**Sesión** **1205**
**Título** **A STATE-SPACE MODEL FOR FERTILITY LOG ODDS**
**Autores** **Cristina Rueda, Pedro C. Alvárez and Pilar Rodriguez**
**Contact** Dpto. Estadística e Investigación Operativa, Facultad de Ciencias, Universidad
**Details**de Valladolid. C/ Prado de la Magdalena s/n, 47005 Valladolid, Spain
E-Mail: crueda@eio.uva.es

## 1.- INTRODUCTION

In this paper we present a new approach designed to analyze age-specific fertility parameters. For each fixed moment in time t we define a linear model for the Log Odds of fertility. From this model a State-Space formulation is obtained that allows smoothing the parameters over time simultaneously.
The flexibility of the proposed method permits to use Odds for each individual age and parity or Odds for age-group and/or all parities together. The parameter series of the model are interpreted in terms of the level and timing of fertility (Bongaarts and Feeney (1998)) and then different scenarios can be used to forecast fertility measures.
The model is a multivariate binomial State-Space model but in this first attempt we use an approximate multivariate Gaussian State–Space model to obtain estimates of the parameters series.
Several data sets are used to illustrate the suitability of the new methodology to analyse fertility patterns.
The layout of the article is the following. In Section 2 some notation is defined. Section 3 is devoted to describe the logistic model that is illustrated with some examples in section 4. The State-Space formulation is described in Section 5 and a final example is analysed in section 6.

## 2.- DATA and NOTATION

For each fixed time t we consider a fertility table that represent a sample of women from a theoretical population. Women are cross-classified on the binary variable maternity between (yes, no) and on the variable age. The tables are obtained using estimates of the age-distribution of women and the observed number of births by age of the mother in the period [t, t+1). To make the presentation simpler we assume that age is grouped in 5 year intervals. In other cases a similar analysis could be done. Then, fertility tables we want to model will have the following aspect:

| Maternity | Woman age | | | | | | |
|---|---|---|---|---|---|---|---|
| | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | |
| Yes | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{14}$ | $n_{15}$ | $n_{16}$ | $n_{1+}$ |
| No | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{24}$ | $n_{25}$ | $n_{26}$ | $n_{2+}$ |
| | $n_{+1}$ | $n_{+2}$ | $n_{+3}$ | $n_{+4}$ | $n_{+5}$ | $n_{+6}$ | $n_{++}$ |

where,

$n_{1j}(t) = \textit{number of women}$ aged $15 + 5(\mathrm{j} - 1)$ to $15 + 5\mathrm{j}$ with births ocurring between $[\mathrm{t}, \mathrm{t} + 1)$ $\mathrm{j} = 1,..6$

$n_{2j}(t) = \textit{number of women}$ aged $15 + 5(\mathrm{j} - 1)$ to $15 + 5\mathrm{j}$ without births between $[\mathrm{t}, \mathrm{t} + 1)$ $\mathrm{j} = 1,..6$

A model for the hypothetical population can be formulated :

- The non-conditional probabilities :

$\pi_{1j}(t) = p(\textit{birth}$ occurs between $[\mathrm{t}, \mathrm{t} + 1)$ and the *woman* is aged $15 + 5(\mathrm{j} - 1)$ to $15 + 5\mathrm{j})$ $\mathrm{j} = 1,..6$

$\pi_{2j}(t) = p(\textit{no birth}$ occurs between $[\mathrm{t}, \mathrm{t} + 1)$ and the *woman* is aged $15 + 5(\mathrm{j} - 1)$ to $15 + 5\mathrm{j})$ $\mathrm{j} = 1,..6$

- the conditional probabilities, ( which are estimated using the period fertility rates)

$m_j(t) = p(\textit{ birth}$ occurs between $[\mathrm{t}, \mathrm{t} + 1)$ given that *woman* is aged $15 + 5(\mathrm{j} - 1)$ to $15 + 5\mathrm{j})$ $\mathrm{j} = 1,..6$

then, taking into account that $n_{+j}\ (t)$ is fixed by design,

$$m_j(t) = \frac{\pi_{1j}(t)}{n_{+j}(t) / n_{++}(t)} \quad \mathrm{j} = 1,...,6$$

- The number of women in the sample, aged 15+5(j-1) to 15+5j with birth in [t,t+1) will follow a binomial distribution: $n_{1j}(t) \to B(n_{1+}(t), m_j(t))$
- And the Odds Ratio: $\theta_j(t) = \dfrac{m_j(t)}{1 - m_j(t)}$ j=1,..6

We will also use the following vectors:

Counts: $\mathbf{N_{1t}} = (n_{11}(t),...n_{16}(t))'$

Totals: $\mathbf{N_t} = (n_{+1}(t),...n_{+6}(t))'$

Probabilities: $\mathbf{\Pi_t} = (m_1(t),..., m_6(t))'$

Odds Ratio: $\boldsymbol{\theta}_\mathbf{t} = \boldsymbol{\theta}(t) = (\theta_1(t),..., \theta_6(t))'$

Observed log Odds Ratio: $\mathbf{Y_t} = \left( \log\left( \dfrac{n_{11}(t)}{n_{21}(t)} \right),..., \log\left( \dfrac{n_{16}(t)}{n_{26}(t)} \right) \right)'$

In some applications fertility analysis begins by defining subgroups of the female population according to parity. The parity specific fertility tables are analysed separately. In these cases a sub index for parity could be added to the quantities defined above.The properties of models are equal for parity grouped data or specific parity data.

