State-Space Models to Estimate and Forecast Fertility

Cristina Rueda *

Pilar Rodríguez

Departamento de Estadística e Investigación Operativa

Universidad de Valladolid

Paseo Prado de la Magdalena, s/n

47005 Valladolid, Spain

Email: crueda@eio.uva.es

pilarr@eio.uva.es

Abstract. We introduce multivariate State-Space Models to estimate and forecast fertility rates that are dynamic alternatives to logistic representations for fixed time points. Strategies for Kalman filter and Quasi-Newton algorithm initialization, that assure convergence of the iterative fitting process, are provided. The broad impact of the new methodology in practice is proven using data series from Spain, Sweden and Australia and comparing the results with a recent approach based on Functional Data analysis and with official forecasts. Very satisfactory short and medium term forecasts are obtained. Besides, the new modeling proposal provides practitioners with

*Corresponding author.

several suitable interpretative tools and this application is one interesting example that shows the usefulness of the State-Space representation to model a real multivariate process.

Keywords: State-Space model, Kalman Filter, Fertility rates, Demographic forecast, Logistic model, Total Fertility Rate.

1 Introduction

The term fertility refers to the occurrence of births to an individual, a group or an entire population. It is determined by several biological, economic and social factors. The problem of estimating and forecasting fertility parameters is one that has a long tradition in demography. The population projections from fertility, mortality and migration components have always had a critical importance for policy-making because they set the basis for medium and long-term planning in many fields. Age-fertility rates are often used as inputs in the most popular population projection models.

Several approaches have commonly been used for projecting rates in demography. The simplest is to use average rates from recent years. Another approach is to suppose that the rates in the population to be projected will converge over time toward those found in another population or chosen by expert judgement.

On the other hand, approaches based on stochastic modeling have also been developed. These approaches have two advantages when compared with the simple extrapolation method: they use more historical information and provide prediction intervals. Projection from time series models, however, are often strongly affected by the structure of the models themselves and by

the changes in rates that occur during the base period. Therefore, many demographers support the use of mixed procedures, where external judgments and information on historical errors are included in the models. Interesting recent proposals along this line are those of Alho et al (2006) and Alders et al (2007). Nowadays, there is not unanimity about what is the best procedure because there are important issues with all the proposals and more statistical research is necessary. This paper is a contribution to this field.

The simplest way of making stochastic forecasts is to use univariate time-series models to analyse separate age-specific rates, but taken together, the separate analyses may not yield a plausible age-pattern (inconsistency problem). Therefore, it seems desirable to use modeling and forecasting methods that capture that smooth shape over age to produce consistent and accurate estimates.

Several approaches have been developed to analyse fertility and mortality patterns using stochastic models. Here, we comment on the two most widely used. The curve fitting approach, which involves fitting parametric curves to the age-specific rates, and the principal components approach, which involves using a matrix decomposition to obtain a linear transformation of the data with a simplified structure. Among the curve fitting models for fertility rates, the more familiar until recently were the Coale and Trussell (Coale and Trussell (1974)) model and the Gamma curve model (Thompson et al (1989)). Both have been used in many applications since their development, by Keilman and Pham (2000) or Scherbov (2002) among others. Schmertamnn (2003) recently proposed a new model based on constrained quadratic splines.

The second approach uses dimensional reduction techniques to linearly transform the rates. One of the most popular models is the Lee-Carter model

to forecast mortality rates (see Lee and Carter (1992)). Several authors have extended the Lee-Carter method (Booth et al (2006) and De Jong and Tickel (2006) among others). Moreover, Hyndman and Ullah (2007) and Hyndman and Booth (2008) have proposed a Functional Data (FD) approach that can also be considered as a successor of Lee-Carter and which has been applied to forecasting fertility and mortality rates. One main difference between the curve fitting and the dimension reduction approaches is that the model for the rates is defined using known parametric functions depending on age for the former but estimated functions depending on age in the latter. For a review of the different approaches see Booth (2006).

The aim of this paper is to propose a new approach to forecast fertility curves that uses the methodology of State-Space (SS) modeling. We will use the model to provide short and medium term forecasts of age-specific fertility rates and other fertility indices. The SS model is based on the Logistic model (LO) proposed by Rueda and Alvarez (2008). It uses simple functional expressions and the modeling and forecasting step is done simultaneously. The parameters of the model can be interpreted as indicators of the level of fertility, and shape of fertility curves. Finally, explanatory variables can also be easily incorporated into the model.

The procedure is validated using data series from different countries and periods and the results are compared with those obtained with the FD approach and with official forecasts. In all cases studied, very satisfactory results are obtained, both for the short and medium term forecast.

The LO model is presented in section 2. The SS model is defined in section 3 where a strategy is designed for Kalman filter and Quasi-Newton algorithm initialization that assure convergence of the iterative fitting process. In section 4, the SS model is applied to data series from Spain, Sweden

and Australia and finally, conclusions are drawn in Section 5.

2 Logistic model for fertility

2.1 Data and initial assumptions

We assume that birth counts and estimates of population at risk are available from vital registration and population census or population registers. To simplify the exposition, single year age data are used although the approach can also be used for other age groups. Let d be the total childbearing ages analysed. In the applications in section 4, d=30, the lowest childbearing are 16 for Spain and Sweden, as appears in the Eurostat data base, while it is 15 for Australia, and the highest 45 and 44 respectively. We use the following data: age-specific birth number for each calendar year, age-specific population numbers at 30th June in each year. For each year $t = 1, ..., n$, and age $j = 1, ..., d$; we define by,

$$\begin{aligned} b_j(t) &= \text{Births in the calendar year t for females of age j} \\ w_j(t) &= \text{Female population of age j exposed to risk in year t (30th June)} \\ m_j(t) &= \frac{b_j(t)}{w_j(t)} = \text{observed fertility rate for females of age j in calendar year t} \end{aligned}$$

Following the general consensus in actuarial modeling, we assume that births are generated by a Poisson process with intensity: $\rho_j(t)$. Under this model, $m_j(t)$ are the MLE of $\rho_j(t)$. The models proposed in this paper give smoothed estimators for $\rho_j(t)$.

In the next subsection, the LO model is defined and some properties of the model are commented on.

