Graduation methods to derive age-specific fertility rates from abridged data: a comparison of 10 methods using HFD data

Ye Liu[1], Patrick Gerland[2], Thomas Spoorenberg[3], Kantorova Vladimira[4], Kirill Andreev[5]

Extended Abstract for the First Human Fertility Database Symposium
(Rostock, Germany, 3-4 November 2011).

Summary:

This paper focuses on the transformation of age-specific fertility rates from five-year age groups into single age. We review and apply different statistical approaches and mathematical models to graduate fertility rates from grouped data into age-specific rates.

We focus on approaches with (a) the most potential to graduate a wide range of fertility patterns (from pre- to post-transition patterns) and (b) the most minimalist data requirements (i.e., only one year of data available). We compare the performances of 10 methods using a sample of HFD countries for which we can compare empirical age-specific fertility rates against graduated ones derived from abridged fertility rates we computed based on annual births and exposure by 5-year age groups of the mother.

Background:

Detailed fertility and population data by single calendar year and single age as used by the HFD are available at best only for countries with good vital registration systems and with regular population censuses providing detailed and regular demographic data to compute accurately age-specific fertility rates. Such detailed data, especially by single year of age, are often lacking for older time periods, and in some instances are less accurate for older birth cohorts due to age heaping or even age exaggeration. More importantly, since the 1950s only about half of the countries in the world are able to provide on a regular basis good and reliable vital registration data. For the rest of the countries, only partial and often deficient information exists for some years, and fertility data often depend on sample survey information rather than vital registration and census data.

Since the most commonly available fertility data are often only published in abridged form, and sometimes only for broader age groups, it is often useful to graduate fertility data from five-year age groups into single-year for a range of analytical and modeling purposes where annual data by single age are more convenient to work with.

The demographic literature offers various graduation methods to transform grouped data into single age, especially in respect to fertility data, but so far no comprehensive evaluation of the most popular methods is available benchmarking them against a wide range of populations and time periods. This paper aims to fill this gap by using a sample of HFD countries to evaluate the performance of 10 transformation methods for age-specific fertility rates.

1. Columbia University, Dept. of Statistics, 1255 Amsterdam Avenue, New York, NY 10027, USA. E-mail: liuyebest@gmail.com
2. United Nations, DESA, Population Division. Estimates and Projection Section. Room DC2-1914 – 2 UN Plaza. New York, NY 10017, USA. E-mail: gerland@un.org
3. United Nations, DESA, Population Division. Estimates and Projection Section. Room DC2-1908 – 2 UN Plaza. New York, NY 10017, USA. E-mail: spoorenberg@un.org
4. United Nations, DESA, Population Division. Fertility Section. Room DC2-1904 – 2 UN Plaza. New York, NY 10017, USA. E-mail: kantorova@un.org
5. United Nations, DESA, Population Division. Estimates and Projection Section Room DC2-1912 – 2 UN Plaza. New York, NY 10017, USA. E-mail: andreev@un.org

Data:

This paper draws extensively on the Human Fertility Database and focuses on seven OECD countries providing a wide range of fertility patterns from pre- to post-transition patterns.

The sample of countries and time periods includes: Sweden (1891-2001), Canada (1921-2007), Germany (1956-2009), Austria (1951-2008), United States of America (1933-2006), Czech Republic (1950-2009) and France (1946-2009).

Annual data by single-year of age and calendar year for "all birth orders combined" were downloaded from the HFD web site (http://www.humanfertility.org/) in July 2011. The series used as inputs were birth counts, female population exposure and age-specific fertility rates.

We aggregated births and female exposure by five-year age groups (from 15-19 to 50-54) and computed corresponding abridged fertility rates for each year and country. We assumed that the five-year fertility rates are centered on the mid of the corresponding age groups.

Analytical strategy:

The methods reviewed here have the most basic or minimal data requirement and can be applied with cross-sectional data available only for one period or year. The choice of methods evaluated was essentially driven by the type of data typically available for most countries (i.e., only abridged rates).

We excluded from this analysis two types of methods due to their data requirements:

- <u>Age-specific methods</u> (all smoothing methods relying on detailed single-age data) ;
- <u>Time series approaches</u> (all methods relying on time series availability, especially Age-Period-Cohort methods (e.g., Lee-Carter approaches and functional analysis (Hyndman and Booth ; Hyndman and Shahid Ullah 2007), state-space logistic models (Rueda-Sabater and Alvarez-Esteban 2008; Rueda and Rodríguez), Support Vector Machines (Kostaki et al. 2009))

We focused on three types of general methods (non-parametric, osculatory interpolations and parametric) to convert ASFR by five-year age groups into single age, and use HFD data as a benchmark with the goal of reproducing the single age rates using only the abridged ones. Other approaches are not as promising as general methods able to deal with a wide range of patterns for both developed and developing countries from pre- to post-transition patterns. To summarize the performance of each model, we computed the sum of square errors (SSE) by age and by year for each country and model, and use the mean SSE as goodness of fit.

1. Standard non-parametric statistical model: Monotone Piecewise Cubic Interpolation

A wide variety of non-parametric statistical approaches (e.g., kernels and splines) exist (de Beer 2011; McNeil, Trussell and Turner 1977; Schmertmann 2003), but one of the most convenient approaches is the Monotone Piecewise Cubic Interpolation (Fritsch and Carlson 1980; Smith, Hyndman and Wood 2004). We tested this approach on the original data and after some transformations:

- The first application uses the original data regrouped into five-year age groups. Due to the five-year aggregation, this approach requires some assumptions at the tails when data are not available below age 17 or above age 52. In these cases, we extrapolated using the decreasing ratio of the last two elements, e.g. for the points smaller than the interpolation range, we let x(i)=x(i+1)*x(i+1)/x(i+2) ; for the points larger than the interpolation range, we let x(i)=x(i-1)*x(i-1)/x(i-2).
- The second application uses a Logit transformation of the original five-year fertility rates (to constraint the interpolated rates to be bounded between [0,1]), and an anti-logit transformation back after the interpolation.

- The third application relies on the cumulative transformation of the data to do the interpolation. The use of cumulated distributions for model fitting, smoothing and interpolation is popular in demography because it helps to smooth noisy data and makes it easier to interpolated data with uneven spacing or open age groups.