## 3.- LOGISTIC MODEL FOR EACH t

(1) $\log(\boldsymbol{\theta}(t)) = \beta_0(t) + \mathbf{A_1}\beta_1(t) + \mathbf{A_2}\beta_2(t) = \mathbf{A}\cdot\boldsymbol{\beta}_\mathbf{t}$

where, $\boldsymbol{\beta}_\mathbf{t} = (\beta_0(t), \beta_1(t), \beta_2(t))'$ is a vector of three parameters and $\mathbf{A} = [\mathbf{A}_0, \mathbf{A}_1, \mathbf{A}_2]$ is a 6x3 matrix, with column-vectors $\mathbf{A_0}$ , $\mathbf{A_1}$ and $\mathbf{A_2}$. These column vectors are defined as follows:

$\mathbf{A_0} = (a_{j0})_{j=1,\ldots,6}$ ; $a_{j0} = 1$; j=1,...,6
$\mathbf{A_1} = (a_{j1})_{j=1,\ldots,6}$ ; $a_{j1} = (j- (6+1)/2)^2 -(6^2-1)/12$; j=1,...,6
$\mathbf{A_2} = (a_{j2})_{j=1,\ldots,6}$ ; $a_{j2} = j-(6+1)/2$; j=1,...,6

Consider also the alternative formulation:
Observed Odds**:** $\theta_j(t) = Q\ (t)T_j(t)$ Where

$Q\ (t) = \exp(\beta_o(t))$ and $T_j(t) = \exp\left(A_{j1}\beta_{j1}(t) + A_{j2}\beta_{j2}(t)\right)$

Two theoretical results illustrate the suitability of the Logistic model to describe fertility patterns and examples in section 4 illustrate the goodness of fit of the model with real data.

*RESULT 1*: Model (1) corresponds to a loglinear model with interaction terms that reflect the ‘quadratic trend’ in fertility schedules.

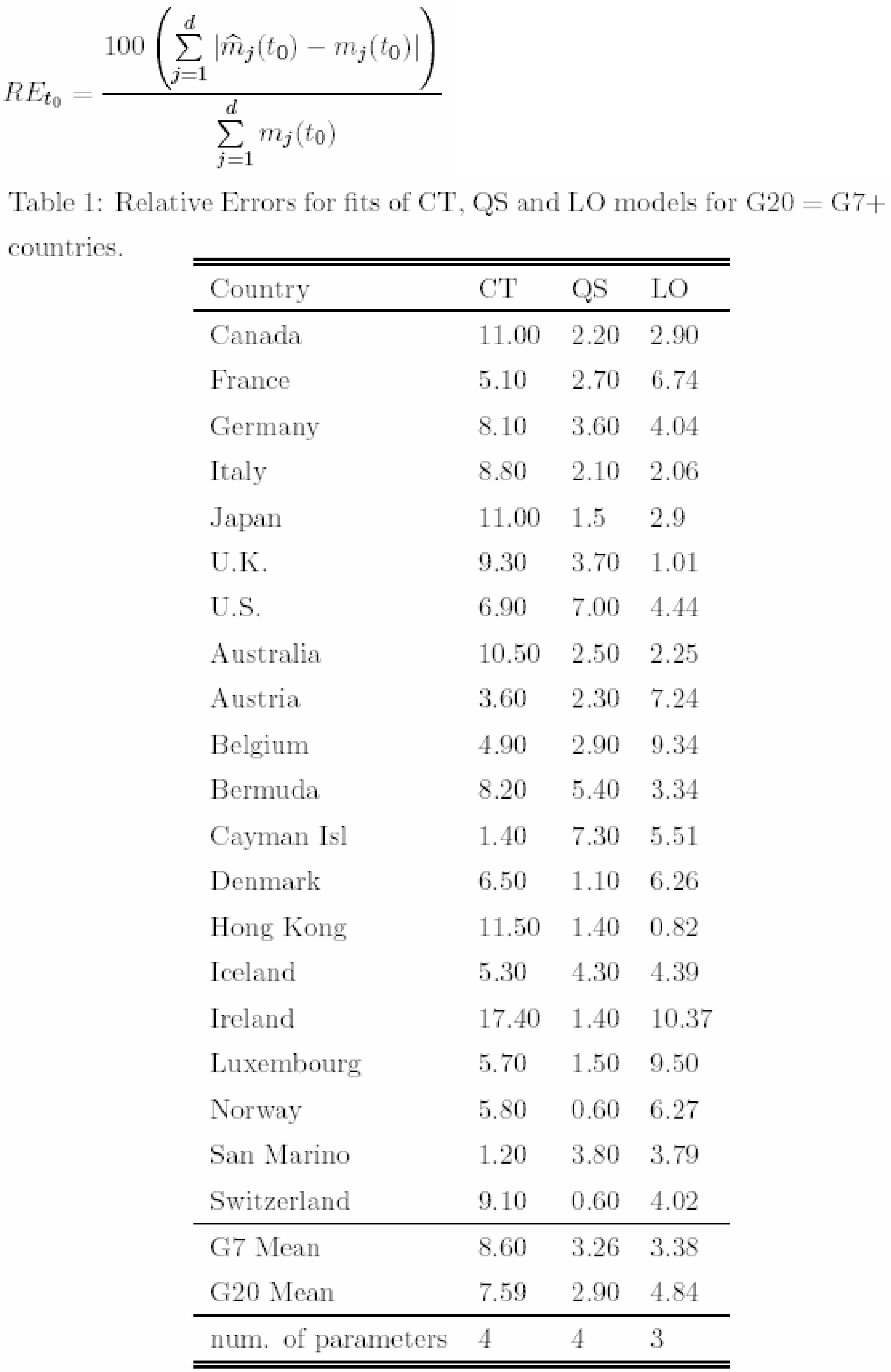
*RESULT 2*: $Q\ (t)$ is a measure of the period fertility level and $T_j(t)$ is a measure of the timing of fertility because in periods where only the level of fertility changes and no anticipation or postponement is observed $T_j(t)$ remains constant and in periods where only changes in the age distributions are observed both $T_j(t)$ and $Q\ (t)$ remain constant.

