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

MPIDR WORKING PAPER WP 2012-022
JULY 2012

Calibrated Spline Estimation of Detailed Fertility Schedules from Abridged Data

Carl P. Schmertmann

This working paper has been approved for release by: Michaela Kreyenfeld (kreyenfeld@demogr.mpg.de), Deputy Head of the Laboratory of Economic and Social Demography.

Calibrated Spline Estimation of Detailed Fertility Schedules from Abridged Data

Carl P. Schmertmann

Max Planck Institute for Demographic Research &
Center for Demography and Population Health,
Florida State University

5 July 2012

Calibrated Spline Estimation of Detailed Fertility Schedules from Abridged Data

ABSTRACT

OBJECTIVE

I develop and explain a new method for interpolating detailed fertility schedules from age-group data. The method allows estimation of fertility rates over a fine grid of ages, from either standard or non-standard age groups. Users can calculate detailed schedules directly from the input data, using only elementary arithmetic.

METHODS

The new method, the calibrated spline (CS) estimator, expands an abridged fertility schedule by finding the smooth curve that minimizes a squared error penalty. The penalty is based both on fit to the available age-group data, and on similarity to patterns of ${}_1f_x$ schedules observed in the Human Fertility Database (HFD) and in the US Census International Database (IDB).

RESULTS

I compare the CS estimator to two very good alternative methods that require more computation: Beers interpolation and the HFD's splitting protocol. CS replicates known ${}_1f_x$ schedules from ${}_5f_x$ data better than the other two methods, and its interpolated schedules are also smoother.

CONCLUSIONS

The CS method is an easily computed, flexible, and accurate method for interpolating detailed fertility schedules from age-group data.

COMMENTS

Data and *R* programs for replicating this paper's results are available online at http://calibrated-spline.schmert.net

1. Introduction

Demographers like precise data for exact ages, but unfortunately we often get the opposite – noisy sample estimates aggregated into wide age groups. Worse, sometimes the age groups do not cover the entire range of interest for the behavior under study. With abridged, partial, or noisy data, demographic calculations often require interpolation and extrapolation of age-specific rates.

In this paper I introduce a method for fitting detailed fertility schedules to coarse, possibly noisy data. The method exploits a large new dataset, the Human Fertility Database (HFD), to identify empirical regularities in fertility schedules by single years of age 12-54. It then uses these regularities in a penalized least squares framework to produce simple rules for expanding grouped data (usually ${}_5f_x$ estimates) into detailed rates over an arbitrarily fine grid of ages that may extend outside the range of the original data (for example, below age 15 or above age 50).

The new method uses spline functions as building blocks, and identifies smooth fertility schedules that match group-level data closely while also conforming to patterns observed in the HFD. I call the result of the procedure a *calibrated spline* (CS) schedule. Its derivation uses some rather dense matrix algebra, but the end result is exceedingly simple: basic arithmetic with the grouped data and a set of predetermined constants.

2. Notation and Derivation of the Calibrated Spline Estimator

In the next two sections I explain and derive the CS estimator. Readers uninterested in the mathematical details may, without difficulty, skip ahead to the penultimate paragraph of the next section, beginning with *The key point is…*.

Suppose that the fertility schedule can be well approximated by a weighted sum of K continuous basis functions

$$(1) \qquad \phi(a) \approx \sum_{k=1}^{K} w_k b_k(a) = \underset{1xK}{b(a)'} \underset{Kx1}{w}$$

over the reproductive age range [α,β]. In many applications demographers use a fine grid of ages $\{a_1 \ldots a_N\}$ and assume that fertility is constant at some level f_i within each small interval $[a_i - \frac{1}{2}\Delta,\ a_i + \frac{1}{2}\Delta)$. In such applications the discrete version of φ is an Nx1 vector

$$(2) \qquad f = \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix} = \begin{bmatrix} b_1' \\ \vdots \\ b_N' \end{bmatrix} w = \mathbf{B}\, w$$

where b_i' is a 1xK vector containing the value of each basis function at $a=a_i$, and **B** is thus an NxK matrix of known constants.

In general, the $\{a_i\}$ grid can be arbitrarily fine, over any age range of interest, and there are many possible choices for the number and form of basis functions $\{b_k\}$. In the calculations in this paper, α=12, β=55, N=86, Δ=.50, there are separate fertility rates for intervals centered at 12.25, 12.75,…54.75. I use quadratic B-spline basis functions (de Boor 1978, Eilers and Marx 1996) over uniform knots at two-year intervals.[1]

[1] Specifically, basis functions come from the *bs*() function in *R* (R Core Development Team 2011), with arguments *x*=seq(12.25, 54.75, .50), *knots*=seq(12,54,2), and *degree*=2. I retain the third through twenty-first columns of the resulting matrix as an 86x19matrix **B.**

When fertility data is reported as averages for age groups (call the groups $A_1 \ldots A_g$), we need multipliers for aggregating *f*. The *N*x1 vector *f* is related to the *g* group averages by

$$(3) \qquad \begin{bmatrix} \bar{f}_1 \\ \vdots \\ \bar{f}_g \end{bmatrix} = \mathbf{G}\, f = \mathbf{G}\,\mathbf{B}\, w$$

where **G** is *g* x *N* with $G_{ij} = I[a_j \in A_i] / \#(a_j \in A_i)$, and I[·] is a 0/1 indicator function. The fine grid *f* is similarly related to single-year rates by

$$(4) \qquad \begin{bmatrix} {}_1 f_\alpha \\ \vdots \\ {}_1 f_{\beta-1} \end{bmatrix} = \mathbf{S}\, f = \mathbf{S}\mathbf{B}\, w$$

where $S_{ij} = \Delta \cdot I[\alpha + i - 1 \le a_j < \alpha + i]$.

