{tr}\left(\left[\boldsymbol{diag}(\hat{\boldsymbol{D}})\, \boldsymbol{B} \right] \; \hat{\boldsymbol{Q}}^{-1} \boldsymbol{B}' \right) \\
&= \text{tr}\left(\hat{\boldsymbol{Q}}^{-1} \boldsymbol{B}' \boldsymbol{diag}(\hat{\boldsymbol{D}})\, \boldsymbol{B} \right) \\
&= \text{tr}\left(\left[\boldsymbol{B}' \boldsymbol{diag}(\hat{\boldsymbol{D}}) \boldsymbol{B} \; + \; \boldsymbol{B}' \boldsymbol{A}' \boldsymbol{V}^{-1} \boldsymbol{A} \boldsymbol{B} \right]^{-1} \; \boldsymbol{B}' \boldsymbol{diag}(\hat{\boldsymbol{D}})\, \boldsymbol{B} \right).
\end{aligned}
$$

This is identical in structure to the matrix derived in Eilers and Marx (1996) (Equations 20 and 26); here the Hessian of the D-spline penalty replaces the Hessian of the penalty for squared differences between consecutive parameters.