## 4.- LOGISTIC MODEL EXAMPLES

Two data set are analysed in this section, example 1 using contemporary data from developed countries and example 2 using historic series from Sweden.

### 4.1- Example 1: Internacional data, comparision with Coale-Trussell and Spline based models

In order to evaluate the suitability of the logistic model (LO) to estimate fertility schedules in real cases we consider a set of fertility rate data for five-year age groups from developed countries. We compare, in table 1, the fit of the LO model with that of the Coale-Trussell (CT) model and the Quadratic Spline (QS) model of Schmertmann (2003).
CT and QS models use four parameters and LO model uses only three. The data set is the same as that used in Schmertmann (2003) and corresponds in the majority of countries to the fertility schedules from 2001. The relative error is used as a measure of goodness of fit.



$$RE_{t_0} = \frac{100\left(\sum_{j=1}^{d} |\widehat{m}_j(t_0) - m_j(t_0)|\right)}{\sum_{j=1}^{d} m_j(t_0)}$$

Table 1: Relative Errors for fits of CT, QS and LO models for G20 = G7+ countries.

| Country | CT | QS | LO |
|---|---|---|---|
| Canada | 11.00 | 2.20 | 2.90 |
| France | 5.10 | 2.70 | 6.74 |
| Germany | 8.10 | 3.60 | 4.04 |
| Italy | 8.80 | 2.10 | 2.06 |
| Japan | 11.00 | 1.5 | 2.9 |
| U.K. | 9.30 | 3.70 | 1.01 |
| U.S. | 6.90 | 7.00 | 4.44 |
| Australia | 10.50 | 2.50 | 2.25 |
| Austria | 3.60 | 2.30 | 7.24 |
| Belgium | 4.90 | 2.90 | 9.34 |
| Bermuda | 8.20 | 5.40 | 3.34 |
| Cayman Isl | 1.40 | 7.30 | 5.51 |
| Denmark | 6.50 | 1.10 | 6.26 |
| Hong Kong | 11.50 | 1.40 | 0.82 |
| Iceland | 5.30 | 4.30 | 4.39 |
| Ireland | 17.40 | 1.40 | 10.37 |
| Luxembourg | 5.70 | 1.50 | 9.50 |
| Norway | 5.80 | 0.60 | 6.27 |
| San Marino | 1.20 | 3.80 | 3.79 |
| Switzerland | 9.10 | 0.60 | 4.02 |
| G7 Mean | 8.60 | 3.26 | 3.38 |
| G20 Mean | 7.59 | 2.90 | 4.84 |
| num. of parameters | 4 | 4 | 3 |