2. Osculatory Interpolations

A second approach we considered for its robustness and computational performance relies on osculatory interpolations. In this section, we have tested four formulas (Beer, Beer Modified, Karup and Sprague) to subdivide age groups into fifth (Swanson and Siegel 2004). One of the drawbacks of high degree polynomial interpolations is that negative points may occur at the first five-group points and the last three five-group points. Consequently we have made the following adjustments for the negative points for each formula:

Assuming j is largest index of points at the first group which is negative, we adjusted it in the way that x(j)=x(j+1)*x(j+1)/x(j+2) from the j index to the first point, and summed up all the points in the group. Then, for j from 1 to 5, we assumed that x(j) equals x(j) times the given value of this group, and divided by the sum obtained from the previous step. The same adjustments are done for the last groups, while the smallest index needs to be detected, and adjustments should be made from the group involved in the final group.

3. Parametric Models

Several parametric functions have commonly been used in demography, especially to model fertility age patterns (e.g., Beta, Gamma, Pearson curves, Hadwiger (Chandola, Coleman and Hiorns 1999; Hoem et al. 1981), etc.) and Coale-Trussel (1974) and Xie (1992) models. Typically these parametric models fit well specific sets of fertility data only for some countries and time periods for which these functional shapes apply. The aim of these models is to reduce to a few interpretable parameters the empirical distributions either to answer analytical questions or to project trends in the parameters in order to predict future fertility age patterns (Rogers 1986; Thompson et al. 1989).

The parsimony of these models and the functional shape they impose on the data are useful to smooth or even adjust poor-quality data by imposing a structure considered more appropriate than empirically observed. These features are not only strengths, but also limitations preventing the general use of any of these parametric functions to all countries and time periods. One of the major problems with this approach is the inability for most of these models to deal with the change of shape that occurs during the first (or even the second) fertility transition experienced by most countries. The dependence on any fixed functional form imposes too many constraints and “rigidity” to any given shape.

In this paper we focus on three general parametric models that have been identified as more flexible for both historical and contemporary age fertility patterns. We fit them using least squares estimates:

- The first model we use here is the Gompertz model (Goldstein 2010), with the form of:

$$f(x)=K*a*exp[-a/b*exp(-b*x)-b*x]$$

- The second model is a Modified Gompertz model, which uses the same function as the Gompertz model when the age x is less than 45. When the age is larger or equal to 45, the modified model is the product of the functions f(x) and g(x), where g(x) takes the value of 1 at 45 and takes the value of 0 at 54, which is a linear function between 45 and 54.
- The third model used is the Two Peak model (Kostaki and Peristera 2007), simply because it can represent the pattern of two bulges and a flat peak. Its formula is:

$$f(x)=c_{1*}exp\{-[(x-u_1)/\sigma_1]^2\}+ c_{2*}exp\{-[(x-u_2)/\sigma_2]^2\}$$

Preliminary results:

With respect to the modeling by age, the best performing methods providing the best goodness of fit across the most ages (i.e., smallest errors from empirical single ASFR) are the Beer and Sprague interpolation methods. The worst approaches are the Monotone Piecewise Cubic Interpolation with cumulative transformation, where greater errors occur at the mid age of the first two age groups, and the Gompertz model (Table 1 and Figure 1 for Sweden).

For the Monotone Piecewise Cubic Interpolation, the original data without transformation provide the smallest error in most instances. In absolute terms, errors become very small after age 45. The application of this approach on transformed rates (either through Logit transformation or using a cumulated distribution) leads to larger errors.

For all osculatory Interpolations, there is no great difference in errors at each age, and errors are smaller after age 45.

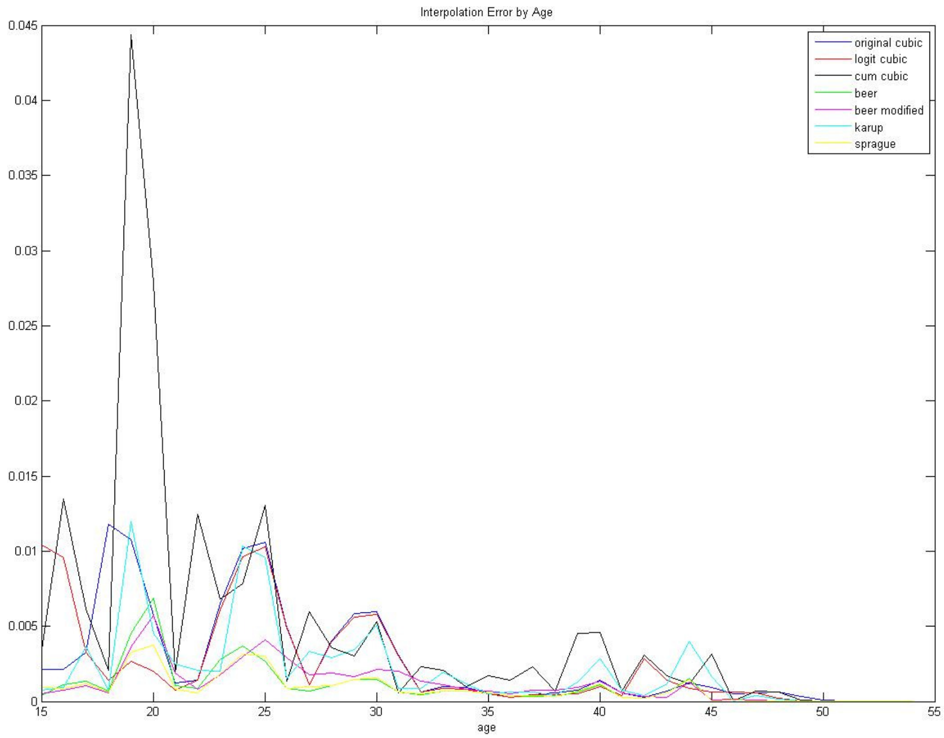
For the Parametric Model, the Modified Gompertz Model does a better job than Gompertz Model after age 45, but both of them are not so efficient to fit older ages compared to other methods. And for most cases, there are larger errors around the age of 20 and 25 in the Two Peak Model.