2.2 Description and properties of LO models

The r-dimensional logistic model to analyse fertility curve for a given moment in time, t, is given by:

$$[LO]_r \qquad \log(\rho(t)/(1-\rho(t)) = A\beta_t \tag{2.1}$$

where $\beta_t = (\beta_0(t), \beta_1(t), ..., \beta_{r-1}(t))'$ is the parameter vector and A is a known $d \times r$ design matrix with orthogonal columns defined as a function of power of age:

$$A = (A_0, A_1, ..., A_{r-1}) \qquad A_k = (A_{1k}, ..., A_{dk})' \qquad 0 \leq k \leq r-1$$

The expression of the first three columns are given below using basic statistics from the age distribution :

$$\begin{aligned} A_{j0} &= 1, A_{j1} = (j - \overline{a}), A_{j2} = (j-\overline{a})^2 - S_a^2, A_{j3} = (j-\overline{a})^3 - \frac{K_a}{S_a^2}(j-\overline{a}) \\ \overline{a} &= \frac{1}{d}\sum_{j=1}^{d} j \quad S_a^2 = \frac{1}{d}\sum_{j=1}^{d}(j-\overline{a})^2 \quad K_a = \frac{1}{d}\sum_{j=1}^{d}(j-\overline{a})^4 \quad j = 1, ..., d \end{aligned}$$

The suitability of the fitted model to describe fertility curves is evaluated in Rueda and Alvarez (2008) with data from 226 countries. In that paper, the fit of the $[LO]_r$ model is compared with that of the Quadratic Spline (QS) model of Schmertmann (2003) and the (CT) model form Coale and Trussell (1974). The logistic model $[LO]_4$ gives better results than the CT model and comparable results to the QS in developed countries (the three models defined using 4 parameters). The incorporation of power of age of higher order ($[LO]_r, r > 4$) significantly improves the fit in many countries but for some countries and years, a model with fewer parameters suffices. We have decided to work with the model $[LO]_7$ as a standard for the dynamic analysis, for single-year age groups, in the following sections.

The parameters of the model can be interpreted as measures of the level (or period *quantum* $\beta_0(t)$), and shape (or *tempo* ($\beta_i(t)$) of fertility curves, as shown in Rueda and Alvarez (2008). For a discussion of *tempo* and *quantum* concepts in demography see Van Imhoff and Keilman (2000) and Sobotka (2003). Then, in particular, changes in the Total fertility Rate (TFR) values in a period can be interpreted as changes in *quantum* and/or *tempo* via the changes in the observed beta series. These interpretative properties are used in practice to describe past and future fertility using real data in section 4.

3 Definition of State-Space models

In this section we present the SS representation to analyse series of rates. SS modelling provides a unified methodology for treating a wide range of problems in time series analysis, allowing considerable flexibility in the specification of the parametric structure for time series processes. In this approach, it is assumed that the development over time of the system under study is determined by an unobserved series of state-vectors: α_t with which are associated a series of observations: Y_t. The linear SS model can be defined using two equations. The first is known as the observation equation and expresses the vector observation as a linear function of a state vector plus a noise. The second equation, called the state equation, determines α_{t+1} in terms of α_t and a noise term. It is assumed that the initial state vector is uncorrelated with all the noise terms, so the state vector then has the Markov property. In a general SS representation, neither the vector observation nor the state vector is assumed to be stationary. A large number of well known time series models have an SS representation. To find the estimates of the state vector, the SS methodology uses the well-known Kalman filter. The Kalman filter is

a recursive algorithm, that is, it is based on formulae in which we calculate the value at time t+1 from earlier values for t, t-1,...,1. The question of how these recursions are started up at the beginning of the series is called initialization. The Kalman filter provides a unified approach to prediction and estimation for all processes that can be given by an SS representation. When the models depend on unknown parameters, the estimation is provided by maximum likelihood. For maximization of the log-likelihood, we use a Quasi-Newton algorithm that starts with a trial value for the parameter vector. For a non expert reader in state space modeling, we recommend the book by Commandeur and Koopman (2007) and chapter 8 from the book by Brockwell and Davis (2002). For a more detailed study, see Durbin and Koopman (2001).

To derive the final expression as a gaussian SS model given below, we have assumed that the logits vector , $Y_t = \left(\log\left(\frac{m_1(t)}{1-m_1(t)}\right), ..., \log\left(\frac{m_d(t)}{1-m_d(t)}\right)\right)'$, is conditionally normally distributed with mean $A\beta_t$ and that β_t are independent ARIMA processes. The logistic model defined in (2.1), the orthogonality of the design matrix A, the standard normal approximation to the Poisson and the time component of the data are the properties that support these assumptions.

The SS model for fertility is written in a usual form with two equations. The observation equation has the structure of a linear regression model with coefficients that depend on time and the state-equation represents the development of the system over time, as follows :

$$\begin{array}{llll} \textit{Observation equation} & Y_t = B\alpha_t + \varepsilon_t & & \varepsilon_t \rightsquigarrow N_d(0, H) \\ \textit{State equation} & \alpha_{t+1} = R\alpha_t + \eta_t & & \eta_t \rightsquigarrow N_p(0, Q) \\ 1 \leq t \leq n & \alpha_1 \rightsquigarrow N(a_1, P_1) & & \end{array} \tag{3.1}$$

To obtain the above SS representation we first calculate the logistic estimators fitting the model (2.1). These latter estimators are obtained using standard software for logistic regression that uses the observed fertility rates and female population figures as inputs and provides estimators for the mutivariate beta process: $\beta_t = (\beta_0(t), \beta_1(t), ..., \beta_6(t))'$ as output. ARIMA processes are then fitted to each component of the output series. The analysis of data sets from different countries points to nonstationary models selected from ARIMA(0,1,0) or ARIMA(1,1,0) for each component; this is nothing new, as fertility series in the literature have been traditionally fitted using small order ARI processes. The p-dimensional state vector α_t, where $p \geq 7$, and the form of the state-equation (R), are derived from β_t and the differences needed to define the ARIMA processes selected. The number of the differences considered determines the exact value of p and the matrix B is obtained from the relation: $A\beta_t = B\alpha_t$, $B = [A, 0]$,where 0 is $d \times (p-7)$ (See Durbin and Koopman (2001), pag 46).

Moreover in the SS representation (3.1), the error terms ε_t and η_t are assumed to be serially independent of each other at all time points and H and Q are unknown diagonal matrices, that do not depend on time, measuring the model errors (disturbance variances). Also, the initial state vector α_1 is assumed to be $N(a_1, P_1)$ independently of $\varepsilon_1, ..., \varepsilon_n$ and $\eta_1, ..., \eta_n$, where a_1 and P_1 are assumed to be known and must be provided to initialize the Kalman filter, a_1 is derived using the logistic estimators and P_1 is initialized as 0.

The vector of model parameters is given by the parameters in R and the $d+$ p disturbance variances in H and Q. Initial guesses for the parameters are also needed to use a Quasi-Newton maximization algorithm to derive MLE. The initial values are for the disturbance variances in Q, the values

derived using the logistic estimators. For H, the mean values, in the observed time, of the asymptotic estimated Poisson variance matrix for the logistic transform. The asymptotic Poisson variance of $log(m(t)/(1-m(t)))$ is by the Taylor approximation: $H_t^p = diag([w_j(t)\rho_j(t)(1-\rho_j(t))^2]^{-1})$ and the estimated variance is given by $\widehat{H}_t^p$, $\widehat{H}_t^p = diag([w_j(t)m_j(t)(1-m_j(t))^2]^{-1})$. Finally, the autoregressive parameters which define matrix R are initialized using nonnegative values (we use 0.5).