### 4.2.- Example 2: Swedish parity specific data.

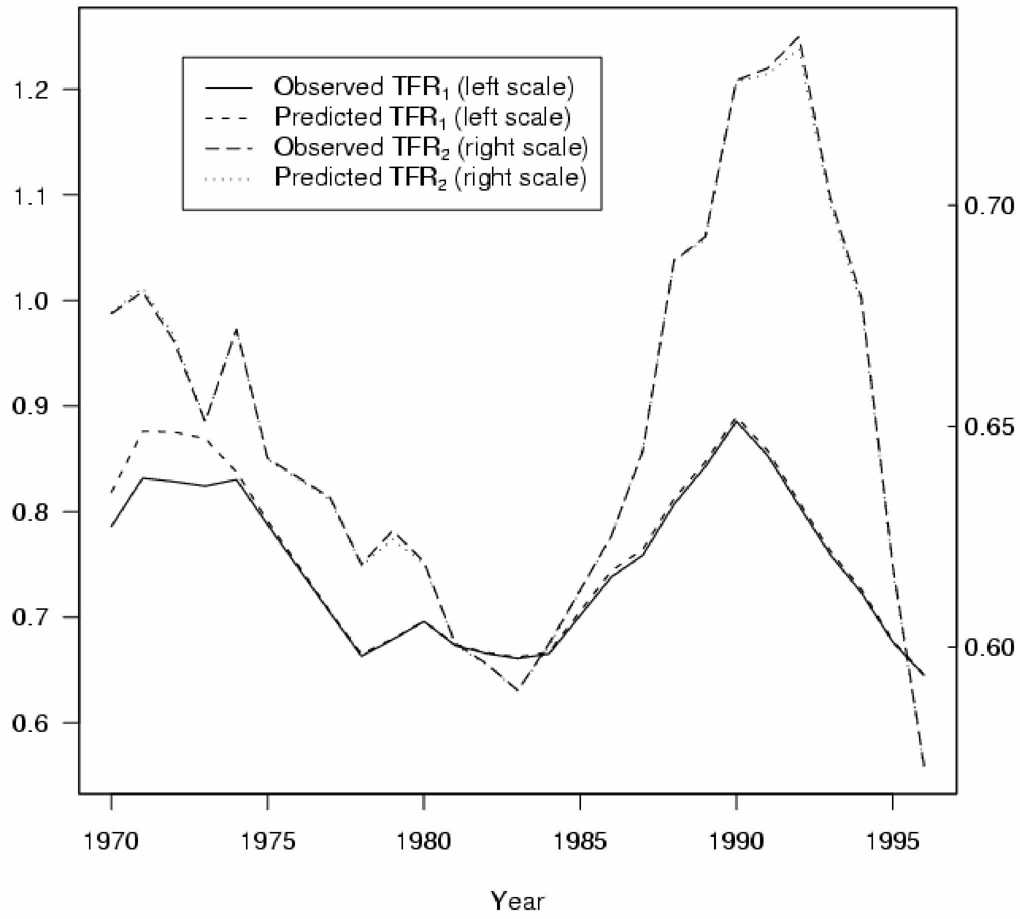
We have used the SAS package to adjust LO models.

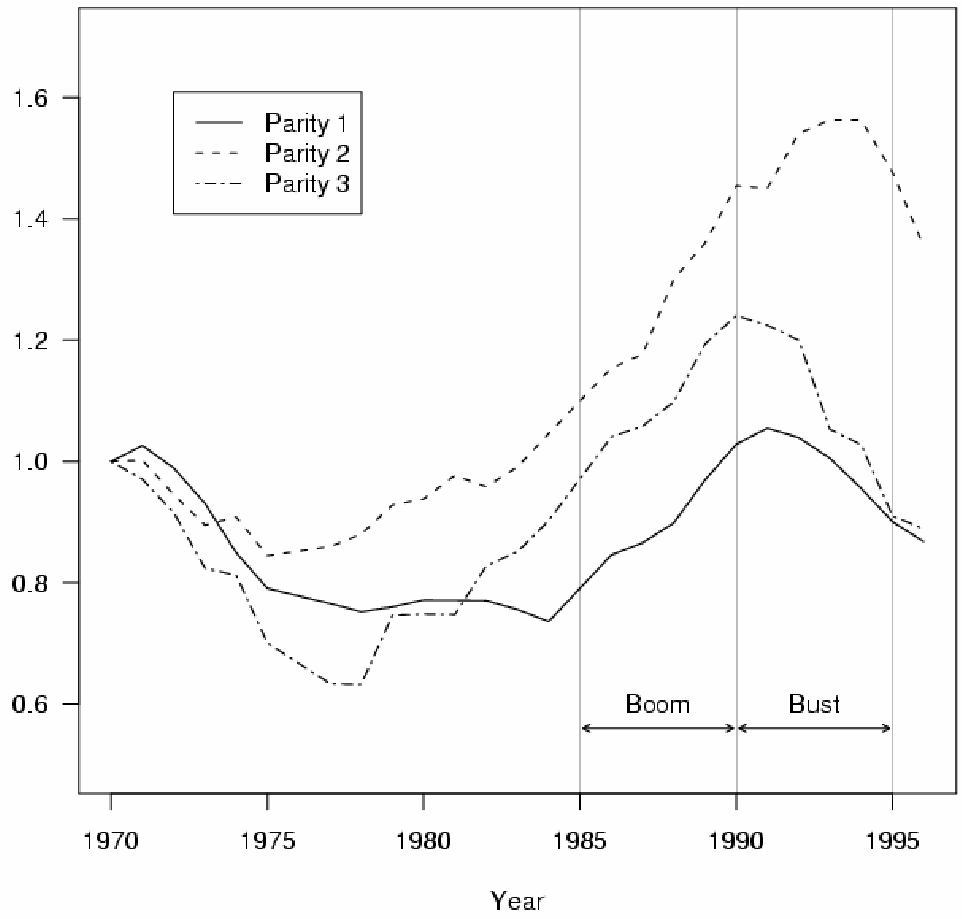
The data used are the tables that cross classify women by births and age. These tables are obtained using data from each parity. This data set has been also analysed by Kohler and Ortega (2002).