3. Objective and Estimation Strategy

Suppose that we have a *g* x 1 vector of sample estimates for age group averages. Call this vector *y*. We want to estimate the *K* spline weights *w* (and ultimately, the *N* elements of the discretized schedule *f*) from the *g* estimates in *y*. This requires additional identifying information of some kind.

I propose two criteria for a good schedule *f*: it should (1) closely fit the observed data *y*, (2) have an age pattern similar to known schedules – specifically, to schedules downloaded from the Human Fertility Database (HFD 2012) and in the US Census International Database (Schmertmann 2003: File III). For these criteria, which I call *fit* and *shape* respectively, one can construct vectors of residuals that should be near zero for good schedules. These vectors are

(5)
$$\begin{aligned} \text{Fit:} \quad & \varepsilon_f = y - \mathbf{G}\, f = y - \mathbf{GB}\, w \\ \text{Shape:} \quad & \varepsilon_s = \mathbf{M}\,\mathbf{S}\, f = \mathbf{MSB}\, w \end{aligned}$$

The **M** matrix for shape residuals has a complicated construction, but a simple interpretation. Construction is as follows. I first assemble a 43x530 matrix **F**, comprising 304 single-year ASFR schedules from the HFD over ages 12…54[2], plus an additional 226 estimated single-year schedules from the US Census International Database (IDB) using the quadratic spline model and coefficients from Schmertmann (2003:File III).[3] Singular value decomposition **F**=**UDV**' yields orthonormal principal component vectors in **U**'s columns. The first three of these columns (call this 43x3 matrix **X**) account for approximately 95% of the variation in **F**, in the sense that projections of any single-year schedule *s* onto the column space of **X** have small errors

(6)
$$e = s - \hat{s} = \left(\mathbf{I}_{43} - \mathbf{P}\right)s$$

where **P**= **X**(**X'X**)$^{-1}$**X'** is the projection matrix for the column space of **X**.

Defining **M**=(**I**$_{43}$-**P**), shape residuals in Equation (5) represent the portion of a single-year schedule that is unexplained by linear combinations of principal components. In other words, shape residuals ε_s in Equation (5) are large for single-year schedules that have age patterns

[2] The HFD version that I used has 1480 single-year schedules, many of which are from the same country in consecutive calendar years. In order to limit the overcounting of highly correlated schedules, I used every fifth year from each population – e.g., Austria 1953, 1958, …, 2008, Bulgaria 1949, 1953, …, 2009, and so on.

[3] It is slightly clumsy to split the five-year IDB schedules into approximate single-year schedules in order to include them in the analysis, but adding these schedules is important. The HFD does not yet include countries from Africa and Asia that have very distinct age patterns – in particular African schedules often have relatively high fertility at ages 35+, and some East Asian schedules have extremely low fertility at ages below 25. Estimation of SVD principal components from a matrix that includes the wider variety of patterns in the IDB produces a much more representative set of "typical" age schedules.

unlike those observed in the HFD and IDB.[4]

Each criterion can be converted into a scalar index of a schedule's "badness" by calculating an appropriately weighted sum of squares. These scalar penalty terms have generic form

(7) $$P_c = \varepsilon_c' \mathbf{V}_c^{-1} \varepsilon_c \qquad c \in \{f, s\}$$

where $\mathbf{V}_c = E[\varepsilon_c \varepsilon_c']$ is the covariance of ε_c.

The covariance matrix of fitting errors ε_f can be approximated logically. Supposing that sample estimates *y* come from groups with 2000 women at each single year of age, and that a typical rate in a five-year interval is about 0.10, then with independent sampling errors across groups the covariance of ε_f is[5]

(8) $$\mathbf{V}_f = E(\varepsilon_f \varepsilon_f') = 2x10^{-6}\, \mathbf{I}_g$$

These assumptions are crude, but the results are not very sensitive to them. The main point is that with large sample sizes, schedules that fit age group averages poorly get extremely heavy penalties.

For the covariance of shape residuals, we refer to the single-year schedules in the HFD. For each of the 1480 schedules (*s*) in the HFD single-year data, one can calculate $e_s = \mathbf{M}s$. The average outer product of these HFD shape residuals serves as a covariance estimate:

(9) $$\mathbf{V}_s = \overline{(e_s e_s')}$$

[4] More precisely, a schedule *f* has large shape residuals when $\mathbf{S}f$ lies far from the column space of $\mathbf{X}$. It is possible for *f* to have low shape residuals even if it is unlike any observed schedule, if *f* is well approximated by a combination of principal components that has no counterpart in the database.

[5] The calculation assumes that the number of births (B) to $5W$ women in a five-year age group with true rate *f* is a Poisson random variable with mean and variance $5Wf$. A sample estimate $y_k = B/5W$ therefore has variance $f/5W$. I assume $5W$=10000, somewhere between the typical sizes in censuses and surveys.