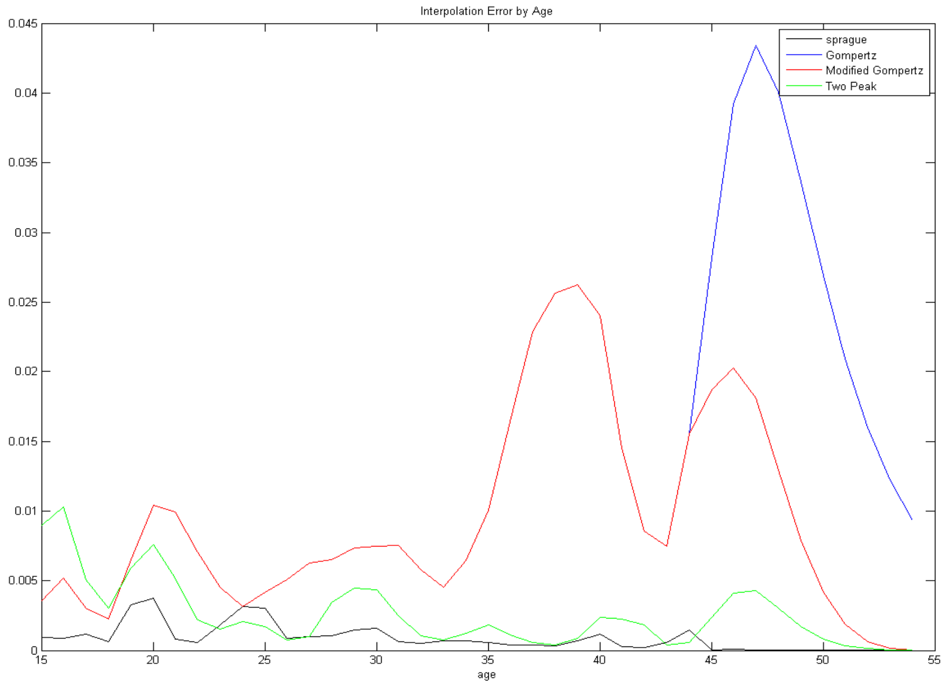
Table 1. Mean of the Sum of Square Errors by *Age (from 15 to 54)*:

	Sweden	**Canada**	**Germany**	**Austria**	**USA**	**Czech**	**France**	**Overall***
Original data with Monotone Cubic interpolation	0.0026	0.0034	0.0011	0.0013	0.0015	0.0028	0.0042	0.0024
Logit transformed data with Monotone Cubic	0.0024	0.0034	0.0014	0.0020	0.0031	0.0046	0.0029	0.0028
Cumulated data with Monotone Cubic	0.0048	0.0063	0.0026	0.0039	0.0071	0.0121	0.0039	0.0058
Beer interpolation method	0.0010	0.0014	0.0004	0.0006	0.0007	0.0020	0.0014	0.0011
Beer Modified interpolation method	0.0012	0.0016	0.0005	0.0007	0.0011	0.0031	0.0016	0.0014
Karup interpolation method	0.0021	0.0027	0.0011	0.0016	0.0027	0.0056	0.0022	0.0026
Sprague interpolation method	0.0009	0.0015	0.0004	0.0005	0.0007	0.0028	0.0015	0.0012
Gompertz function	0.0140	0.0086	0.0022	0.0026	0.0049	0.0040	0.0032	0.0056
Modified Gompertz function	0.0093	0.0072	0.0021	0.0024	0.0046	0.0039	0.0030	0.0047
Two Peak function	0.0026	0.0024	0.0013	0.0023	0.0033	0.0102	0.0039	0.0037

(*) unweighted average of 7 countries

Figure 1. Sweden: Errors by Age for selected graduation models





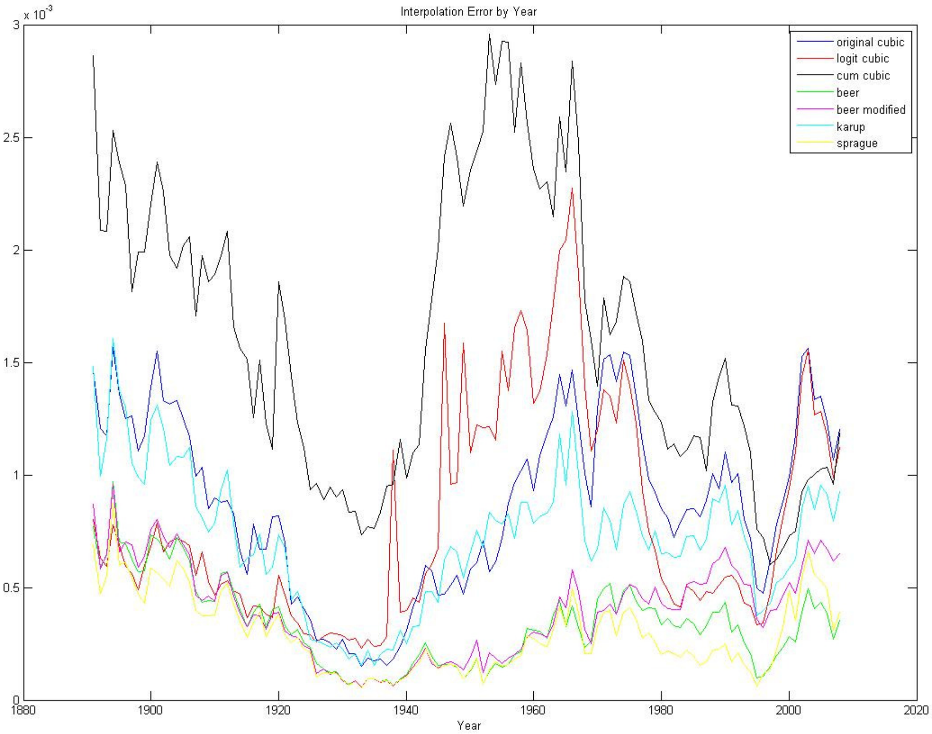
With respect to the modeling by time, we reach the same conclusion. The best methods are the Beer Method and Sprague Method which provide the best goodness of fit over the most years. The worst approaches are the Monotone Piecewise Cubic Interpolation with cumulative transformation and Gompertz model (Table 2 and Figure 2 for Sweden).