We illustrate the way the matrices B and R are derived and also the initial value a_1 using the model fitted to the Swedish data analysed in section 4 as an example. In this case, the ARIMA processes for the initial series of logistic estimators are as follows: ARIMA(1,1,0) for $\beta_i(t), i \leq 3$ and ARIMA(0,1,0) (random walk without drift) for $\beta_i(t), i \geq 4$. Then,

$$\alpha_t = (\beta_0(t), ..., \beta_6(t), \nabla\beta_0(t+1), \nabla\beta_1(t+1), \nabla\beta_2(t+1), \nabla\beta_3(t+1))'$$

where,

$$\nabla\beta_i(t+1) = \beta_i(t+1) - \beta_i(t) \quad i = 0, 1, 2, 3$$

The initialization is given by,

$$\alpha_1 = a_1 = \left(\widehat{\beta}_0(1), ..., \widehat{\beta}_6(1), \nabla\widehat{\beta}_0(2), \nabla\widehat{\beta}_1(2), \nabla\widehat{\beta}_2(2), \nabla\widehat{\beta}_3(2)\right)'$$

where $\widehat{\beta}_i(1)$ and $\widehat{\beta}_i(2)$ are the logistic estimators given by (2.1) for the first two years. Therefore, matrix R is 11×11 and depends on 4 parameters which are the autoregressive coefficients of the beta process models for $\beta_i(t), i = 0, 1, 2, 3$, as follows :

$$\mathbf{R} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \phi_0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \phi_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \phi_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \phi_3 \end{pmatrix}$$

B is a 30×11 matrix where the first 7 columns equals those in matrix A and the last 4 columns have values equal to zero. Q is diagonal 11×11 with 7 parameters:

$$Q = diag(0, 0, 0, 0, \sigma^2_{\eta 5}, \sigma^2_{\eta 6}, ..., \sigma^2_{\eta 11})$$

Alternatively, a model with the same structure as model (3.1), where $\widehat{H}^p_t = diag([w_j(t)m_j(t)(1-m_j(t))^2]^{-1})$, can be fitted to the data. The results are similar to those obtained using H. To fit the state space model, we use the software package SsfPack in Ox computing environment (Koopman et al (1998)). The inputs to start the program are : the observed rates, the exact form of matrices B and R , the initial values for the parameter vector, and the distribution of the initial state $\alpha_1 \rightsquigarrow N(a_1, P_1)$. The outputs are the parameter estimates and the h-step ahead forecast together with their estimated standard errors. we summarize below the steps of the fitting process:

1. Model Definition: ARIMA processes for the betas are derived using the logistic estimators from (2.1). The selected processes determine the dimensionality of the multivariate alpha process and the exact form of matrices R and B.

2. Kalman filter Initialization: an initial guess for the vector a_1 is derived using the logistic estimators for $t = 1$ and $t = 2$ and P_1 is initialized as 0.

3. Estimation of parameters: MLE are derived using a Quasi-Newton maximization algorithm. The values used to initialize the algorithm are for the disturbance variances in Q, the values derived using the logistic estimators. For H, the mean values, in the observed time, of $\widehat{H}_t^p$. The autoregressive parameters in matrix R are initialized as 0.5.

4. Forecasting: smoothed estimators for the beta process, forecasted values and prediction intervals are derived using the Kalman filter and smoother.

The iterative estimation process converged in all cases we tried. This is not necessarily the case if other initial values are used. To measure the prediction capacity, the model is fitted reserving the last three (five) years for each country and period and the corresponding SSE for these three (five) years alone is computed.

Through the selection of specific ARIMA processes fitted to beta series, different fertility scenarios for the future can be assumed. This means that the SS approach allows demographers to draw the form of the fertility curve for the following years selecting models that stagnate, accelerate, or decelerate current trends in fertility levels (controlling $\beta_0(t)$) and other important characteristics (controlling $\beta_i(t)$, $i \geq 1$).) This can be also done in a similar

way to Lee's(1993) proposal to constrain the ultimate level of the TFR forecast. Therefore, the SS approach permits changes to be incorporated into the age pattern of fertility which are expected to be different in the future from in the past.

As we focus in this paper on the comparison of the SS and the FD approach, we introduce next the main features of the FD approach and the main differences with the SS approach. The FD approach uses a similar model structure to that of the SS approach : $g(m(j,t)) = \sum_{k=0}^{r-1} \beta_k(t)\phi_k(j) + e_{tj} + \epsilon_{tj}$, where e_{tj} and ϵ_{tj} are the model and observational error terms respectively and g a Box-Cox transformation, often the logarithm, and where $\phi_k(j)$ is a set of orthonormal basic functions estimated from the data using functional data analysis in a similar manner to Ramsay and Silverman (1997, Chapter 6). The prediction intervals in the FD approach are obtained by forecasting the beta coefficients, using univariate time series models, and from the estimation of the observational error and model error variance. (See the papers of Hyndaman and Ullah (2007) and Hyndman and Booth (2008) for details). The main difference between the SS and FD approaches is that in the former the base functions $\phi_k(j)$ are fixed but in the latter they are data dependent. In our application, at least, this appears to make the SS approach somewhat less dependent on the choice of the data period than the FD. Moreover, as the parametric series, in the SS approach, are interpreted as measures of changes in the level and shape of fertility curves, a comparison of the beta series predictions from different base periods is a good strategy to select a reasonable base period for medium term forecasts: long enough to provide good estimators but also short enough to reject the non relevant data for the near future. We illustrate these ideas with the analysis of real data series in the next section.

4 Examples

The cases of Spain, Australia and Sweden will be analyzed in this section. These countries have been selected for several reasons. The analysis of the Spanish fertility is interesting as the drop in the birth rate in Spain has occurred in part through the adoption of general European patterns, but with three essential points: a much greater relative and absolute, a noticeably smaller final TFR rate and also a considerable smaller fertility rate below thirty (Cabré(2003) and Fernández de la Mora y Varela (2000)). For the case of Sweden, demographic researchers have paid great attention to the study of fertility in this country, because good quality data is available, and also because Sweden was one of the first countries where fertility levels under the replacement level (2.1 in developed countries) were observed. See Kohler and Ortega (2002), Andersson (2004), Hoem (2005) and references in these papers. Finally the data from Australia has also been extensively analysed by an important group of demographers and statisticians from the country who have produced several of the most interesting recent papers in the field (Booth (2006), Hyndman and Ullah (2007) and Hyndman and Booth (2008)). This research checked the proposed FD approach with data series from this country. Then, a fair comparison with results from the SS approach is also feasible and of special interest in this case.

The European data have been obtained from Eurostat data base

(http://epp.eurostat.ec.europa.eu), for 1971-2005 in Spain and for 1955-2005 in Sweden. The Australian data for 1921-2003 comes from the R package 'Addb' from the personal web page of Rob Hyndman. Also from this web page the 'demography' R package is used to implement the FD approach (Hyndman (2006)) and obtain summary statistics that are good to compare different aspects of both approaches. Moreover, official TFR forecasts from

Eurostat, the Spanish National Statistical Institute, and the Australian Bureau of Statistics are also used for comparison with the predicted TFR from the SS approach.

For Spain, as short series are available, only a set of forecasts up to 2020 are provided using the complete period 1971-2005. For Sweden, two sets of predictions up to 2020 were constructed: one based on the annual data series 1955-2005, another based on annual figures observed during the period 1975-2005. For Australia, the country with the longest series, we construct three series of predictions based on data from the periods: 1921-2003, 1955-2003 and 1975-2003.