$\mathbf{V}_s$ provides information about which ages are likely to have large or small residuals, and about the age patterns among those residuals.[6]

Summing the penalties produces a single index that is appropriately calibrated to the available information about errors[7]:

$$
\begin{aligned}
P(w) &= P_f + P_s \\
&= \varepsilon_f' \mathbf{V}_f^{-1} \varepsilon_f + \varepsilon_s' \mathbf{V}_s^{-1} \varepsilon_s \\
&= (y - \mathbf{GB}w)' \mathbf{V}_f^{-1} (y - \mathbf{GB}w) + (\mathbf{MSB}w)' \mathbf{V}_s^{-1} (\mathbf{MSB}w) \\
&= w' \mathbf{C}_1 w - 2 w' \mathbf{C}_2 y + y' \mathbf{V}_f^{-1} y
\end{aligned} \tag{10}
$$

where

$$\mathbf{C}_1 = \mathbf{B}'\mathbf{G}'\mathbf{V}_f^{-1}\mathbf{GB} + \mathbf{B}'\mathbf{S}'\mathbf{M}'\mathbf{V}_s^{-1}\mathbf{MSB} \tag{11}$$

and

$$\mathbf{C}_2 = \mathbf{B}'\mathbf{G}'\mathbf{V}_f^{-1} \tag{12}$$

Because $\mathbf{C}_1$ is positive definite, expression in Equation (10) has a unique minimum when weights are

$$w^* = \mathbf{C}_1^{-1}\mathbf{C}_2 y = \mathbf{K}_w y \tag{13}$$

Thus the combination of basis function that minimizes the joint criterion in Equation (10) is a vector that I call the *calibrated spline* (CS) fit:

$$f^* = \mathbf{B}\, w^* = \mathbf{B}\mathbf{K}_w y = \mathbf{K}\, y \tag{14}$$

The key point is that this complex derivation leads to a simple result: *the optimal schedule f is a linear function of the observed data y*. The *N* x g matrix **K** contains predetermined constants, so that can write

---

[6] Adding a small constant to each diagonal element of $\mathbf{V}_s$ before inverting stabilizes results considerably. I add 0.1 times the median value of the diagonal elements from Equation (9).

[7] There is also a natural Bayesian interpretation for this index: the fitting penalty comes from the log likelihood of a multivariate normal distribution, and the shape penalty terms come from an improper multivariate normal prior.

the CS vector f* as a weighted sum of *g* columns:

$$(15) \qquad f^* = \begin{bmatrix} \vdots \\ \mathbf{K}^{(\text{column } 1)} \\ \vdots \end{bmatrix} y_1 + \cdots + \begin{bmatrix} \vdots \\ \mathbf{K}^{(\text{column } g)} \\ \vdots \end{bmatrix} y_g$$

In principle, this framework allows a demographer to create simple arithmetical rules for transforming fertility estimates from any set of *g* age groups into a schedule over an arbitrarily fine grid of *N* rates over any age span of interest. The method is particularly simple because the "parameters" for the empirical model are the estimated age-group fertility rates themselves, so that fitting the model requires only multiplication and addition.

4. Example Fits with IDB and HFD data

The CS method outlined above works for any set of age groups, but I deal with two specific examples in the rest of this paper – cases in which (a) data are available for *g*=7 age groups 15-19 through 45-49, as in the US Census International Database (IDB) and many other datasets, or (b) data are available for *g*=9 five-year age groups 10-14 through 50-54, as in the HFD. For the *g*=7 case, the 86x7 matrices of constants **K** and $\mathbf{K}_w$ appear in comma-delimited supplemental files *K7.csv* and *Kw7.csv*, respectively; for the *g*=9 case, the corresponding 86x9 matrices appear in *K9.csv* and *Kw9.csv*. Readers can also adapt the supplemental programs to construct constants for other combinations of age grids and age groups.

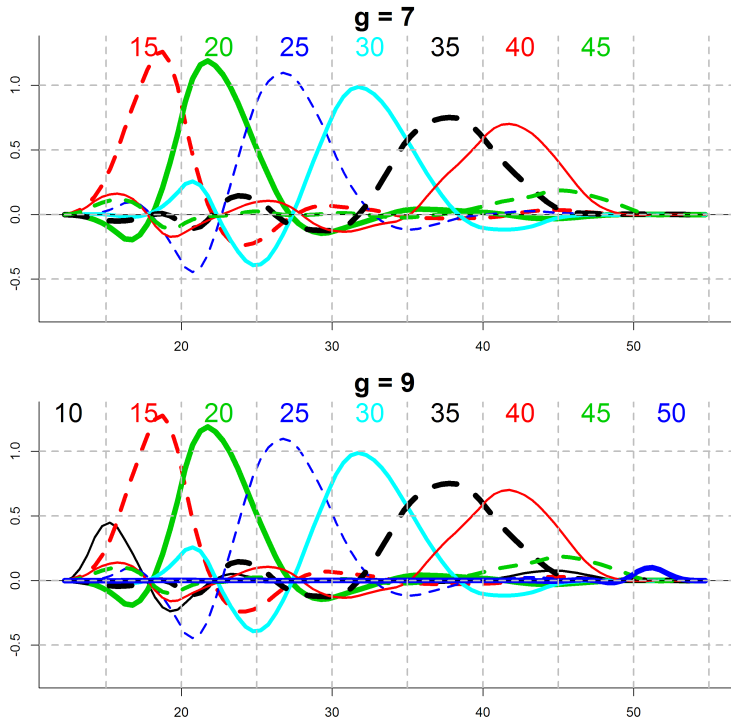

Figure 1. Empirical basis functions for a fitted schedule at half-year intervals over [12,55]. Each line represents one column of K. Input data are estimated average rates for five-year age groups (g=7 and g=9 in top and bottom panels, respectively).