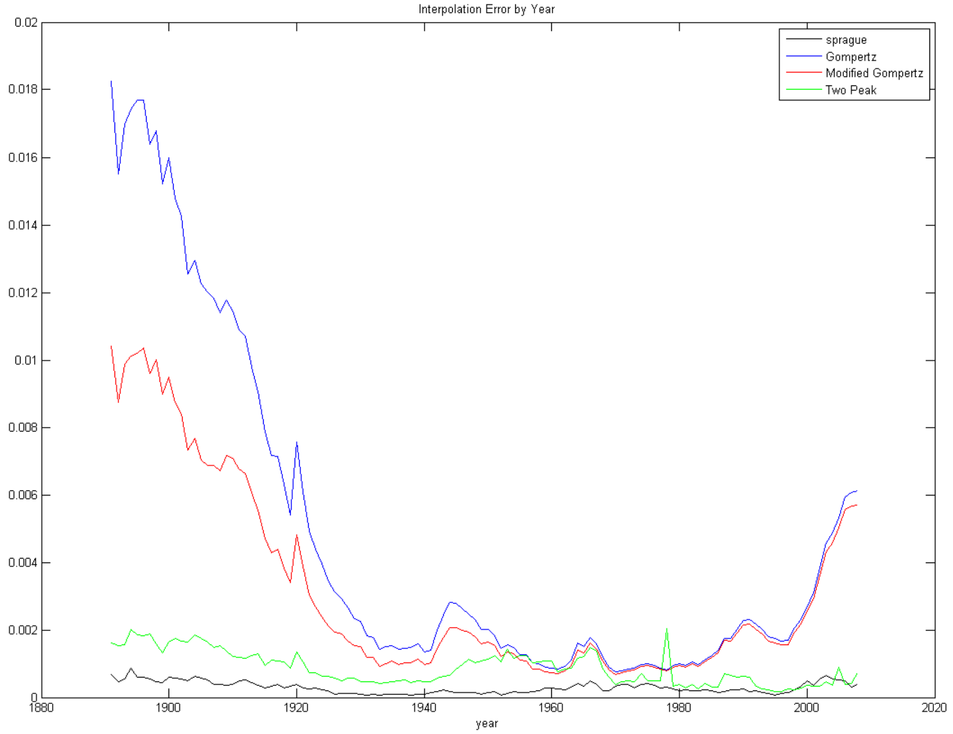
For most methods, relative large errors occur around the 1950's and 1960's. Particularly, the errors of the Two Peak Model are unstable for some countries. In the Parametric Models, Two Peak Model has a very good fit for the recent twenty years, while at the same time, Gompertz and Modified Gompertz Model are so not convincing.

Table 2. Mean of Sum Square Error by *Year*:

	Sweden	**Canada**	**Germany**	**Austria**	**USA**	**Czech**	**France**	**Overall***
Original data with Monotone Cubic interpolation	0.0009	0.0016	0.0008	0.0009	0.0008	0.0019	0.0026	0.0013
Logit transformed data with Monotone Cubic	0.0008	0.0016	0.0010	0.0013	0.0017	0.0030	0.0018	0.0016
Cumulated data with Monotone Cubic	0.0016	0.0029	0.0019	0.0027	0.0038	0.0081	0.0024	0.0034
Beer interpolation method	0.0003	0.0007	0.0003	0.0004	0.0004	0.0013	0.0009	0.0006
Beer Modified interpolation method	0.0004	0.0007	0.0004	0.0005	0.0006	0.0021	0.0010	0.0008
Karup interpolation method	0.0007	0.0013	0.0008	0.0011	0.0015	0.0037	0.0014	0.0015
Sprague interpolation method	0.0003	0.0007	0.0003	0.0004	0.0004	0.0018	0.0009	0.0007
Gompertz function	0.0047	0.0040	0.0017	0.0018	0.0027	0.0027	0.0020	0.0028
Modified Gompertz function	0.0032	0.0033	0.0016	0.0017	0.0025	0.0026	0.0019	0.0024
Two Peak function	0.0009	0.0011	0.0010	0.0016	0.0018	0.0068	0.0024	0.0022

Figure 2. Sweden: Errors by Year for selected graduation models





Conclusion

Using the HFD data for selected developed countries covering a range of historical and contemporary fertility patterns by age, we can conclude that the Beer and Sprague formulas are the best methods to transform fertility rates from standard five-year age groups into single age rates.

Estimation of non-parametric models with limited age groups can be at time problematic due to the large number of parameters to estimate with a limited number of empirical observations. In this context, parametric models are easier to estimate since they are more parsimonious, but they often depend on an initial guess that can sometime be problematic to find the optimal parameters. The more parameters involved, the more errors may occur during the estimation ; therefore, the more parsimonious approaches should be favored with limited data. Finally, nearly all the methods examined in this paper are inefficient to capture the peak of the fertility age pattern (occurring often in the middle of a five-year age group – see Annex plots showing Swedish fertility age pattern graduation for selected years).

References:

Chandola, T., D.A. Coleman, and R.W. Hiorns. 1999. "Recent European Fertility Patterns: Fitting Curves to 'Distorted' Distributions." *Population Studies* 53(3):317-329.

Coale, A.J.and T.J. Trussell. 1974. "Model Fertility Schedules: Variations in The Age Structure of Childbearing in Human Populations." *Population Index* 40(2):185-258.

de Beer, J. 2011. "A new relational method for smoothing and projecting age-specific fertility rates: TOPALS." *Demographic Research* 24(18):409-454.

Fritsch, F.N.and R.E. Carlson. 1980. "Monotone Piecewise Cubic Interpolation." *SIAM Journal on Numerical Analysis* 17(2):238-246.

Goldstein, J.R. 2010. "A behavioral Gompertz model for cohort fertility schedules in low and moderate fertility populations." in *MPIDR Working Paper.* Rostock: Max Planck Institute for Demographic Research.

Hoem, J.M., D. Madsen, J.L. Nielsen, E.-M. Ohlsen, H.O. Hansen, and B. Rennermalm. 1981. "Experiments in Modelling Recent Danish Fertility Curves." *Demography* 18(2):231-244.

