SPLINE INTERPOLATION FOR DEMOGRAPHIC VARIABLES: THE MONOTONICITY PROBLEM

Len Smith,† The Australian National University
Rob J. Hyndman, Monash University
Simon N. Wood, The University of Glasgow

In demography, it is often necessary to obtain a monotonic interpolation of data. A solution to this problem is available using the Hyman filter for cubic splines. However, this does not seem to be well known amongst demographers, and no implementation of the procedure is readily available. We remedy these problems by outlining the relevant ideas here, and providing a function for the R language.

Keywords: interpolation, splines, monotonic, age groups, age distribution, deaths, data bank

In 1977, McNeil, Trussell and Turner published a paper in *Demography* outlining the use of cubic splines for interpolating demographic variables. Splines are polynomials of specified degree (usually cubics) which are fitted to each of the intervals in the data, and forced to be smooth and continuous at the joining points or 'knots'. The term 'spline' comes originally from the name of a flexible laminated wooden draughtsman's rule, used for example by railway engineers to design smooth curves in railway tracks.

A typical application of spline interpolation in demography is the estimation of a single-year age distribution from five-yearly or other regularly or irregularly grouped data. In these applications the only exact values of the variable known are at the boundaries of the age groups, and for this reason it is necessary to interpolate the cumulative or integrated function and obtain the individual age groups by differencing.

Suppose we observe data at points (x_i, y_i) for $i=1,\ldots,n$. For example, x_i may represent ages in five-year intervals and y_i may represent cumulative deaths for ages up to and including x_i. Then a cubic spline is a curve $f(x)$ interpolating all points and consisting of cubic polynomials between each consecutive pair of knots x_i and x_{i+1}. The parameters of the cubic polynomials are constrained so that $f(x)$ is continuous and has continuous derivatives $f'(x)$ and $f''(x)$, thus causing it to be smooth. To define the curve uniquely, it is necessary to specify two additional constraints. Usually, this is done by specifying the form of the function at the two

† Address for correspondence: Australian Centre for Population Research, The Australian National University, ACT 0200, Australia. Email: len@coombs.anu.edu.au.

extremes, x_1 and x_n. For example, a natural spline is obtained by requiring $f''(x_1)=f''(x_n)=0$, thus making the curve linear at the extremes.

Wilmoth (2002) noted in a recent technical paper which uses splines to interpolate deaths, that problems can arise in sections of the curve where the slope changes rapidly, such as where the mortality curve flattens out in the second year of life. A spline fitted to the cumulative distribution of deaths will often not be monotonic at this point, so the differences (the single year of age estimates) will be negative, clearly an impossible result.

Wilmoth chose two additional constraints in order to reduce the likelihood of non-monotonicity. Specifically, he imposed a constraint on the slope of the function at age 1 to equal half the increment between 1 and 5. He based this on an observation that about half the deaths between 1 and 5 occur during the second year of life. At upper ages he constrains the slope to be zero. However, these constraints do not guarantee monotonicity and Wilmoth (2002: 53) concludes that there seems to be no reliable solution at older ages: the spline function often starts to decline with the open age group.

In fact, methods for ensuring monotonicity of an interpolating spline already exist. One solution developed has involved the use of the 'Hyman filter' (Hyman 1983; Dougherty, Edelman and Hyman 1989) which ensures that the interpolating spline remains monotonic by imposing alternative constraints on the derivatives of the piecewise cubics at the expense of possibly sacrificing some smoothness. Specifically, the second derivative of the curve will no longer be continuous at those knots at which the filter has changed first derivatives. If the interpolating spline is already monotonic, the Hyman filter leaves it unchanged.

Although smooth monotonic interpolation is implemented in some statistical code libraries, for example in the NAG library, it is not available in the main statistics packages, and it is not available as a standard option in R. We believe that R is the language of choice for modern applied statistics including statistical demography (Dalgaard 2002; Maindonald and Braun 2003). R is an open-source programming environment for data analysis and graphics. In only a decade it has become a *de facto* standard for statistical analysis, and provides cutting-edge statistical methods at no cost and for a number of platforms.