Figure 1 illustrates **K** for the g=7 and g=9 cases, by plotting each column as a function of age. For example, a unit increase in estimated ${}_5f_{15}$ changes f^* values at various ages by the height of the line labeled "15". A unit increase in estimated ${}_5f_{20}$ changes f^* according to the line labeled "20", and so on. Note that the range of estimated fertility f^* may extend beyond that spanned by the input data: in the g=7 case the procedure produces estimated ASFRs below age 15 and above age 50, based on known regularities in the age pattern of rates.

Using Equation (14) or (15), basis functions in Figure 1 are multiplied by the observed y values and then summed to produce complete CS schedules over [α,β]. The top panel of Figure 2 illustrates the expansion of a set of g=7 five-year estimates into half-year intervals, using IDB data from Uruguay. The input data for Uruguay are

$$y_{URU} = 10^{-3} \times (49\ 116\ 135\ 99\ 54\ 16\ 2)'$$

Multiplying these values by the columns of **K** and summing produces an 86x1 vector $f^*=\mathbf{K}y$ for rates at half-year intervals over 12-55, shown in the top panel.

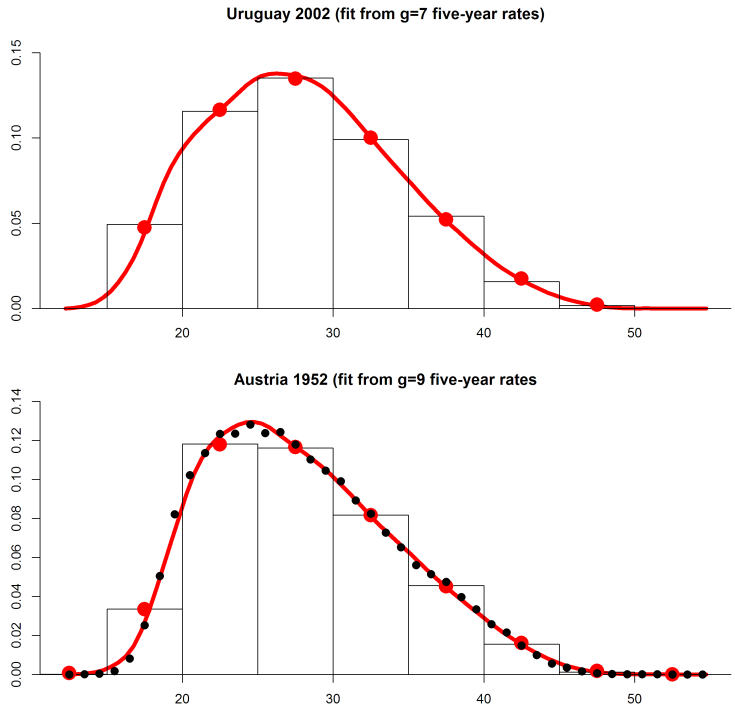

Figure 2. Calibrated spline (CS) schedules for Uruguay 2002 (g=7, top panel) and Austria 1952 (g=9, bottom panel), estimated at half-year intervals over [12,55]. Input data y in both cases are five-year rates in the histograms. Large circles represent the average of the CS schedule over a five-year interval. Small dots in the bottom panel represent the original single-year data from Austria, from which the five-year rate vector *y* was calculated.

By comparing the height of the histogram to that of the large dots, one can see that the age-group averages for the CS model do not exactly replicate the input data. For example, the average of the CS schedule over ages 35-39 in Uruguay is slightly lower than the original ${}_5f_{35}$ value of .054. This occurs because minimizing the penalty index in Equation (10) requires tradeoffs between model fit and the shape of schedule. The tradeoff for Uruguay was typical, in the sense that over all 226 IDB schedules, Uruguay's mean squared fitting error was closest to the

median: half of IDB schedules have better CS fits to the ${}_5f_x$ data, and half have worse.

The bottom panel of Figure 1 illustrates the CS schedule for Austria's 1952 period fertility, calculated from g=9 five-year rates for age groups 10-14 through 50-54:

$$Y_{AUT1952} = 10^{-3} \times (.14\ 34\ 118\ 116\ 82\ 46\ 16\ 1\ .02)'$$

In this case one can check the accuracy of the CS fit, because Austria 1952 is one of 586 HFD schedules with ${}_1f_x$ values over x=12…54 that come directly from original data (rather than being interpolated from ${}_5f_x$ or other group averages). These original ${}_1f_x$ values appear as black dots in the lower panel of Figure 1, and it is clear that for this schedule the CS fit to the histogram matches the single year data well: the root mean squared error (RMSE) across all 43 ages is .0021. This is identical to the median RMSE over all of the 586 complete single-year schedules in the HFD, so that the Austria 1952 fit is also typical: half of CS fits from five-year data match the original single-year schedule less accurately, while half are more accurate.[8]

5. Comparative Accuracy of CS vs. Other Methods

Researchers from Columbia University and the UN Population Division (Liu et al. 2011) recently used HFD data to compare the accuracy of several interpolation methods for fertility schedules. They concluded that the best overall method for recovering single-year age-specific rates

[8] 99% of fitted single-year rates with the CS model are within .01 of the equivalent HFD data. The largest CS fitting error over the 586 complete single-year schedules is for 19-year-olds in East Germany 1965: true and fitted rates were .173 and .139, respectively. This error arises because East German 1965 rates had an unusually steep rise over ages 16-20, which the CS model does not replicate precisely. East Germany 1965 also had the highest RMSE over all ages: .0068.

from five-year averages was a variant[9] of Beers's ordinary osculatory interpolation method (Shryock and Siegel 1975:Table C3).