Hyndman, R.J.and H. Booth. "Stochastic population forecasts using functional data models for mortality, fertility and migration." *International Journal of Forecasting* 24(3):323-342.

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

Kostaki, A., J. Moguerza, A. Olivares, and S. Psarakis. 2009. "Graduating the age-specific fertility pattern using Support Vector Machines." *Demographic Research* 20(25):599-622.

Kostaki, A.and P. Peristera. 2007. "Modeling fertility in modern populations." *Demographic Research* 16(6):141-194.

McNeil, D.R., T.J. Trussell, and J.C. Turner. 1977. "Spline Interpolation of Demographic Data." *Demography* 14(2):245-252.

Rogers, A. 1986. "Parameterized Multistate Population Dynamics and Projections." *Journal of the American Statistical Association* 81(393):48-61.

Rueda-Sabater, C.and P.C. Alvarez-Esteban. 2008. "The analysis of age-specific fertility patterns via logistic models." *Journal of Applied Statistics* 35(9):1053-1070.

Rueda, C.and P. Rodríguez. "State space models for estimating and forecasting fertility." *International Journal of Forecasting* 26(4):712-724.

Schmertmann, C. 2003. "A system of model fertility schedules with graphically intuitive parameters." *Demographic Research* 9(5):81-110.

Smith, L., R. Hyndman, and S. Wood. 2004. "Spline interpolation for demographic variables: The monotonicity problem." *Journal of Population Research* 21(1):95-98.

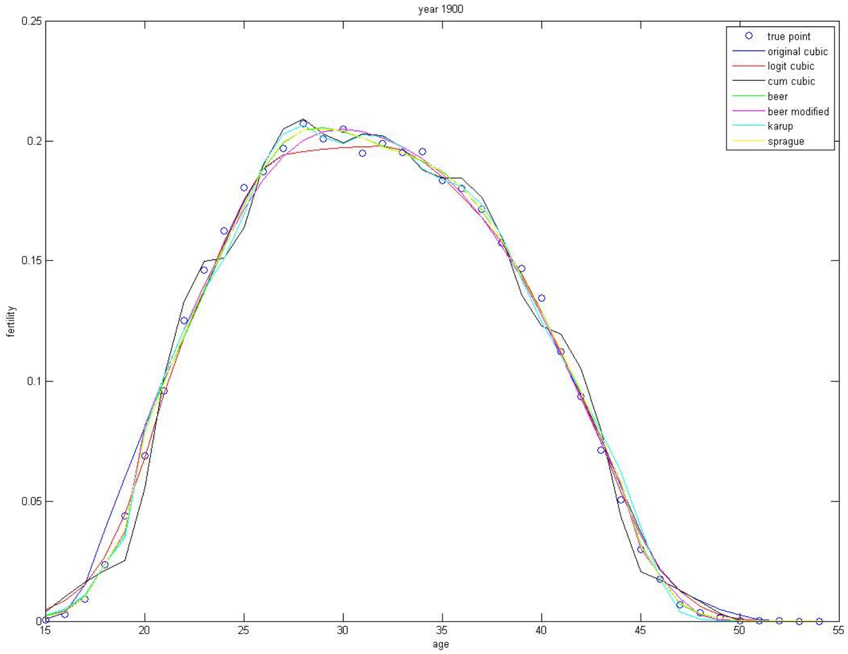
Swanson, D.and J.S. Siegel. 2004. "The methods and material of demography." Pp. 640. San Diego, CA: Academic Press.

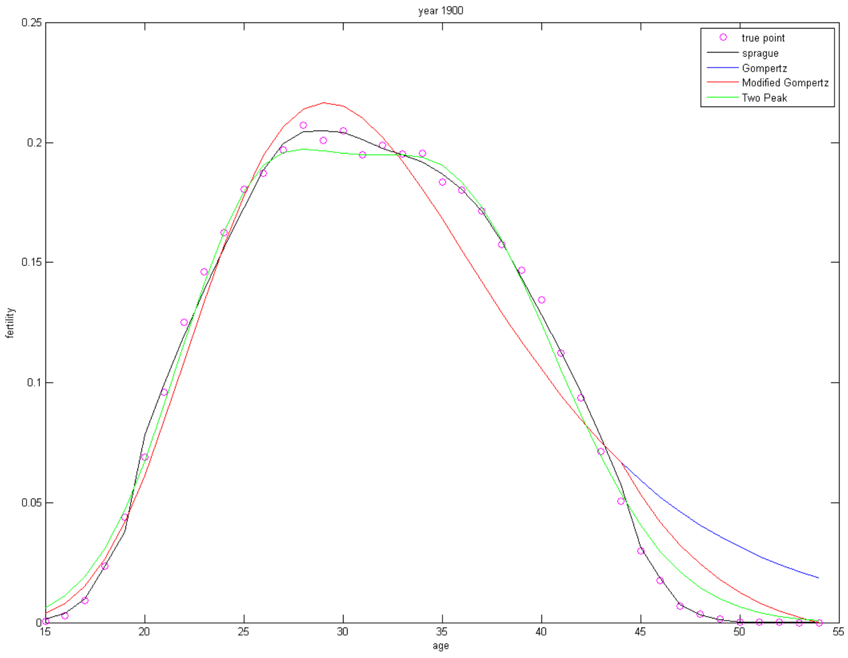
Thompson, P.A., W.R. Bell, J.F. Long, and R.B. Miller. 1989. "Multivariate Time Series Projections of Parameterized Age-Specific Fertility Rates." *Journal of the American Statistical Association* 84(407):689-699.

Xie, Y.and P. Ellen Efron. 1992. "Age Patterns of Marital Fertility: Revising the Coale-Trussell Method." *Journal of the American Statistical Association* 87(420):977-984.

