Forecasting fertility by age and birth order using time series from the Human Fertility Database

Han Lin Shang*
Alessandra Carioli †
Guy J. Abel‡

December 15, 2015

1 Abstract

This article presents a comparison of two approaches in forecasting total fertility, age specific fertility, and birth order age specific fertility rates. We employ an auto-arima and a functional time series robust forecasting model to project fertility for 23 countries using the Human Fertility Database time series. The comparison of the two models aim at demonstrating the advantages of forecasting both age specific rates as opposed to the common practice of simply forecasting total rates. We use a functional principal components analysis to project smoothed age-specific fertility rates in the short run (15 years). We compare the empirical accuracy of the approach by Hyndman and Ullah (2007) to auto-ARIMA forecasts and we validate our results through the one-step-ahead to 20-step-ahead forecast error measures (MFE and MAFE).

2 Data and Methods

Fertility forecasting has recently experienced a renewed interest due to plummeting childbearing trends in recent decades, making policy makers wonder how much longer will low fertility persist and if such a downward trend will ever reverse. The new accessibility of long fertility time series together with their detaledness made possible by the Human Fertility Database allow to try new approaches in forecasting. The recent study by Hyndman and Ullah (2007) provides a robust forecasting method, which allows to use age specific fertility rates to forecast total fertility and age specific fertility rates at the same time.

*Research School of Finance, Actuarial Studies and Applied Statistics, The Australian National University
†Netherlands Interdisciplinary Demographic Institute and University of Groningen
‡School of Sociology and Political Science, Shanghai University

This paper stems from the idea of comparing fertility forecasts using age specific fertility rates and total fertility, all broke down by parity from first to fifth or higher birth order, to assess the robustness of Hyndman and Ullah (2007) projection model and contrast the results with auto-ARIMA forecasts.

2.1 Data

The data sets used in this study come from the Human Fertility Database (2015). Fertility indicators consist of age-specific fertility rates, for total and parity specific fertility, obtained from the ratio between number of live births by mothers age (occurrences) and number of women by age at childbirth among the female resident population (exposures) for a specific calendar year. We selected 23 countries for the projection of ASFR, while the number of countries used for parity specific fertility projections had to be reduced to 11 due to shorter data series. We selected countries with data series to be at least 50 years long to obtain consistent sample estimators (Box 2008, ch. 1).
The input data can be consulted and visualized at `https://gjabel.shinyapps.io/fertdagg`.

2.2 Method

Hyndman and Booth (2007) projection method combines non-parametric smoothing, functional data analysis and principal components analysis. Principal components methods are not new in fertility forecasts: the advantage of Hyndman and Ullah (2007) and Hyndman and Booth (2008) is to provide an interpretation of which factors have played a substantial role in shaping fertility in the past and which factors are going to be relevant in the future.
The technique chosen for this study follows the model introduced by Hyndman and Ullah (2007) and improved in Hyndman and Booth (2008) using the demography package in R (Hyndman and Booth 2013). Hyndman and Booth(2008) and Hyndman and Ullah (2007) suggest to smooth the age-specific fertility schedules in order to erase random fluctuations of fertility rates over time, which would result in erratic age patterns if forecasted without being first smoothed. Smoothing age-specific fertility schedules allows a reduction of the inherent randomness in the observed data (Ramsay and Silverman, 2005).
Considering $y_t^*(x)$ the observed fertility rates to be modeled, Hyndman and Booth (2008) first proposes a Box-Cox transformation (Box and Cox, 1964) of $y_t^*(x)$, to allow for variation that increases with the value of $y_t^*(x)$ so that the variability in rates is greater when the rates are higher:

$$y_t(x) = \tfrac{1}{\lambda}([y_t^*(x)]^\lambda - 1) \text{ if } 0 < \lambda < 1 \quad (1)$$

where λ determines the strength of the transformation and is set at 0.4 as it gives relatively small out of sample errors Hyndman and Booth (2008), which differs from the first approach used in Hyndman and Ullah (2007) that used a log-linear transformation of the fertility rates, to allow for variation that increases with the value of $y_t^*(x)$.