In addition, the HFD project itself has a protocol for splitting age-group averages into single-year rates. HFD interpolation calculates the logit of standardized cumulative fertility $Y_x=ln[F_x/(TFR-F_x)]$ at age group boundaries, interpolates values between the boundaries using a Hermite cubic spline, and then differences anti-logits to arrive at single-year rates:

$$_1 f_x = TFR \cdot \left[(1+e^{-Y_{x+1}})^{-1} - (1+e^{-Y_x})^{-1}\right]$$

Jasiolioniene et al. (2011:27-31) has additional details, and the reader can see my exact implementation of the protocol in the supplemental program files. The principal disadvantage of the HFD splitting procedure is that the underlying model implies an $f(x)$ schedule that is not smooth, because it has discontinuous slopes at the breaks between age groups. The next figure illustrates this problem.

Because these two interpolation approaches have been selected in earlier "competitions", it is valuable to compare them to the CS approach over a wide range of schedules[10]. Figure 3 offers an initial example for a single schedule, showing the interpolated fits from the three methods for Scotland in 2004, and a summary of the fitting errors.

---

[9] The Beers method often generates negative rate estimates at ages <20 and 40+. In the Liu et al. (2011) variant, negative rates are replaced with exponential curves, which are then rescaled so that the five-year age group totals match the input data.

[10] I experimented with cross-validation in the construction of the **X** and **K** matrices. For example, cross-validated fits for Scotland came from a model built without any Scottish data, and so on for each country. Cross-validated results are not reported here, but they made only a trivial difference: RMSEs were identical to 5 decimal places, with or without the inclusion of own-country data in model construction. I conclude that inclusion of own-country data is not an important concern in evaluation of the CS model's performance.

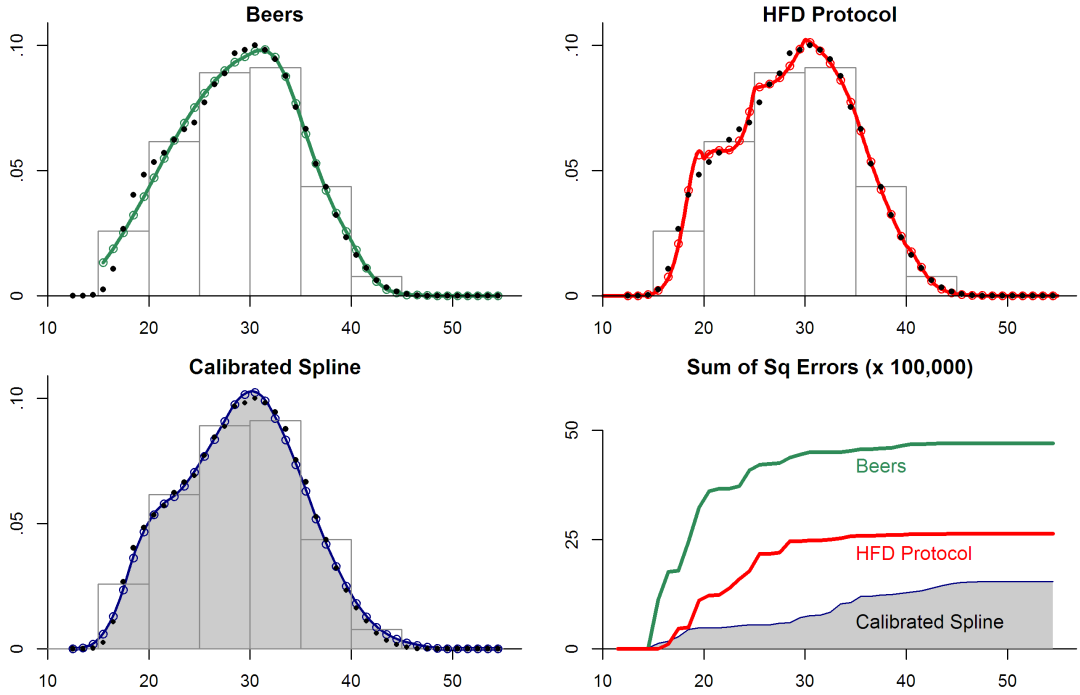

Figure 3. Alternative fits from the g=9 five-year rates for Scotland 2004. Open circles are interpolated ${}_1f_x$ values. Solid dots are original single-year data from which five-year rates were calculated. Bottom right panel illustrates cumulative sum of squared fitting errors over age.

Several features of Figure 3 deserve mention. All three methods produce interpolated schedules that fit the single-year rates well. The HFD schedule is notably less smooth than the other two fits, because of slope discontinuities at the boundaries of age intervals.[11] For the Scotland 2004 schedule the CS method is generally more accurate at ages below 30, and unlike the other two approaches it captures the subtle inflection in rates the early 20s. The Beers and HFD models fit the single-year data better at ages 40+ (in part because of the Beers adjustment that Liu et al. make for negative predicted rates at ages 48-52 with these input data). Overall, the CS errors are smallest, and Beers errors are largest.