The general features of past and future fertility are analyzed based on the beta series in subsection 4.1 for Spain, Sweden and Australia. The short and medium term forecasts with the 80% prediction intervals are also included. In subsection 4.2, the official forecast and forecasts from the FD approach are compared with the SS forecasts from different base-periods, using TFR series. The prediction capacity is also calculated, reserving the last three (five) years in each country.

4.1 Past and future Fertility by the State Space approach

In the following presentation we will discuss the interpretative properties of the beta series estimates from fitting model (3.1) to real data. The R matrix is derived using the longest period for each country. This matrix depends on several autoregressive parameters which are then estimated for each period. Alternatively, different ARIMA models could have been selected for each period. However, we have checked that the predictions are quite similar with different ARIMA models, which fit the data reasonably well. We reproduce

here only the most significant estimated series $\beta_i(t)$ i=0,1,2. We also use the TFR series to illustrate the comments.

In Spain, the TFR has gone down from values around 2.9 in 1975 to 1.34 in 2005, one of the lowest values in the world. Figure 1 illustrates this fact. Beta series are also drawn in figure 1, which explain that the fertility change in Spain is mainly due to a *quantum* effect ($\beta_0(t)$). Also, in the last 15 years, changes in the shape of the fertility curve have been observed, as illustrated by the $\beta_1(t)$ trend in this period. The consequence is that lower TFR values, than expected without shape changes, have been observed. By instances, figure 1 shows that while $\beta_1(1991)$ and $\beta_1(2002)$ are quite similar, $TFR(1991)$ is larger than $TFR(2002)$.

The opinion of Bijak (2004), among other experts, is that a slow recovery on fertility level is predictable. This is the scenario that our forecast gives; $\beta_i(t)$ trends show that the recovery with the SS approach will be due mainly to the increasing values for $\beta_0(t)$. The TFR for 2020 is 1.57. A slight increase of other $\beta_i(t)$ is also predicted in the future. The result is that higher rates are predicted in the early to mid-twenties with respect to rates in 2005. This is an interesting feature of Spanish fertility curves that started in 2000 and that fits the model and is also predicted to continue in the future (see figure 4).

Swedish fertility (Figure 2) has experienced special behavior: after decreasing in the early seventies, around 1977 this trend stopped, so that for several years, the TFR remains more or less constant. Then, in the late 1980s it increased, decreasing again in the 1990s to 1.51 in 1996. Since then, the TFR has increased to 1.75 in 2005. In this country, change in fertility has also been characterized by important changes in the level ($\beta_0(t)$) but also the mean age of childbearing has increased since 1975 with a stable period in

the late 1980s (the same pattern as $\beta_1(t)$) and important changes were also observed in $\beta_2(t)$ from 1970 to the late eighties. Again, as in Spain, lower TFR values than expected without shape changes, have been observed, in the period from 1995 to 2002.

To obtain the forecast in Sweden we have used two base periods: 1955-2005 and 1975-2005. Results are very similar in both cases (figure 2). The *quantum* component is predicted to be stable in the future. As the longer base period is more informative in forecasting future trends, then this is taken into account by us when obtaining the fertility parameters in 2020. The TFR value will increase very slowly as a consequence of changes in the *tempo* component. Bijak (2004) says that for Sweden one can expect the recent high TFR (1.85 in 2006) values to be quite good predictors of comparatively high fertility in the future, and other experts share this opinion.

Australian fertility has experienced important changes in the level and the shape components throughout the long period starting in 1921. The trends in beta series illustrates this fact (figure 3). Having reached a TFR of 3.0 during the early-1920s, Australian fertility was relatively low during the 1930s, falling to 2.1 children per woman in 1934. In 1961, it peaked at 3.5 children per woman. Since then, fertility has declined to 1.73 in 2001 (1.76 in 2002 and 1.75 in 2003). However, the *quantum* component $\beta_0(t)$ decreased until the late eighties and has since increased. The low TFR values observed since 1990 are again a consequence of *tempo* changes. The reverse effect is observed in the period 1940-1960.

Australian fertility data from 1921 is available for analysis and the SS model is fitted using three different base periods: 1921-2005,1955-2005 and 1975-2005. As in Sweden, small differences are observed in the future depending on the base period selected (figure 3). As current childbearing behavior

is very different from that of women in the 1930s and data from 1955 gives us a sufficiently large series and good estimators, we have selected this period as the base period to get and interpret future fertility parameters. A behavior similar to that of Sweden is predicted for future Australian fertility. The TFR value will increase as a consequence of changes in the *tempo* component up to 1.81 in 2020 (figure 3). Besides, it is known that the TFR has had an upward trend, reaching 1.81 babies per woman in 2006. Hugo (2007), in agreement with other experts, says that the most reasonable interpretation of recent trends is that there is a degree of stability around 1.8 births per woman. The forecasted values with the SS approach are also interpreted in the same way.

Figures 4,5 and 6 show the forecast fertility curve along with 80% prediction intervals for one step ahead and 2020 for the three countries under study. The Spanish pattern is significatively different from the other two countries for ages 15 to 30. This fact agrees with the hypothesis of several experts that pronounced regional differences in European fertility are likely to prevail.

4.2 A comparative study

Figures 7, 8 and 9 show the TFR forecasted values obtained with the SS, FD and official forecasts. The FD approach has been implemented using k=6 basic functions and state space exponential smoothing time series models. Different base periods going until 2005 for Spain and Sweden and 2003 in Australia have been considered. The official forecasted values have been obtained from data until 2004. To simplify the graphical representation we have only drawn the series for the medium fertility assumptions. It is interesting to note that the recent TFR values in the three countries are higher than

those observed in 2004, being, 1.33(2004), 1.34(2005), 1.37(2006) in Spain, 1.75 (2004), 1.77(2005), 1.85 (2006) in Sweden, and, 1.76(2004), 1.77(2005), 1.81(2006) in Australia. According to this, the next official forecasts are likely to be higher.

For the three countries and selected base periods of section 4.1, the SS approach provided higher TFR forecasted values in 2020 than the official forecasts, but close to them. In addition, the forecasted TFR for alternative base periods are not far from each other in all the cases. On the other hand, the FD approach also gives values close to official forecasts in the cases of Sweden and Australia for selected periods (1921-2003 in Australia and 1955-2005 in Sweden). However, the forecasted TFR for Spain are far from the results obtained by either official or SS approaches and also the influence of the base periods is stronger, as the forecasted TFR values from different periods are far from each other. The choice of the base period has also consequences for the predicted age patterns. Let us consider the example of Australia. Figure 10 shows the forecast fertility schedules for 2020 using data from periods 21-03 , 55-03 and 75-03. The differences in the FD forecast patterns are stronger than those for the SS approach. Moreover, in the latter case, the differences in the predicted slope for $\beta_1(t)$ explains the differences in forecasting patterns. Those from the 75-03 and 55-03 data, where the predicted slope of $\beta_1(t)$ is increasing, results in a pattern that corresponds to an increasing trend in the mean age of fertility. Meanwhile, The one from the 21-03 data, where the predicted slope of $\beta_1(t)$ is more or less constant, results in a pattern that corresponds to no trend in the mean age of fertility.