The following model is developed for the transformed quantity of $y_t(x)$:

$$y_t(x) = s_t(x) + \sigma_t(x)\epsilon_{t,x} \text{ (2)}$$

$$s_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x) \text{ (3)}$$

where $s_t(x)$ is the smoothed age-specific fertility schedule of age x=[15,49], $\epsilon_{t,x}$ is a iid standard normal random variable and $\sigma_t(x)$ allows the amount of noise to change with x.
Within the smoothed function $s_t(x)$, $\mu(x)$ average of the age-specific fertility schedules, a robust estimate of the mean of $s_t(x) over estimated years$t =[$t_0, T$]. The $\phi_k(x)$ terms represent the set of k orthogonal basis functions, which are obtained through principal components decomposition.
Hyndman and Ullah (2007) introduce a two-step algorithm to obtain robust principal component decomposition to derive $\phi_k(x)$ using weighted principal components (Ramsay and Silverman, 2005 chapter 8) and a projection pursuit algorithm RAPCA from Hubert et al. (2002). This two-step algorithm has the advantage of being faster for highly-dymensional data and numerically more stable and when combined with weighted approach to principal components analysis is also more efficient as it ensures a robust method to estimate the basis functions. The $\beta_{t,k}$ terms are independent sets of coefficients for each kth component in years $t = [t_0, T]$. The last element of the equation, $e_t(x) \sim N(0, \vartheta(x))$, is the serially uncorrelated error model.
This model is a generalization of the Lee and Carter (1992) model for mortality forecasting. There are several differences between the Lee and Carter(1992) and Hyndman and Ullah(2007) and Hyndman and Booth (2008) that help improve the fertility forecast. Hyndman and Ullah(2007) and Hyndman and Booth(2008) assume a smooth fertility schedule, thus eliminating fluctuations in the fertility rates. Second, in Lee and Carter(1992) $\mu(x)$ is defined as the average of $y_t(x)$ and hence only one component is used, K=1. Thus, $\beta_{t,k}$ and $\phi_k(x)$ are computed from one component: $[y_t(x) - \mu\hat{(}x)]$.
The two-step principal component method yields the decomposition:

$$\hat{f}_t(x) = \hat{\mu}_t(x) + \sum_{k=1}^{K} \hat{\beta}_{t,k}\phi_k(x) + \hat{e}_t(x) \text{ (4)}$$

By definition the basis functions $\phi_k(x)$ are uncorrelated for k?l, thus as Hyndman and Ullah(2007) showed, robust forecast of the fertility schedule can be based on univariate time series models fitted to each of the coefficients $\beta_{t,k}$, k=1,?,K for t=1,?,n+h where h is the forecast horizon. It follows that the estimated variance of the error term is used to compute the prediction intervals for the forecast.
The third step is to fit univariate time series models to each of the coefficients $\hat{\beta}_{t,k}$, k = 1, . . . , K . For each set of the coefficients , k=1,?,K univariate robust ARIMA models are fitted following a RAPCA projection pursuit algorithm based on Chen and Liu (1993), which includes a method that deals with outliers so that unusual observations do not contaminate the forecast, as specified in Hyndman and Ullah (2007). If we combine (2) and (3) we obtain a forecast equation for future transferred fertility

schedules as:

$$y_t(x) = mu(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x_i) + e_i(x_i) + \sigma_t(x_i)\epsilon_{t,i} \text{ (5)}$$

Conditioning on the observed data and on the basis functions we obtain the h-step ahead forecasts of $y_{n+h}(x)$:

$$\hat{y}_{n,h}(x) = E[y_{(n+h)}(x) \mid \Im, \Phi] = \hat{\mu}(x) + \sum_{k=1}^{K} \tilde{\beta}_{n,k,h}\hat{\phi}_k(x) \text{ (6)}$$

where $\Im$ denotes observed data $y_t(x_i)$. Hyndman and Booth (2008) suggests selecting K to minimize the mean integrated squared forecast error, as a small K would decrease forecast accuracy. Hyndman and Booth (2008) find that the method is insensitive of the choice of K, provided that K is large enough.