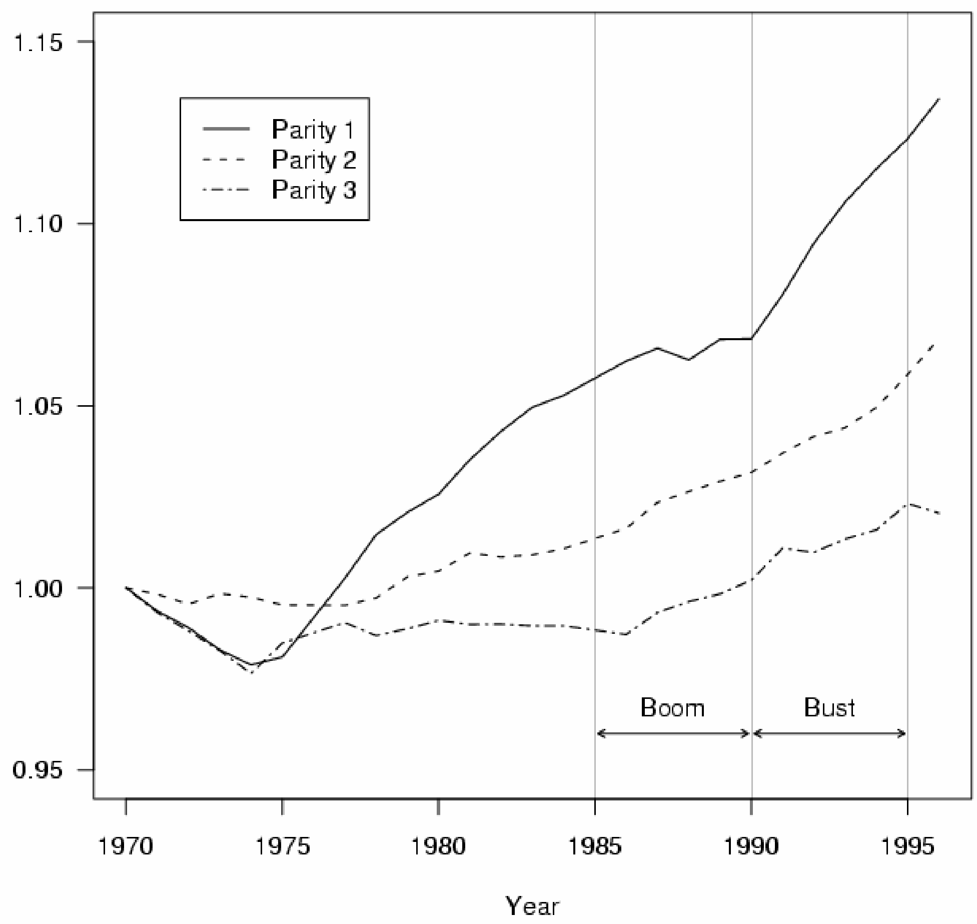
The three figures below illustrate the performance of the model for these data. Figure 1 shows Swedish fertility levels, measured using the TFR, for parities 1 and 2 from year 1970 to 1996. For each parity two series are pictured, the observed TFR and the predicted TFR using the LO model applied to the observed odds. Both cases illustrate the high quality fit of the model in terms of the TFR.

Figure 2 shows the relative evolution of the quantum component, Q(t), for the first three parities using 1970 as reference year. The quotient Q(t)/Q(1970) has been calculated and plotted for each year t in order to compare the relative changes between parities. The general trends are similar in the three series with some differences. An initial decreasing period is observed, followed by a period, of different lengths depending on parities, where there are no important relative changes. Then an increasing period initiated at different moments in time but shared during the baby boom. Finally a decreasing period initiated at the beginning of the baby bust (1995), except in the case of parity 2 which shows a delay of four years. One of the main conclusions that can be drawn from the plot is that the relative changes in the quantum for parity 1 are lower than those for parities 2 and 3. While the maximum change for parity 1 is around 32% (0.73 in 1984, 1.05 in 1991), the variations in parity 2, 72% (0.84 in 1975, 1.56 in 1994), or in parity 3, 61% (0.63 in 1977, 1.24 in 1990), are considerably greater.