Finally, in Table 1 we have included the SSE values, for 2003-2005 in Spain and Sweden and for 2000-2003 in Australia, after the model is fitted, reserving the last three years and Table 2 shows the results reserving five

years for each country and period. The prediction capacity for the short term across periods and countries is very high with both SS and FD approaches, the SS being the best predictor in the cases where longer periods are used.

5 Conclusions

From a Statistical point of view the SS approach has the advantage that the modeling and forecasting steps are done simultaneously and simple functional expressions are used; the model permits the analysis of parity-age-specific data or grouped data and the inclusion of covariables. From a demographic point of view, a useful feature is that the most important parameters have natural interpretations. In this paper, we have only begun to exploit these possibilities by explaining past and future fertility and selecting the base period.

We have focused in this paper on the comparison of the SS approach with the FD approach from a statistical point of view. For some periods implausible forecasts have arisen with the FD approach. In practice, such forecasts would not be used, and the forecasting approach itself would be modified in any one of the many available ways. However, a potential advantage of the SS approach is that it seems to be less sensitive to the choice of the data period.

To carry out the analysis we have used the SsfPack software by Koopman et al (1998). The software can be obtained freely at http://www.ssfpack.com. SsfPack is a suite of C routines for carrying out computations involving the statistical analysis of univariate and multivariate models in state space form. The full implemented link is Ox, which is an object-oriented statistical system. We have prepared programs to analyse and forecast fertility using the

logistic-SS models with this software framework (the sample programs and our advisor are available by emailing us).

Acknowledgements: The authors thank the referees and the associate editor for useful comments that led to improvements in the presentation of this article.

References

Alders, M.; Keilman, N. and Cruijsen, H (2007). Assumptions for long-term stochastic population forecast in 18 European countries. *Eur.J.Population.* **23**.33-69.

Andersson, G. (2004). Childbearing Developments in Denmark, Norway, and Sweden from the 1970s to the 1990s: A Comparision. *Demographic Research*. Special collection 3, Article 7.

Alho, J; Alders, M.; Keilman, N.; Cruijsen, H; Nikander, T. and Pham, D.Q. (2006). New forecast: Population decline postponed in Europe. *Statistical Journal of the United Nations. ECE* **23**.1-10.

Australian Bureau of Statistics (2006). Birth, Australia, 2006, various issues. *Catalogue no. 3301.0.*Canberra.ABS

Bijak,J. (2004). Fertility and Mortality Scenarios for 27 European Countries, 2002-2052. CEFMR, Working paper 3/2004.

Booth, H., Hyndman, R. J., Tickle, L., De Jong, P. (2006). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research* **15(9)**,289-310.

Booth, H., Maindonald, J. and Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies* **56**, 325-336.

Booth, H (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting.* **22**(3), 547-581.

Brockwell, P.J. and Davis, R.A. (2002). *Introduction to Time Series and Forecasting.* Springer-Verlag: New York.

Cabré, A.(2003).Facts and Factors on Low Fertility in Southern Europe: The Case of Spain. *Journal of Population and Social Security (Population), Supplement to Volume 1*, 309-321.

Coale, A.J. and Trusell, T.J. (1974). Model Fertility Schedules: variations in the age structure at childbearing in human populations. *Population Index*, **40**, 185-258.

Commandeur, J.J.F. and Koopman, S.J. (2007). *An introduction to State Space Time Series Analysis.* Oxford University Press.

De Jong, P. and Tickle, L. (2006). Extending Lee-Carter forecasting. *Mathematical Population Studies*,**13**, 1-18.

Durbin, J. and Koopman, S.J. (2001). *Time Series Analysis by State-Space Models.* Oxford: Oxford University Press.

Fernández de la Mora y Varela, G. (2000).La despoblación de Espana. *Razón Espanola.*, **101**, 281-295.

Hoem, J.M. (2005).Why does Sweden have such high fertility?. *Demographic Research.***13**(22), 559-572.

Hyndman, R.J. (2006). Demography: Forecasting mortality and fertility data, R package version 0.97. *URL:http://www.robhyndman.info/Rlibrary/demography.*

Hyndman, R.J. and Booth, H. (2008). Stochastic Population Forecast using funtional data models for mortality, fertility and migration *International Journal of Forecasting.***24(3)**, 323-342.

Hyndman, R.J. and Ullah, Md.S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach *Computational Statistics & Data Analysis*,51,4942-4956.

Hugo, G.(2007). Recent Trends in Australian fertility *O & G*, Vol 9,No 2. Winter 2007.

Lee, R.D. and Carter, L.R. (1992). Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, **87**, 659-671. ck *O & G*, Vol 9,No 2. Winter 2007.

Lee, R.D.(1993). Modeling and Forecasting the time series of U.S. Fertility: Age distribution, range and ultimate level. *International Journal of Forecasting* , **9**, 187-202.

Keilman, N. and Pham, D.Q. (2000). Predictive Intervals for Age-Specific Fertility. *European Journal of Population*, **16**, 41-66.

Kohler, H.P. and Ortega, J.A. (2002). *tempo*-adjusted period parity progression measures: Assessing the implications of delayed childbearing for fertility in Sweden, the Netherlands and Spain. *Demographic Research*,**6**, 145-190.

Koopman, S.J, Shephard,N. and Dornik,J.A. (1998). Statistical algorithms for models in state space using SsfPack2.2. *Econometrics Journal* 1.1-55

Ramsay, J.O. and Silverman, B.W.(1997). *Functional Data Analysis.* Springer-Verlag: New York.

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

Sanderson, W.C.; Scherbov, S.; O'Neill, B.C. and Lutz, W. (2004). Conditional probabilistic population forecasting. *International Statistical Review, 75*: 157-166.

Scherbov, S. and Vianen, H.V. (2002). Period Fertility in Russia since 1930: an application of the Coale-Trussell fertility model. *Demographic Research* ,6: 456-470.

Schmertmann, C.P. (2003). A system of model fertility schedules with graphically intuitive parameters. *Demographic Research, 9*: 88-110.

Sobotka, T. (2003). *tempo-quantum* and period-cohort interplay in fertility changes in Europe. Evidence from the Czech Republic, Italy, the Netherlands and Sweden. *Demographic Research, 8*: 152-213.

Thompson, P.A.; Bell, W.R.; Long, J.F. and Miller, R.B. (1989). Multivariate time series projections of parameterized age-specific fertility rates. *Journal of the American Statistical Association, 84*: 689-699.

Van Imhoff, E and Keilman, N. (2000). On the Quantum and Tempo of Fertility: Comment. *Population and Development Review, 26*: 549-553.