[11] Discontinuities in the slope of the HFD interpolation arise by construction. In the HFD approach, cumulative fertility at age x is $F(x)=\mathrm{TFR}\cdot g[Y(x)]$, where $g(u)$ is a continuous function $(1+e^{-u})^{-1}$ and $Y(x)$ is a piecewise Hermite cubic spline that has (1) continuous first derivatives and (2) discontinuous second derivatives at age-group boundaries . As the derivative of $F(x)$, age-specific fertility is therefore continuous, with discontinuous first derivatives.

Moving from a single example to a global summary, Figure 4 summarizes the errors for the three methods over all 586 HFD schedules with known single-year rates, disaggregated by age. Notice

1. The vertical scale shows that average errors are very small for all methods.

2. The sawtooth pattern of errors at ages below 35 shows that all interpolation methods fit single-year data better in the middle of five-year intervals than they do at the edges. This is an arithmetical property of interpolation when the underlying curve is approximately linear over five-year intervals: both the fitted and true schedules are likely to be close to the age-group average at the center of the age range.

3. The pattern of comparative errors by age seen for Scotland 2004 in Figure 3 holds up across all schedules: calibrated spline fits are much better at ages below 40, while HFD and Beers fits (after fixing Beers negative values) are slightly better at ages above 40.

4. Most importantly, the total of average errors (all ages combined) is lowest for the CS approach.

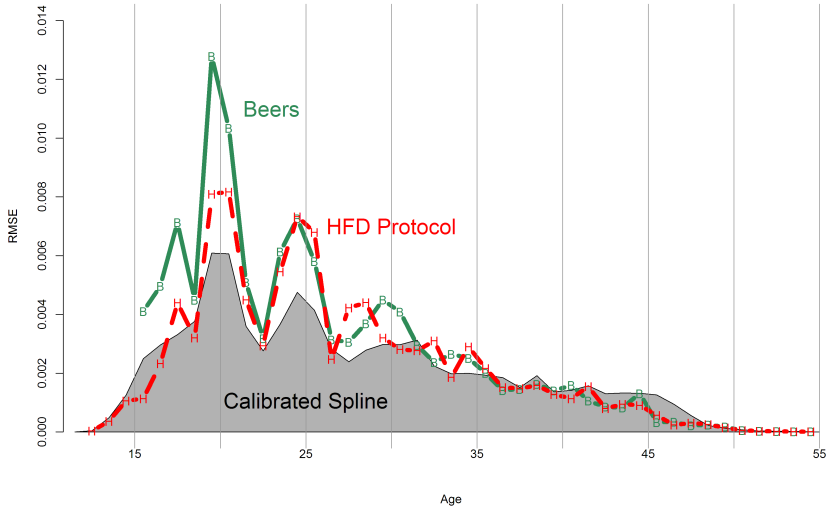

Figure 4. Root mean squared fitting errors by age. Calculated over HFD cells with original (rather than estimated) single-year rates.

It is also useful to summarize errors over different dimensions. Figure 5 offers a second global comparison of the methods, this time aggregating over ages and showing the average RMSE by country. Average interpolation errors are lowest for the CS method in 17 of the 20

populations, and for the HFD protocol in the other three (East Germany, Czech Republic, Netherlands). Once again, all three methods perform very well, and again the overall ranking places the CS method first, HFD second, Beers third.

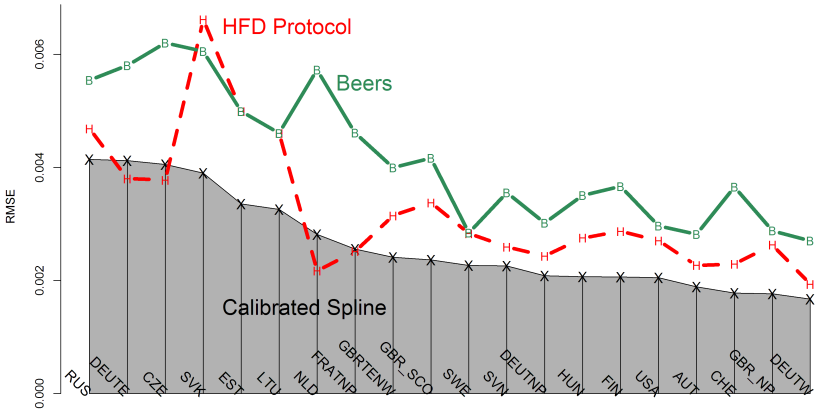

Figure 5. Root mean squared fitting errors by country. Calculated over HFD cells with original (rather than estimated) single-year rates. Abbreviations from HFD.

Table 1 provides a final comparison of the methods, with slightly more quantitative detail about some of the potential problems that may occur when interpolating rates from abridged data. Section A of the table contain fitting errors ($\cdot 10^4$) by age group and interpolation method, for (age,period,country) cells where the HFD's ${}_1f_x$ values come from original data sources rather than from a splitting algorithm. The CS method performs best overall, but at high maternal ages its fits are slightly worse than those of the adjusted Beers or HFD splitting algorithms.