To illustrate the relative changes in tempo, the series $\frac{\exp(\beta_2(t))}{\exp(\beta_2(1970))}$ is plotted in figure 3 for parities 1, 2 and 3. In this case the relative changes in tempo are clearly greater for parity 1 than for parity 2 and 3. Another interesting feature can be observed in this plot. While in the baby boom the relative changes are similar in all parities, during the baby bust the relative changes in parity 1, 6% (1.06 in 1990, 1.12 in 1995), are three times those in parity 2 and 3.


Observed TFR$_1$ (left scale)
Predicted TFR$_1$ (left scale)
Observed TFR$_2$ (right scale)
Predicted TFR$_2$ (right scale)
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.70
0.65
0.60
1970
1975
1980
1985
1990
1995
Year


Parity 1
Parity 2
Parity 3
1.6
1.4
1.2
1.0
0.8
0.6
Boom
Bust
1970
1975
1980
1985
1990
1995
Year



## 5.-SMOTHING PARAMETERS OVER TIME: STATE-SPACE MODELS

The book by Durbin and Koopman (2001) is an important reference to understand this section. Using the vectorial notation the LO model can be easily formulated as a state-space model,

$$\text{BINOMIAL} \quad \begin{cases} \log \boldsymbol{\theta}_{\mathbf{t}} = \mathbf{B}\boldsymbol{\alpha}_{\mathbf{t}} \\ \boldsymbol{\alpha}_{\mathbf{t+1}} = \mathbf{R}\boldsymbol{\alpha}_{\mathbf{t}} + \mathbf{v}\boldsymbol{\eta}_{\mathbf{t}} \qquad \boldsymbol{\eta}_{\mathbf{t}} \to N_{3p}(\mathbf{0},\mathbf{Q}) \\ p(\mathbf{N}_{\mathbf{1t}} / \boldsymbol{\theta}_{\mathbf{t}}) \to B_6(\mathbf{N}_{\mathbf{t}}, \boldsymbol{\Pi}_{\mathbf{t}}) \end{cases}$$

Where the vector $\boldsymbol{\alpha}_t$ is defined in a general form by:

$$\boldsymbol{\alpha}_{\mathbf{t}} = \left(\beta_0(t), \beta_0(t-1), ...\beta_0(t-p), \quad \beta_1(t), \quad \beta_1(t-1), ...\beta_1(t-p), \beta_2(t), \quad \beta_2(t-1), ...\beta_2(t-p) \qquad \right)'$$

the matrix **B** is obtained from the relationship: $\mathbf{A}\boldsymbol{\beta}_{\mathbf{t}}\mathbf{=}\mathbf{B}\boldsymbol{\alpha}_{\mathbf{t}}$

$B_6(\mathbf{N_t}, \mathbf{\Pi_t})$ Means a 6x1 vector whose components are independent binomial distributions. Here, $\mathbf{N_t}$ and $\mathbf{\Pi_t}$ are 6x1 vectors. Each component of these vectors has the parameters of the corresponding binomial distribution.

We are now working with the software to analyse observations from a multivariate non Gaussian State-Space model but at the moment it is not ready. An approximated Gaussian model can be formulated and the analysis can be obtained from existing software.

GAUSSIAN
$$\begin{cases} \mathbf{Y_t} = \mathbf{B}\ \boldsymbol{\alpha}_\mathbf{t} + \boldsymbol{\varepsilon}_\mathbf{t} & \boldsymbol{\varepsilon}_\mathbf{t} \to N(\mathbf{0}, \mathbf{H}) \\ \boldsymbol{\alpha}_\mathbf{t+1} = \mathbf{R}\boldsymbol{\alpha}_\mathbf{t} + \mathbf{v}\boldsymbol{\eta}_\mathbf{t} & \boldsymbol{\eta}_\mathbf{t} \to N(\mathbf{0}, \mathbf{Q}) \\ p(\mathbf{Y_t} \ / \log \boldsymbol{\theta}_\mathbf{t}) \to N_6(\log \boldsymbol{\theta}_\mathbf{t}, \mathbf{H}) & \end{cases}$$

## 6. – A STATE-SPACE MODEL EXAMPLE

To analyse the data in this section we have use the package SsfPack in Ox computing environment ( Koopman et al (1999)).