Country and Period	State-Space	Funtional Data
75-00 Australia	**0.000396**	0.001264
55-00 Australia	**0.000398**	0.003290
21-00 Australia	**0.000368**	0.000528
71-02 Spain	0.000220	**0.000166**
75-02 Sweden	0.000801	**0.000702**
55-02 Sweden	**0.000726**	0.002067

Table 1: Prediction capacity of SS and FD approaches. SSE for the last three years

Country and Period	State-Space	Funtional Data
75-98 Australia	**0.00084921**	0.00186771
55-98 Australia	**0.00107156**	0.0179468
21-98 Australia	**0.00108486**	0.00127282
71-00 Spain	0.00625957	**0.00296848**
75-00 Sweden	0.00621306	**0.00611426**
55-00 Sweden	**0.00653331**	0.00921913

Table 2: Prediction capacity of SS and FD approaches. SSE for the last five years

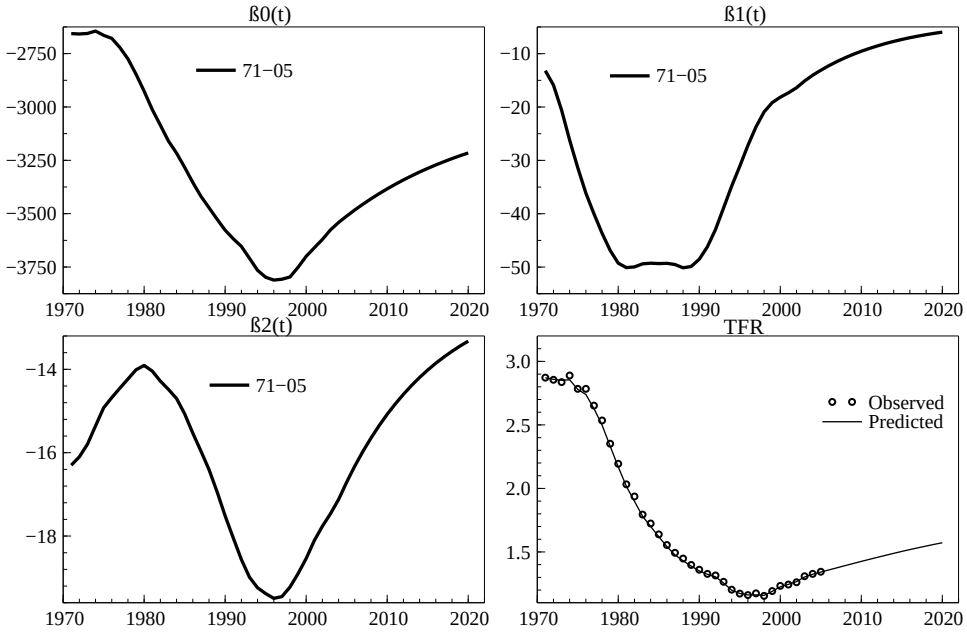

Figure 1: Spain: SS beta series.Observed and predicted TFR values.

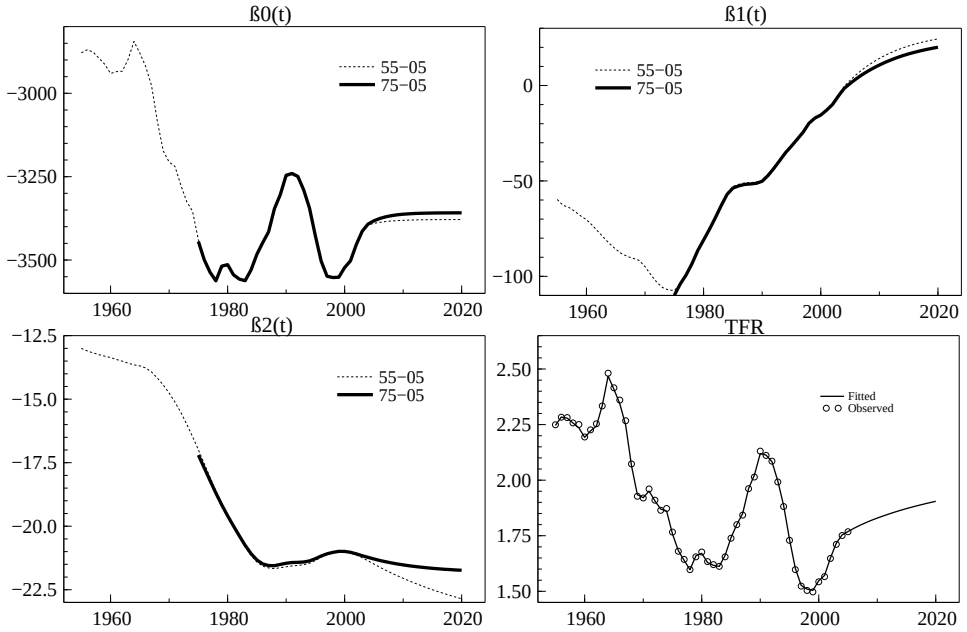

Figure 2: Sweden: SS beta series.Observed and predicted TFR values.

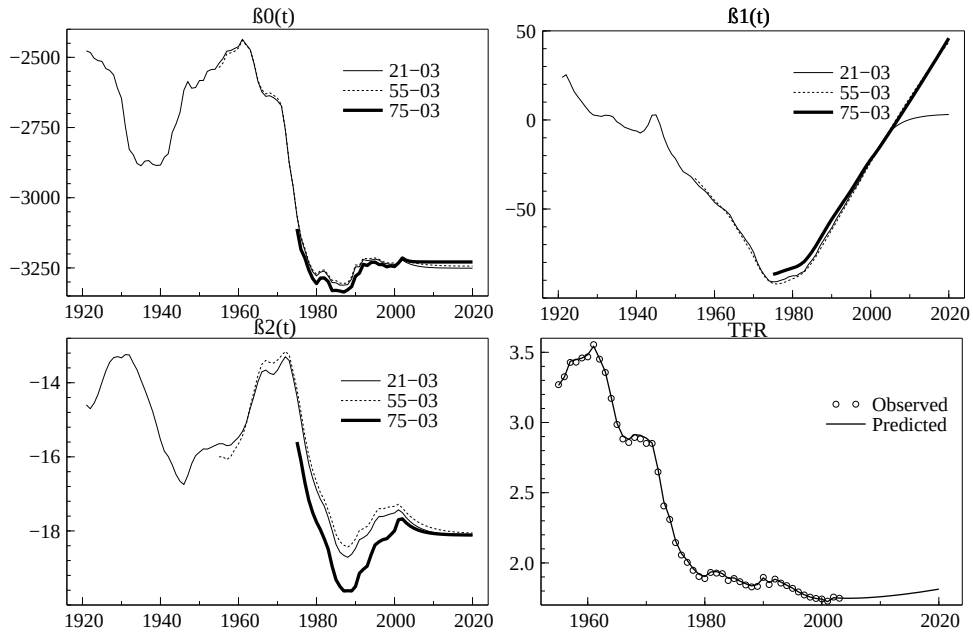

Figure 3: Australia: SS beta series.Observed and predicted TFR values.