2.3 Validation

Validation is carried out truncating progressively the time series for each country in 1990 and projecting fertility rates for a 20 years horizon using both Hyndman and Ullah (2007) and ARIMA models.
In order to measure forecast accuracy and bias, we employ mean forecast error, MFE, and mean absolute forecast error, MAFE. While the MFE is a measure of bias and can be positive or negative depending on whether the projection overestimates or underestimated the actual data, MAFE measures forecast precision regardless of its sign and is the average of absolute errors across years in the forecasting horizon.
Given f_i the predicted value and y_i the observed value, MFE and MAFE are defined as:

$$MFE = \tfrac{1}{h}\sum_{i=1}^{h}(f_i - y_i) = \frac{1}{h}\sum_{i=1}^{h}(e_i) \text{ (7)}$$

$$MAFE = \tfrac{1}{h} \mid \sum_{i=1}^{h} \mid f_i - y_i \mid = \frac{1}{h}\sum_{i=1}^{h} \mid e_i \mid \text{ (8)}$$

where $\mid f_i - y_i \mid$ is the average of absolute errors, $\mid e_i \mid$.

2.4 Preliminary results

The forecast have been implemented using R, through the demography package (Hyndman 2014) with the coherent.fdm function, and the forecast package, with the auto.aria function (Hyndman and Kandahar 2008).
We proceeded to forecast central rates for:

- tfr based on auto-arima and tfr data;
- asfr and tfr by birth order by auto-arima;
- asfr and tfr based and asfr data on Hyndman and Ullah (2007) functional model;
- asfr and tfr by birth order by Hyndman and Ullah (2007) functional model;

Preliminary results of the validation for central rates show the better fit of Hyndman and Ullah (2007) model.
Nevertheless, uncertainty needs to be further investigated.