In this example we consider fertility tables from Spain during the period 1971-2001.
We consider a Gaussian State-Space model as defined in section 5:

$$\begin{cases} \mathbf{Y_t} = \mathbf{B}\ \boldsymbol{\alpha}_\mathbf{t} + \boldsymbol{\varepsilon}_\mathbf{t} & \boldsymbol{\varepsilon}_\mathbf{t} \to N(\mathbf{0}, \mathbf{H}) \\ \boldsymbol{\alpha}_\mathbf{t+1} = \mathbf{R}\boldsymbol{\alpha}_\mathbf{t} + \mathbf{v}\boldsymbol{\eta}_\mathbf{t} & \boldsymbol{\eta}_\mathbf{t} \to N(\mathbf{0}, \mathbf{Q}) \\ p(\mathbf{Y_t} \ / \log \boldsymbol{\theta}_\mathbf{t}) \to N_6(\log \boldsymbol{\theta}_\mathbf{t}, \mathbf{H}) & \end{cases}$$

To guess the definitive value of the matrix **R** and the exact definition of $\boldsymbol{\alpha}_\mathbf{t}$ **w**e have fitted several ARIMA models to the univariate parameter series $\beta_0(t), \beta_1(t)$ and $\beta_2(t)$, which are given, for each t, from the logistic model fit. The best ARIMA models for the state-space fit are:

$\beta_0(t)$ ARIMA(1,1,0)

$\beta_1(t)$ ARIMA(0,1,0)

$\beta_2(t)$ ARIMA(1,1,0) and then

$\boldsymbol{\alpha}_\mathbf{t} = \left(\beta_0(t), \quad \beta_1(t), \quad \beta_2(t), \quad \beta_0(t-1), \quad \beta_2(t-1)\right)'$

After trying the Gaussian State Space formulation with Spanish data we have observed that the autoregressive coefficient of $\beta_0(t)$ could have the same value as the autoregressive coefficient of $\beta_2(t)$ (denoted as $\phi$ below). Then, the **R** matrix we are using in Gaussian State Space formulation (and the one we propose initially in Binomial SS) is:

$$\mathbf{R} = \begin{pmatrix} 1+\phi & 0 & 0 & -\phi & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1+\phi & 0 & -\phi \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Figures below illustrate the result. Figures 1 and 2 compares the series for $\beta_i(t)$ where the values for years 2002-2007 are forecasted. Figures 3-6 compare the observed and adjusted values for specific age fertility 'rates' ($m_j(t)$ values).

Two approaches are being compared. In one side the State-Space model approach as it has been defined in the paper. On the other side a standard approach that fit a Box-Jenkins model separately for each series $\beta_i(t)$ obtained after adjusting the logistic model separately for each t. The corresponding ARIMA models obtained in the latter case are:

$\beta_0(t)$ ARIMA(2,1,0)

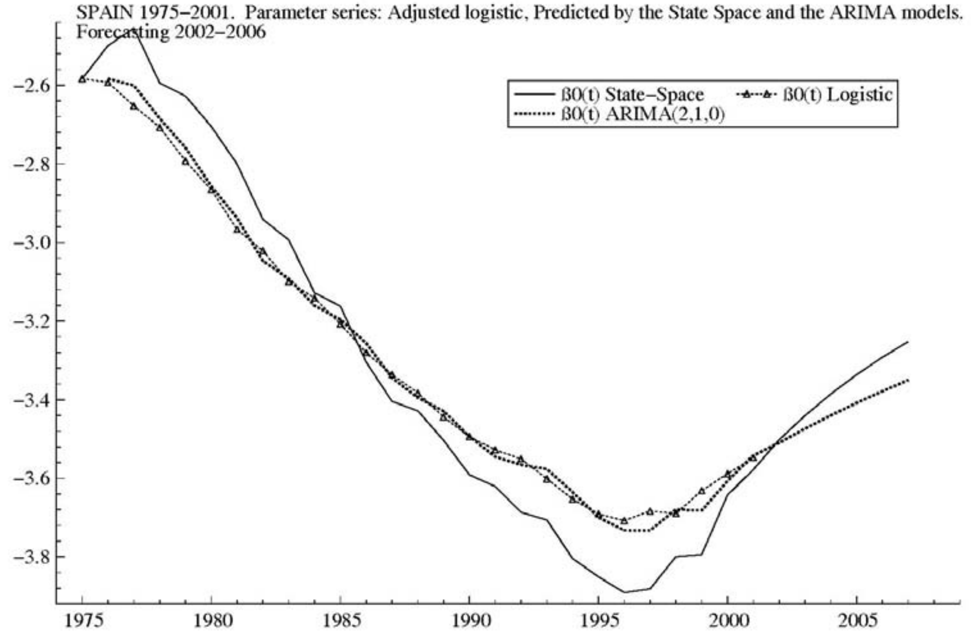
$\beta_1(t)$ ARIMA(1,1,0)

$\beta_2(t)$ ARIMA(1,1,0)