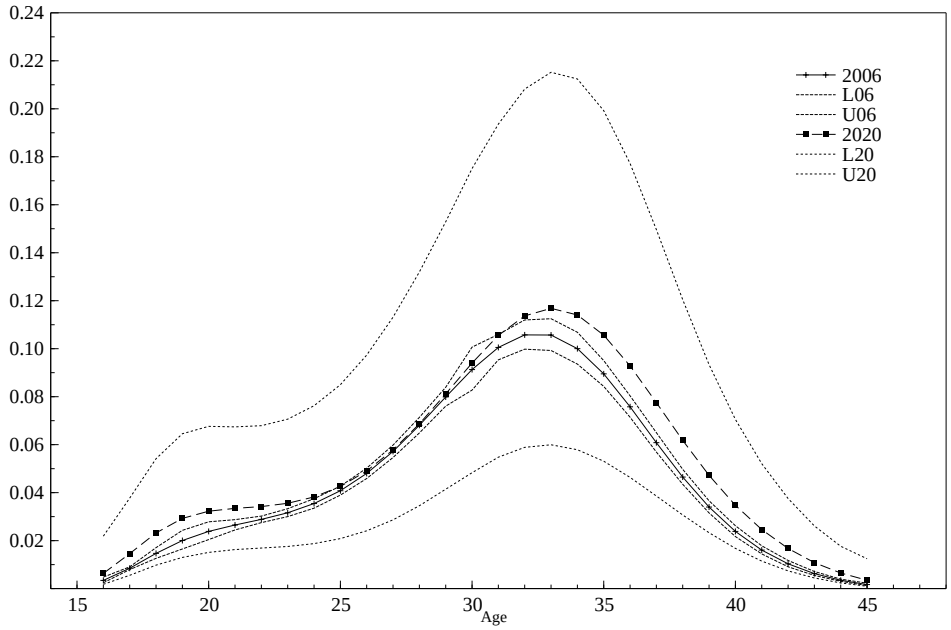

Figure 4: Spain: Forecast fertility rates for 2006 and 2020, along with 80% prediction intervals.

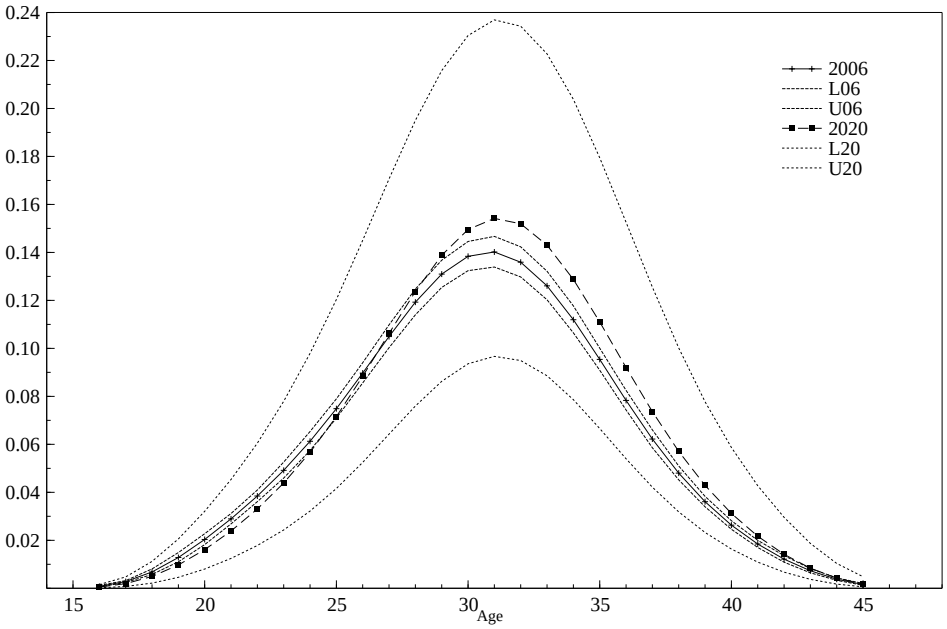

Figure 5: Sweden: Forecast fertility rates for 2006 and 2020, along with 80% prediction intervals.

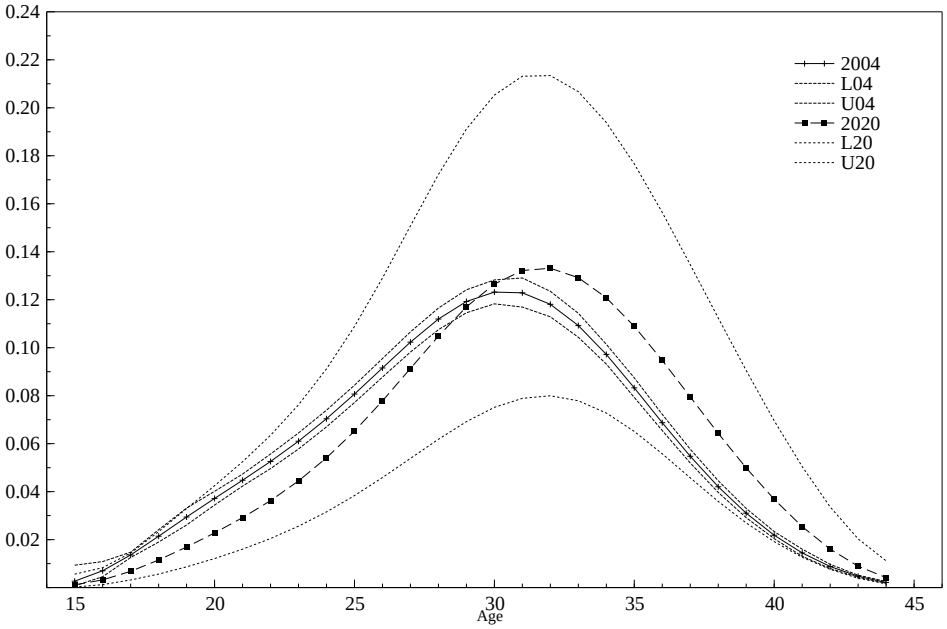

Figure 6: Australia: Forecast fertility rates for 2004 and 2020, along with 80% prediction intervals.

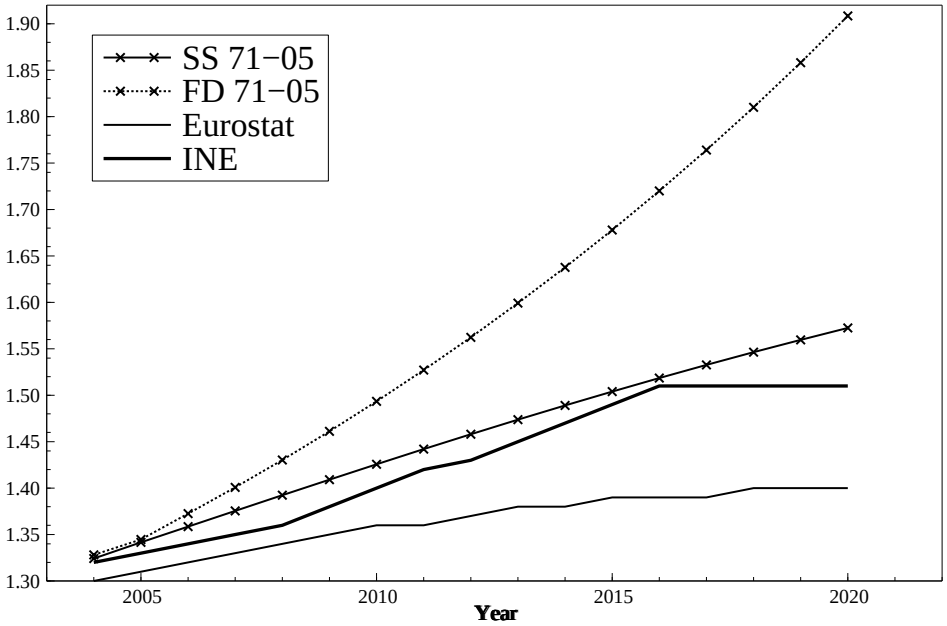

Figure 7: Spain: SS, FD and official Forecast from Eurostat and INE (Spanish Statistical National Institute) of TFR for 2006-2020