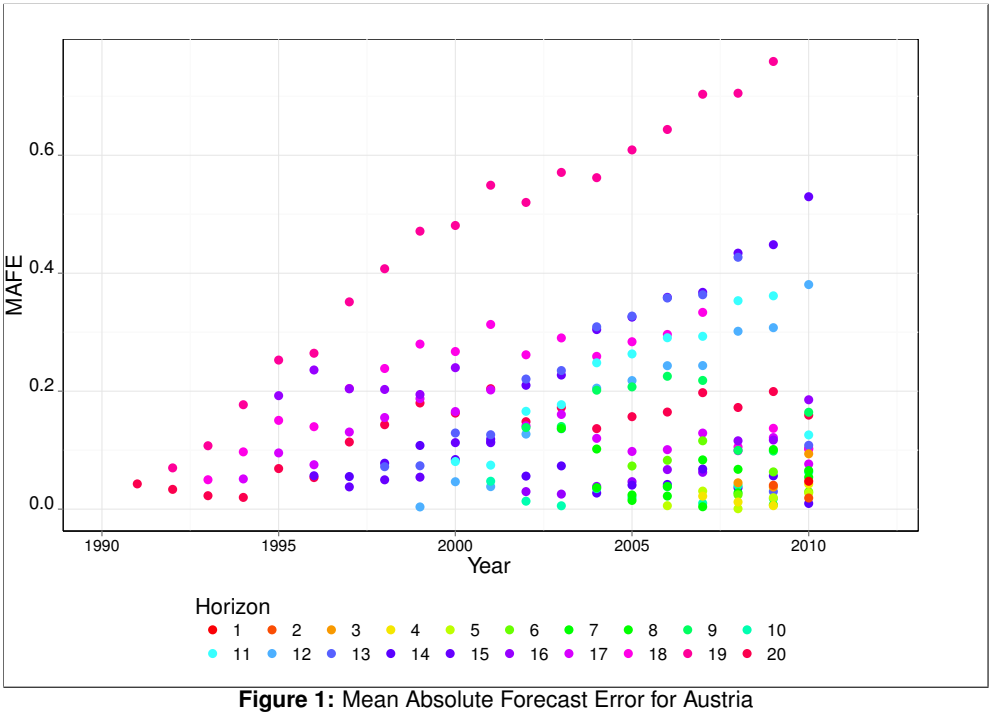

Figure 1: Mean Absolute Forecast Error for Austria

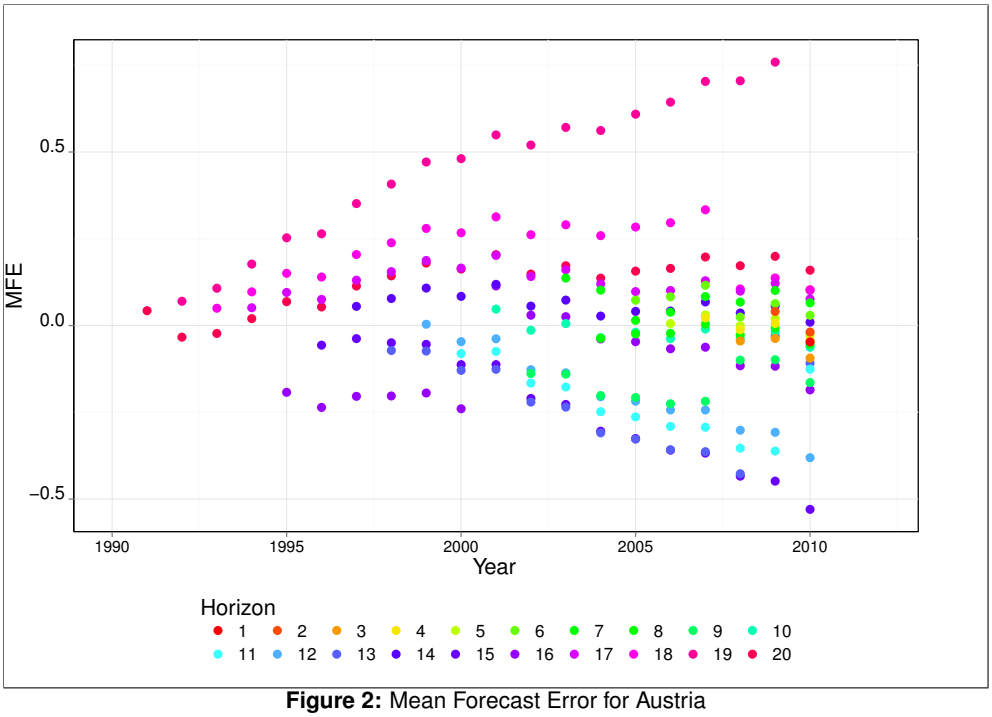

Figure 2: Mean Forecast Error for Austria

2.5 References

Box, G. E. P., and Cox, D. R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26(2), 211–252.

Chen, C., and Liu, L. M. (1993). Joint estimation of model parameters and outlier effects in time series. Journal of the American Statistical Association, 88, 284–297.

Hyndman, R.J. and Khandakar, Y. (2008) "Automatic time series forecasting: The forecast package for R", Journal of Statistical Software, 26(3).

Hubert, M., Rousseuw, P. J., and Verboven, S. (2002). A fast method of robust principal components with applications to chemometrics. Chemometrics and Intelligent Laboratory Systems, 60, 101–111.

Hyndman, R. J., and Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. International Journal of Forecasting, 24(3), 323–342.

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

Lee, R. D., and Carter, L. R. (1992). Modeling and Forecasting U.S. Mortality. Journal of the American Statistical Association, 87(419), 659–671.

Ramsay, J. O., and Silverman, B. W. (2005). Functional Data Analysis (Second Edition). Springer New York.

Shang, H. L. (2015). Statistically tested comparisons of the accuracy of forecasting methods for age-specific and sex-specific mortality and life expectancy. Population Studies, 69(3), 317–335.