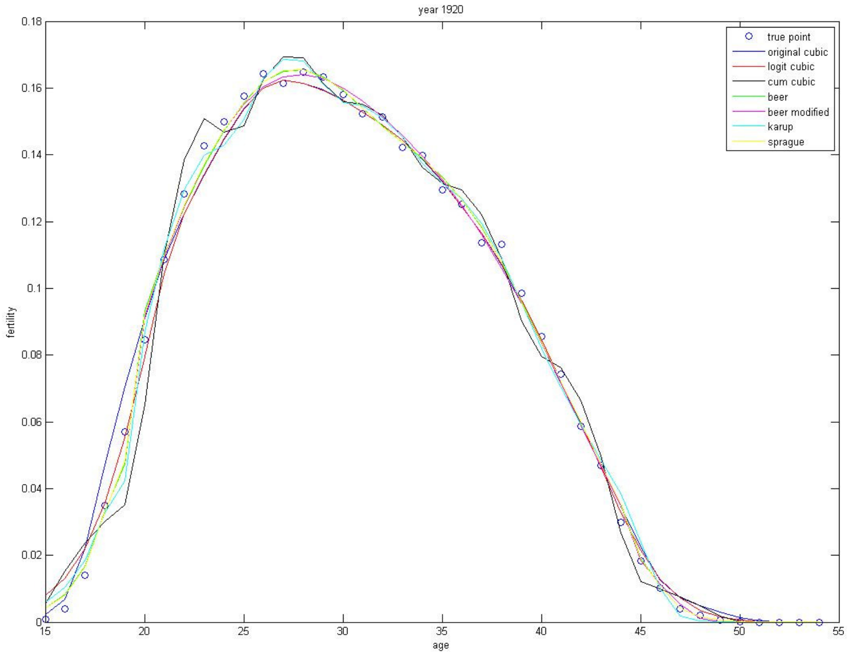
Annex: Swedish fertility age pattern for selected years

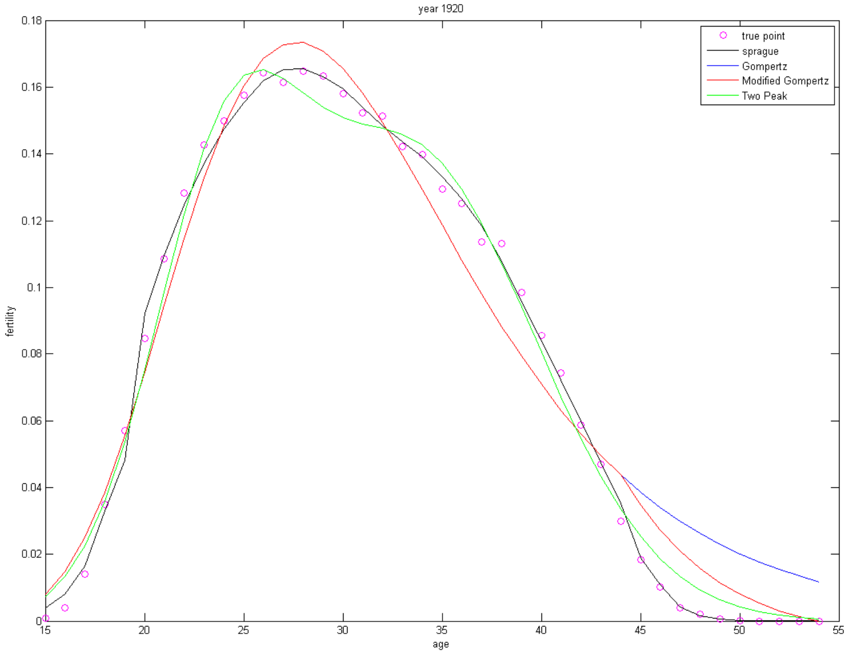
Sweden 1900 fertility age pattern graduation



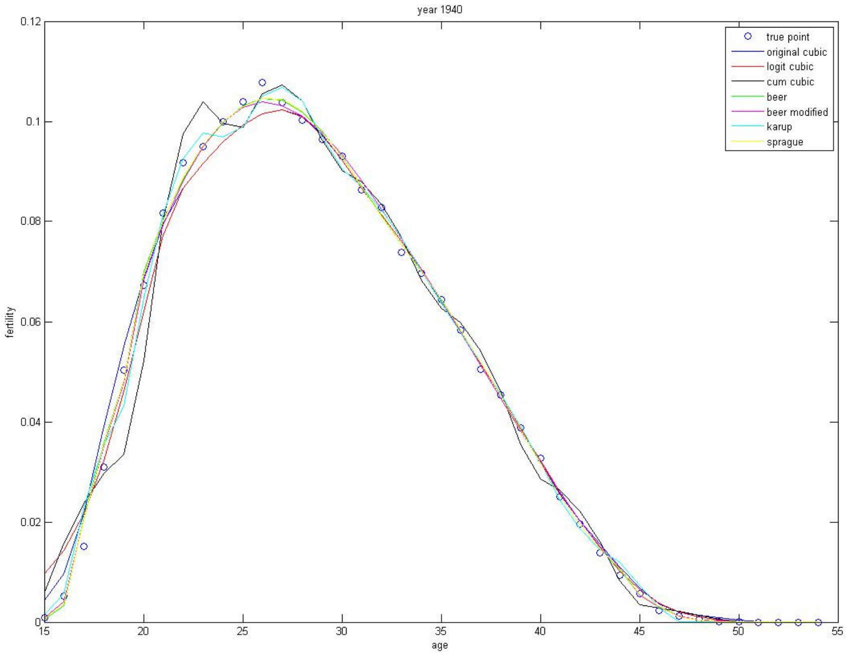


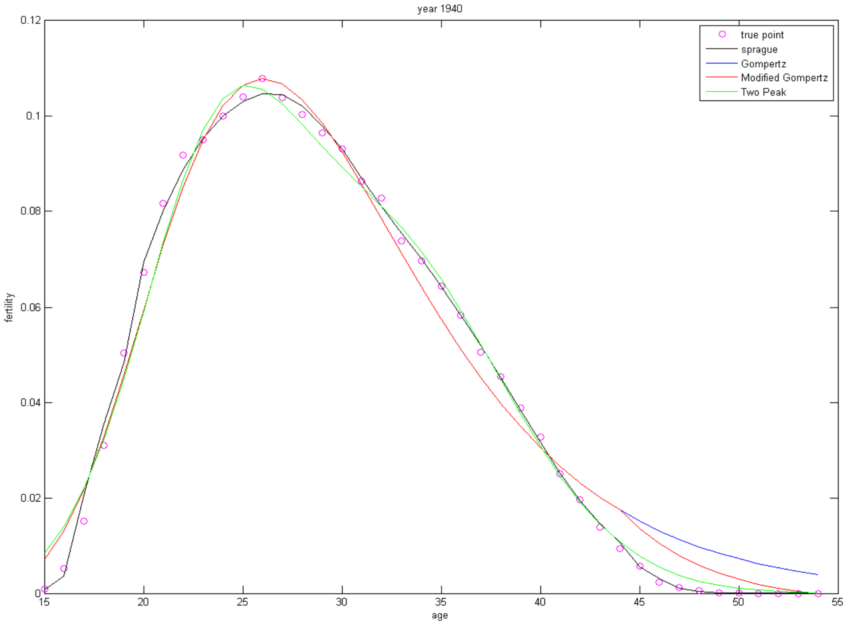
Sweden 1920 fertility age pattern graduation