Section B reports measures of the roughness or wiggliness of interpolated schedules, summarizing second differences by age $\left({}_1f_{x+2} - {}_1f_{x+1}\right) - \left({}_1f_{x+1} - {}_1f_x\right)$ with root mean squared values ($\cdot 10^4$) across all the 1480 HFD schedules (interpolation from g=9 age groups) and all 226

IDB schedules (g=7). Lower index values in Section B correspond to sets of interpolated schedules with fewer up-and-down wiggles and fewer local maxima in the interpolated single-year rates. Again the CS method performs best, producing smoother schedules.

Section C of Table 1 includes information on a performance criterion for which the CS method is inferior to the other two approaches: negative rate estimates. With the test data at hand, each method produces 1706x43=73358 single-year rate estimates. The HFD splitting method uses logarithmic pre-processing before interpolating, so that by construction none of its 73358 estimates are negative. In the original Beers approach (not shown in the table) approximately 12% of the estimates are negative and 3% are below -.005. However, the Liu et al. variant used here eliminates all negative values through a post-processing algorithm.

Table 1. Error summaries for alternative interpolation methods. RMSEs calculated over cells with known single-year data. All other calculations refer to interpolated fits over ages 12-54 from all 1706 available $_5f_x$ schedules (1480 in HFD + 226 in IDB). Shaded cells correspond to the best- performing method for each error criterion.

	HFD	Beers	Calibrated Spline
A. Fitting Errors (RMSE x 10^4)			
All Ages	34	42	27
12-24	49	72	38
25-34	37	36	28
35+	11	11	13
B. Roughness of Fitted Schedule (Root Mean Squared 2nd Difference x 10^4)			
HFD (g=9)	61	81	40
IDB (g=7)	152	62	42
C. Negative Values (Percent of all estimated rates)			
< 0	0	0	3
< -.0005	0	0	1
< -.0050	0	0	0

In contrast, without adjustment 3% of the CS-estimated fertility rates are negative. Although this is of course logically impossible, the vast majority of these negative CS rate estimates are negligibly different from zero. For example, one of the 43 CS-interpolated ${}_1f_x$ values for the Scotland 2004 schedule in Figure 3 is negative (${}_1f_{49}$ = -.0000034), but is so close to zero that its direct use in calculations such as TFR, mean age of childbearing, etc. would cause no meaningful problems. As seen in Section C, only 1% of CS rates are below -.0005 (i.e., negative even rounding to three decimal places) and none are below -.005.

Small negative estimates are a minor problem for the CS method, small enough that I have not applied any post-processing to the CS rates in any of this paper's tables or figures. However, it is possible to use a very simple post-processing procedure on CS rates – namely, after calculating $f^*=\mathbf{K}y$, replace any negative values with zeroes. This is computationally much simpler than the Liu et al. post-processing algorithm for Beers rates, and it would not alter any of the comparisons in Sections B and C of Table 1.[12]

In sum, all three methods are very good, but the CS method performs slightly better – over almost all countries, and over the ages at which fertility rates are highest. Interpolated CS schedules are smoother and fit known data better. CS calculation is also much simpler than the HFD splines or the Beers variant used by Liu et al. (2011), because it does not require complex adjustments for edge effects and negative values.

[12] With truncation at zero, the Calibrated Spline column of Table 1 would remain unchanged, except that the percentages in Section C would all be zero, and the IDB smoothness measure would change to from 42 to 43.

6. Discussion

I have presented applications of the calibrated spline model for only two specific cases, but the general framework is extremely flexible. In principle one can construct expansion constants **K** that map input data from any set of age groups onto any fine grid of ages. The input age groups may be incomplete (e.g., {25-29,35-39,40-44,45-54}), irregularly spaced ({12-14,15-19,20-24,25-34,…}), or even overlapping ({15-17,15-24,…}).

The CS model fits observed schedules well, outperforming some alternative methods that have done well in earlier research. It is also much simpler to estimate. Given the **K** constants (which in most cases are the ones already provided in this paper and the accompanying data files), fitting a detailed ASFR schedule requires only basic arithmetic. Unlike the Beers method and other generic polynomial fitting methods that are not designed specifically for fertility estimation, post-estimation tweaks for negative fitted rates at the highest and lowest maternal ages are rarely necessary. Unlike the HFD splitting protocol, it does not require the user to perform a multi-step mathematical procedure to get from data *y* to fitted schedule *f**.

Although not explicitly Bayesian, the CS estimation approach makes heavy use of *a priori* information. The penalized least squares criterion gives priority to fertility schedules that not only fit input data well, but that also match historical or contemporary patterns seen in large databases. The technique of identifying such patterns through singular value decomposition of a large data array is not new in demography (for example, it is the basis of the Lee-Carter [1992] mortality model), but to

my knowledge researchers have not previously used such patterns in a simple, least-squares method like that presented here.

References

C de Boor, 1978. *A Practical Guide to Splines*. Springer-Verlag. New York.

PHC Eilers and BD Marx, 1996. "Flexible Smoothing using B-Splines and Penalized Likelihood". *Statistical Science* 11:89-121.

A Jasilioniene, DA Jdanov, T Sobotka, EM Andreev, K Zeman, and VM Shkolnikov, 2011. "Methods Protocol for the Human Fertility Database". 28 Oct 2011. Online at http://www.humanfertility.org/Docs/methods.pdf

RD Lee and LR Carter, 1992. "Modeling and Forecasting U.S. Mortality". *Journal of the American Statistical Association* 87(419):659-671.