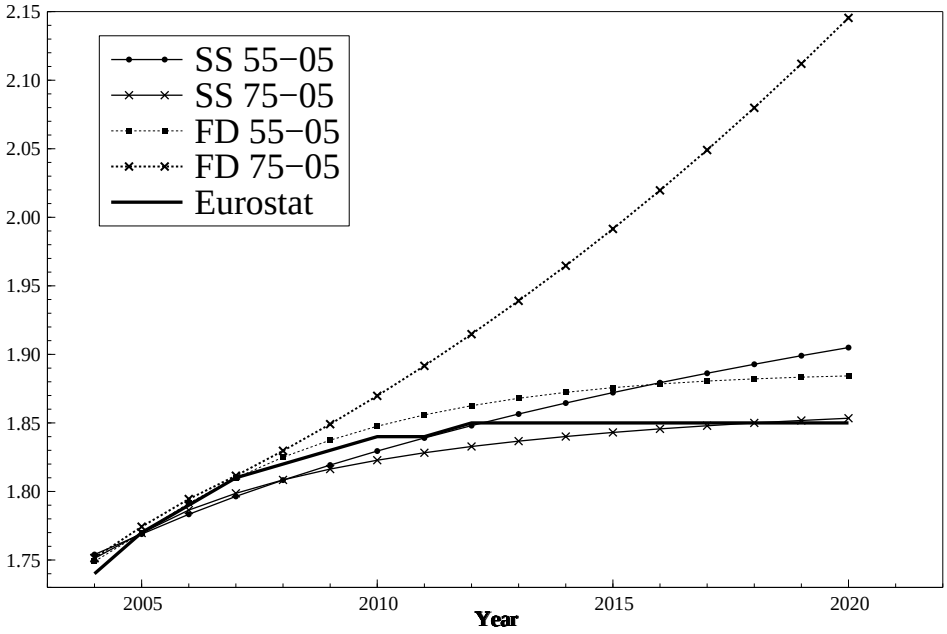

Figure 8: Sweden: SS, FD and official Forecast from Eurostat of TFR for 2006-2020. Using different base periods.

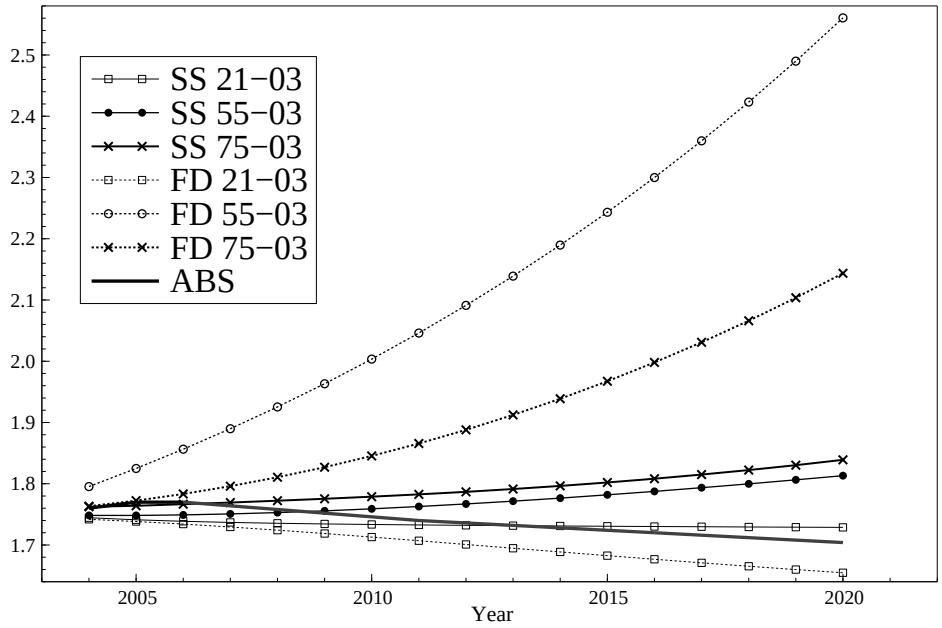

Figure 9: Australia: SS, FD and official Forecast from Australian Bureau of Statistics (ABS) of TFR for 2006-2020. Using different base periods.

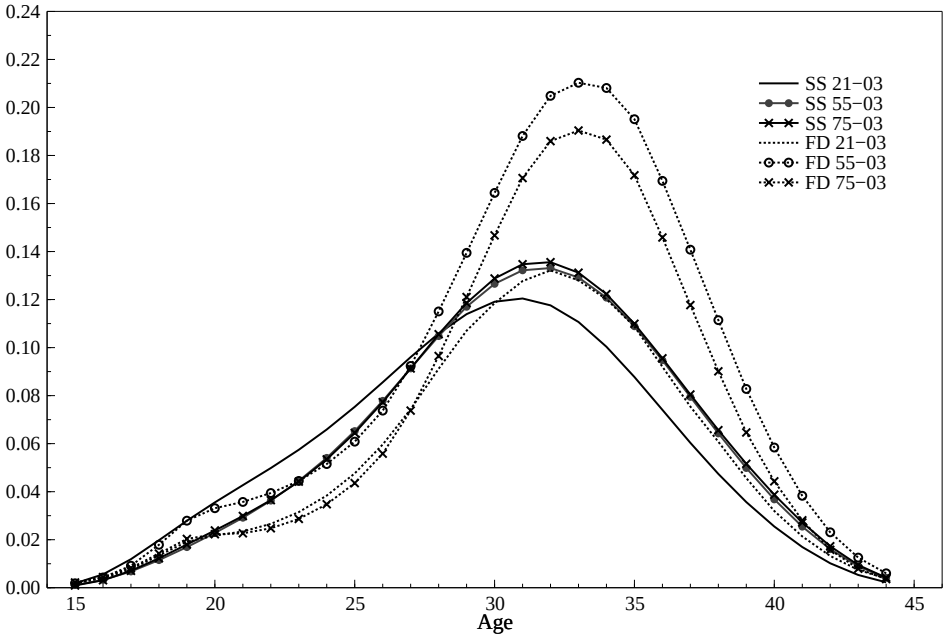

Figure 10: Australia: SS, FD Forecasts for 2020 using different base periods.