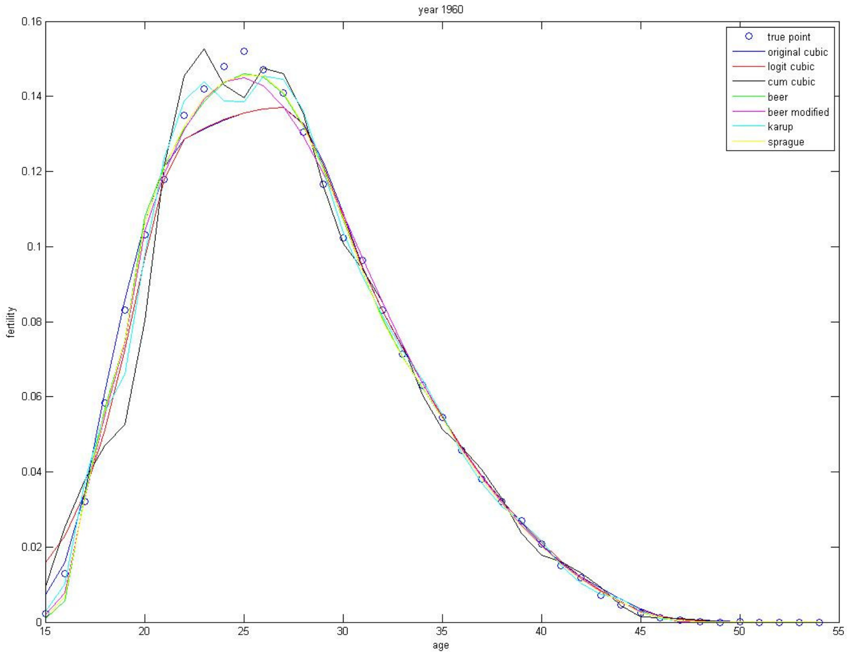


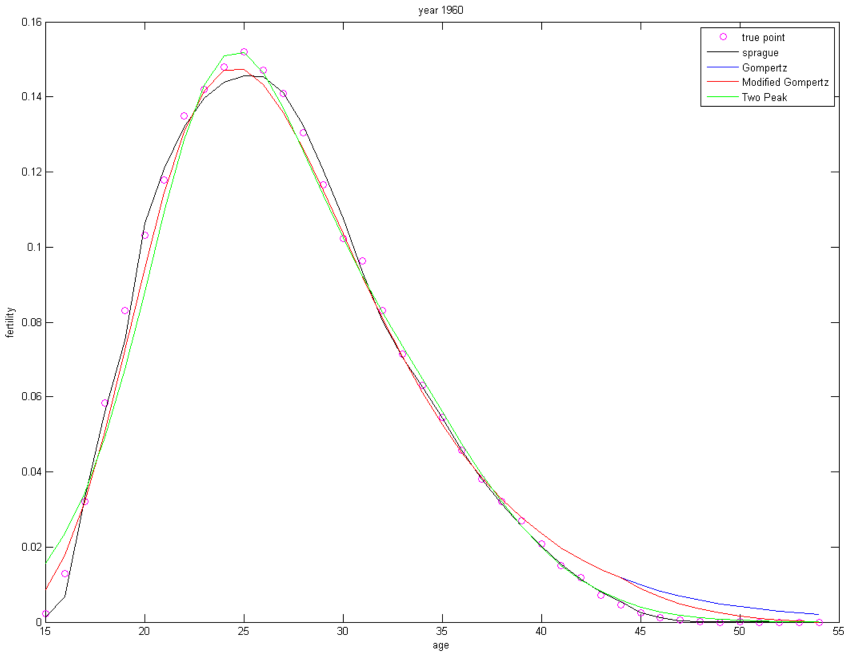
Sweden 1940 fertility age pattern graduation