Y Liu, P Gerland, T Spoorenberg, K Vladimira, and K Andreev, 2011. "Graduation methods to derive age-specific fertility rates from abridged data: a comparison of 10 methods using HFD data". Presentation at the First Human Fertility Database Symposium, Max Planck Institute for Demographic Research, Rostock, November 2011. Extended abstract at http://www.humanfertility.org/Docs/Symposium/Liu-Gerland%20et%20al.pdf (downloaded 10 June 2012).

R Development Core Team, 2011. *R: A Language and Environment for StatisticalComputing*. R Foundation for Statistical Computing. Vienna. http://www.R-project.org

CP Schmertmann, 2003. "A system of model fertility schedules with graphically intuitive parameters". *Demographic Research* 9/5:81-110. http://dx.doi.org/10.4054/DemRes.2003.9.5

HS Shryock and JS Siegel, 1975. *The Methods and Materials of Demography*, Vol 2. Third Printing (rev.). US Bureau of the Census, US Government Printing Office. Washington DC.

Project Website with Data, Programs, and additional Details

http://calibrated-spline.schmert.net

Appendix: Moment Calculations from Age Group Data

One possible use of the empirical model is estimation of moments of the continuous fertility schedule from grouped data. This type of approximation might be especially useful with indirect methods.

Begin by defining the function

(A1) $$Q_n(x) = \int_{\alpha}^{x} a^n \ \phi(a)\ da$$

which can be approximated as

$$\begin{aligned} Q_n(x) &\approx \sum_{i:\,a_i<x} a_i^n \ \phi(a_i)\Delta \\ &= \sum_{i:\,a_i<x} a_i^n \ f_i \ \Delta \\ &= \sum_{i:\,a_i<x} a_i^n \, b_i' \, w \, \Delta \qquad \text{(A2)} \\ &= \left(\sum_{i:\,a_i<x} a_i^n \, b_i' \mathbf{K}_w \Delta \right) y \\ &= c_n(x)' \, y \end{aligned}$$

where $c_n(x)$ is a g x 1 vector of known constants.

With different (x,n) combinations, Equation (A2) produces different moments of the fertility function. Table A1 shows some of the calculated constants for the g=7 case; a more complete set of constants is available in supplemental file *Cdata.csv*.

By definition $Q_0(\infty)$ is a schedule's total fertility (TFR), and $Q_1(\infty)/Q_0(\infty)$ is its mean age of childbearing μ. In the case of the Uruguay 2002 data shown earlier, for example, we can approximate these quantities as

TFR = $Q_0(\infty)$ ≈ 4.99(.049) + … + 1.45(.002) = 2.360

μ = $Q_1(\infty)$ / $Q_0(\infty)$ ≈ [89.11(.049) + … + 60.21(.003)] / 2.360 = 28.04

Similarly, one can approximate conditional moments such as average parity of women 30-34 [$Q_0(32.5)$] and the average age at which they had their previous births [$Q_1(32.5)$/ $[Q_0(32.5)]$. With the Uruguay data these moments would be

$$P_{30\text{-}34} \approx Q_0(32.5) \approx 4.98(.049) + \ldots + 0.17(.002) = \mathbf{1.781}$$

$$\mu_{30\text{-}34} \approx Q_1(32.5) / Q_0(32.5) \approx [88.14(.049) + \ldots + 2.19(.002)] / 1.781 = \mathbf{25.21}$$

Calculations like this can be important for time allocation with indirect methods. For example, from the five-year rate schedule for Uruguay, moment approximations imply that with a cohort fertility schedule with this shape, women 30-34 interviewed in a survey would have had their births an average of 32.50-25.21 = 7.29 years earlier.

Table A1. Some *c* multipliers for the g=7 case

	15-19	20-24	25-29	30-34	35-39	40-44	45-49
n=0 (TFR)							
x = 17.5	1.90	-0.46	0.20	-0.03	-0.16	0.50	0.36
x = 22.5	5.52	2.85	-1.00	0.60	-0.34	0.02	0.11
x = 27.5	4.73	5.47	2.68	-0.65	-0.03	0.37	0.15
x = 32.5	4.98	4.95	5.33	2.70	-0.42	-0.08	0.17
x = 37.5	4.94	5.05	4.92	5.49	2.18	0.11	0.09
x = 42.5	4.85	5.12	4.89	5.15	5.01	2.89	0.41
x = 47.5	4.98	5.00	4.97	4.95	5.42	4.88	1.23
x = ∞	4.99	4.99	4.97	4.95	5.44	4.94	1.45
n=1 (TFR · μ)							
x = 17.5	30.84	-7.40	3.19	-0.47	-2.50	7.71	5.55
x = 22.5	100.00	62.13	-21.41	12.42	-6.20	-1.73	0.71
x = 27.5	80.57	124.92	73.05	-18.63	1.23	7.20	1.65
x = 32.5	88.14	109.44	150.09	83.68	-10.28	-6.57	2.19
x = 37.5	86.59	113.06	135.82	179.67	82.04	0.78	-0.27
x = 42.5	82.78	115.68	134.83	165.62	194.25	112.98	12.81
x = 47.5	88.96	110.22	138.08	157.01	212.21	201.26	49.51
x = ∞	89.11	110.13	138.09	156.65	213.10	203.75	60.21