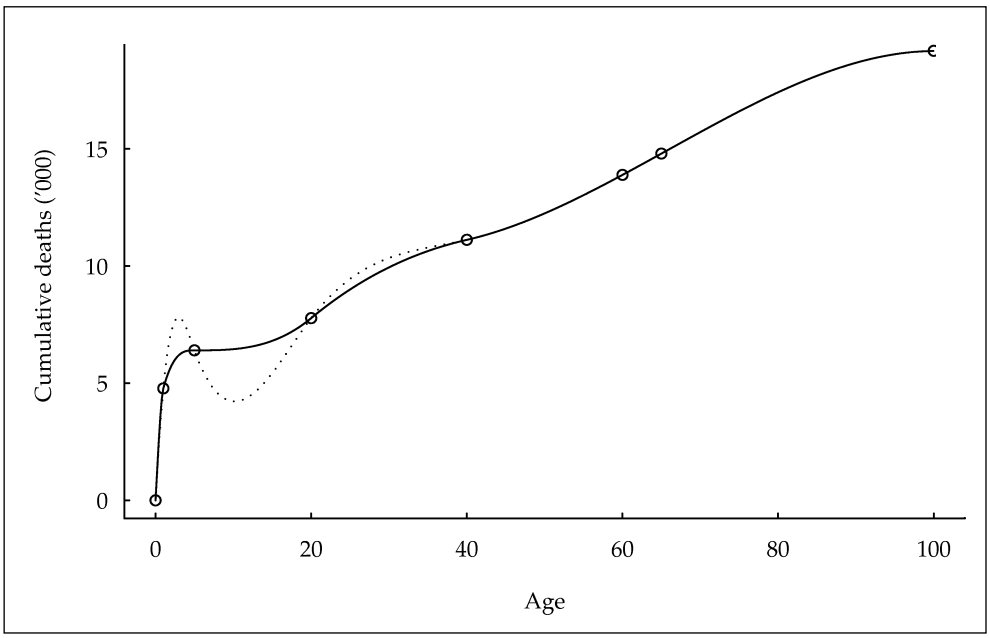
We provide an implementation of the Hyman filter in R code on the website http://www.personal.buseco.monash.edu.au/~hyndman/Rlibrary/. The application of the method is illustrated in the following section with an Australian data set.

Application

As part of a project to extend the Australian Demographic Databank (Brown and Hall 1978; Krishnamoorthy and Derrick 1983) back to 1901, it was necessary to obtain single-year age distributions of deaths and populations from published data which used grouped ages. In the process of producing single-year death rates from the published grouped data we faced problems similar to those described by Wilmoth. For both deaths and populations, we needed to bring a variety of irregularly grouped data to single-age format.

Figure 1 shows the cumulative deaths of Australian females in 1901 with an unconstrained spline interpolation shown as a dotted line. The available data (shown as circles) were for age groups 0, 1–4, 5–19, 20–39, 40–59, 60–64 and 65+. We

Figure 1 Spline interpolations[a] of cumulative female deaths, Australia, 1901



a The unconstrained spline interpolation is shown as a dotted line. The monotonic spline interpolation is shown as a solid line. For most ages, the curves coincide.

have arbitrarily set the upper age limit to 100 in computing spline interpolation. Some interpolated single-year differences between ages one and 20 will clearly be negative because of the negative slope in the spline function. In contrast, the monotonic spline function (shown as a solid line) produces single-year differences which, of necessity, are all positive. Similar interpolation was required for data in other years.

References

Brown, H.P. and A.R. Hall. 1978. *Australian Demographic Databank, Volume I: Recorded Vital Statistics 1921–1976. Volume II: Population Estimates and Demographic Rates 1921–1976*. Canberra: Department of Economics, Research School of Social Sciences, The Australian National University

Dalgaard, P. 2002. *Introductory Statistics with R*. New York: Springer.

Dougherty, R.L., A. Edelman and J.M. Hyman. 1989. Nonnegativity-, monotonicity-, or convexity-preserving cubic and quintic Hermite interpolation. *Mathematics of Computation* 52: 471–494.

Hyman, J.M. 1983. Accurate monotonicity preserving cubic interpolation. *SIAM Journal on Scientific Computing* 4(4): 645–654.

Krishnamoorthy, S. and B. Derrick. 1983. *Australian Demographic Databank, Volume III: Recorded Vital Statistics, Population Estimates and Demographic Rates 1976–1981*. Canberra: Department of Demography, Research School of Social Sciences, The Australian National University.

Maindonald, J. and J. Braun. 2003. *Data Analysis and Graphics using R: An Example-based Approach*. Cambridge: Cambridge University Press.

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

Wilmoth, J.R. 2002. Methods protocol for the human mortality database, revised 1 October 2002. http://www.mortality.org/. Accessed 11 November 2003.