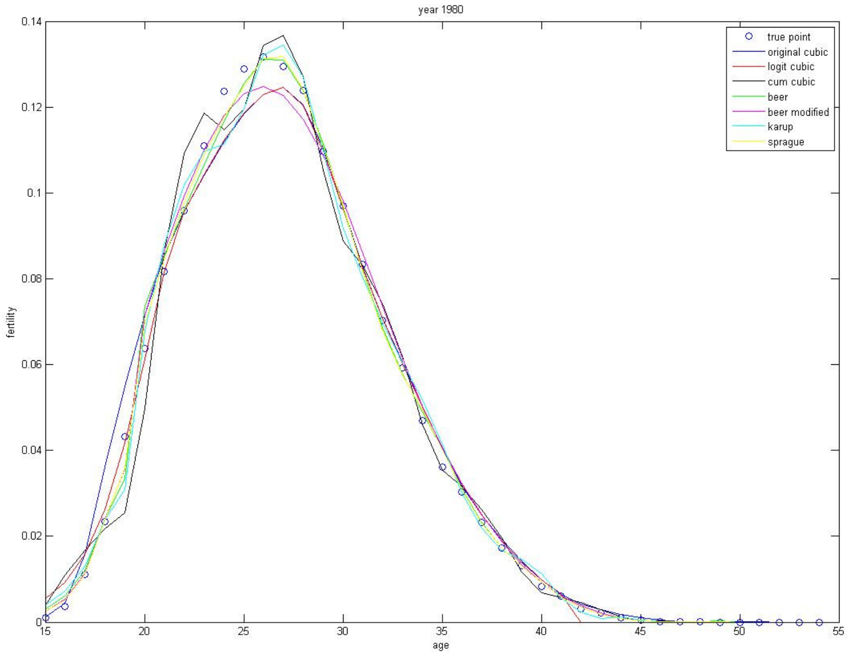


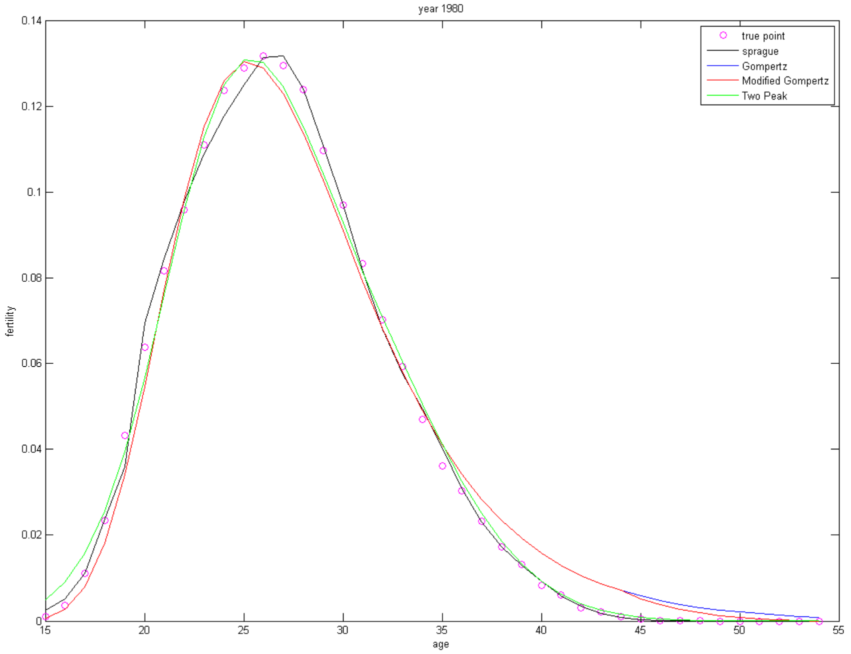
Sweden 1960 fertility age pattern graduation



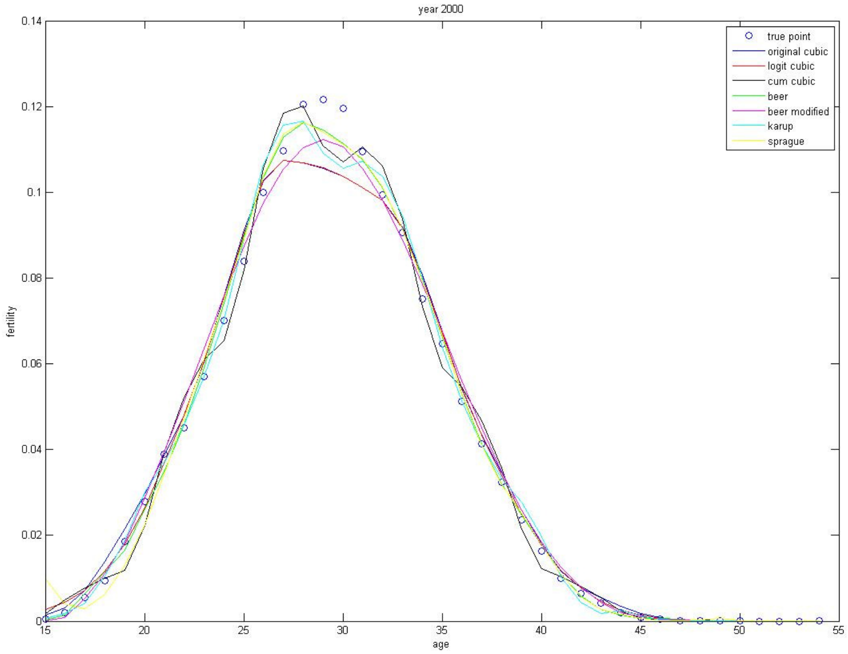


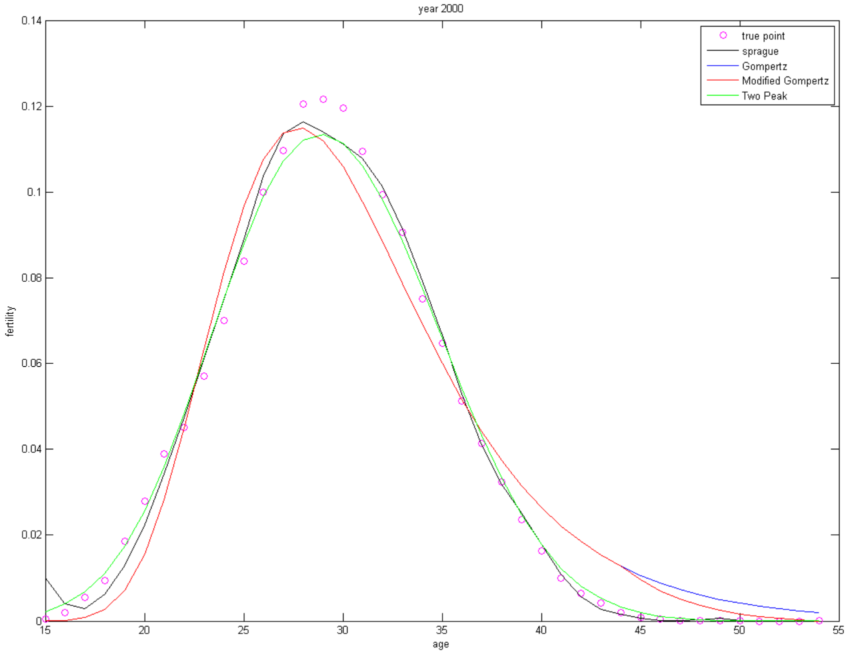
Sweden 1980 fertility age pattern graduation





Sweden 2000 fertility age pattern graduation


year 2000
true point
original cubic
logit cubic
cum cubic
beer
beer modified
karup
sprague
fertility
age


year 2000
true point
sprague
Gompertz
Modified Gompertz
Two Peak
fertility
age