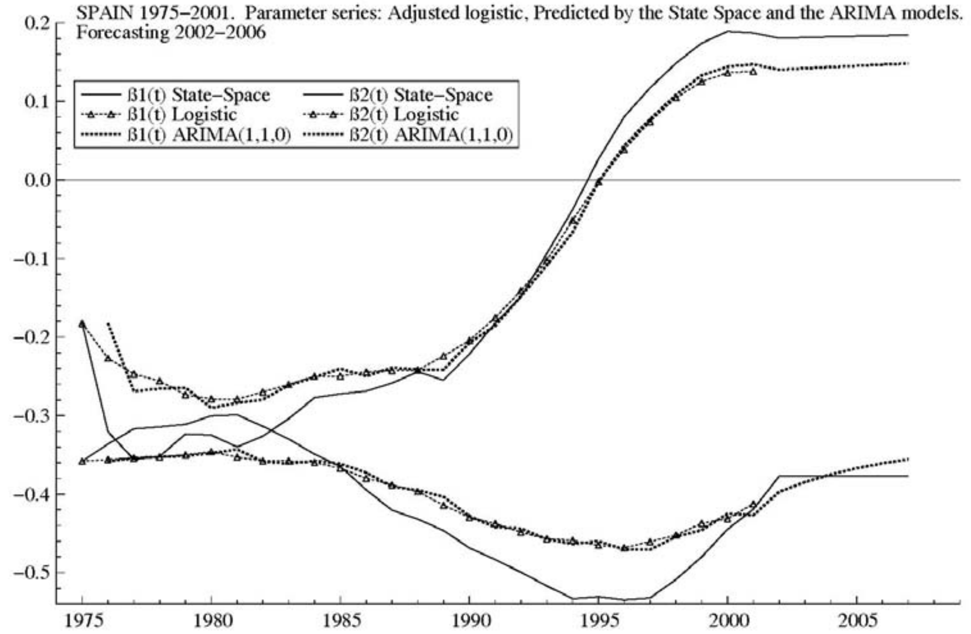
The most important primary conclusions are:

- The time series models for $\beta_i(t)$ are simultaneously adjusted using the State-Space approach, gives simpler models than the Box-Jenkins approach and comparable behaviour to model the response values.
- The procedure successfully model the trends displayed by the series of age-specific values and internally consistent forecasted values for the specific-age parameters are obtained.
- With the state-space procedure it is possible to obtain confidence band for forecasted values that take into account the different source of errors.

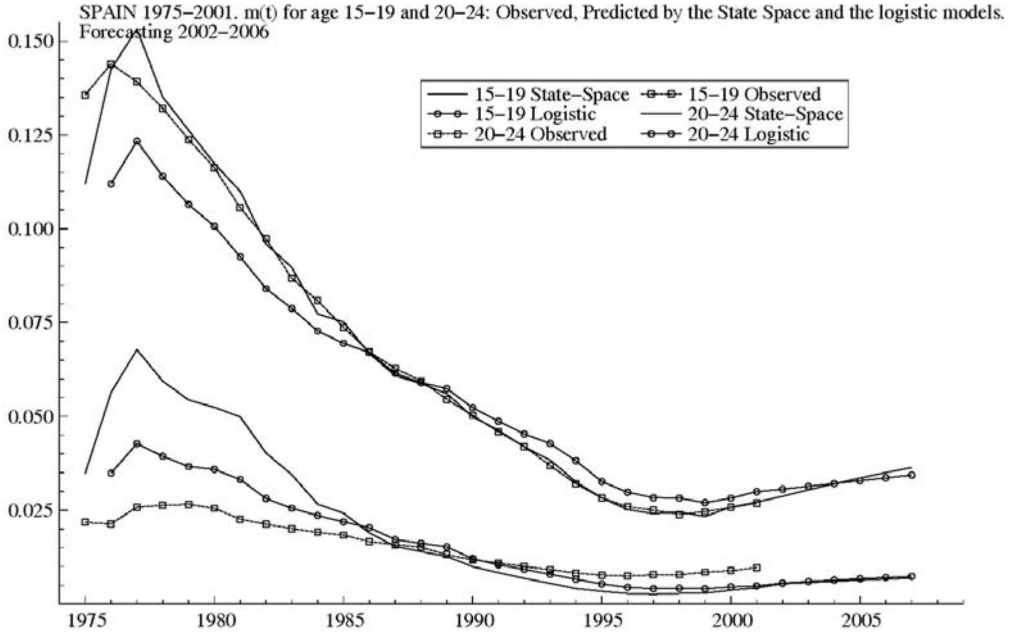
There is a lot of work ahead for getting definitive conclusions by comparing the new method with standard approaches to forecast fertility rates and by applying the methodology to data from other countries. Moreover some other aspect of the model should also be investigated in the future as the ability of the model to forecast response values under different scenarios and the flexibility of the model to incorporate external information. However in the light of these preliminary results we conjecture that the State-Space approach presented in this article succeed in simultaneously modelling the fertility age pattern and in smoothing the parameters over time

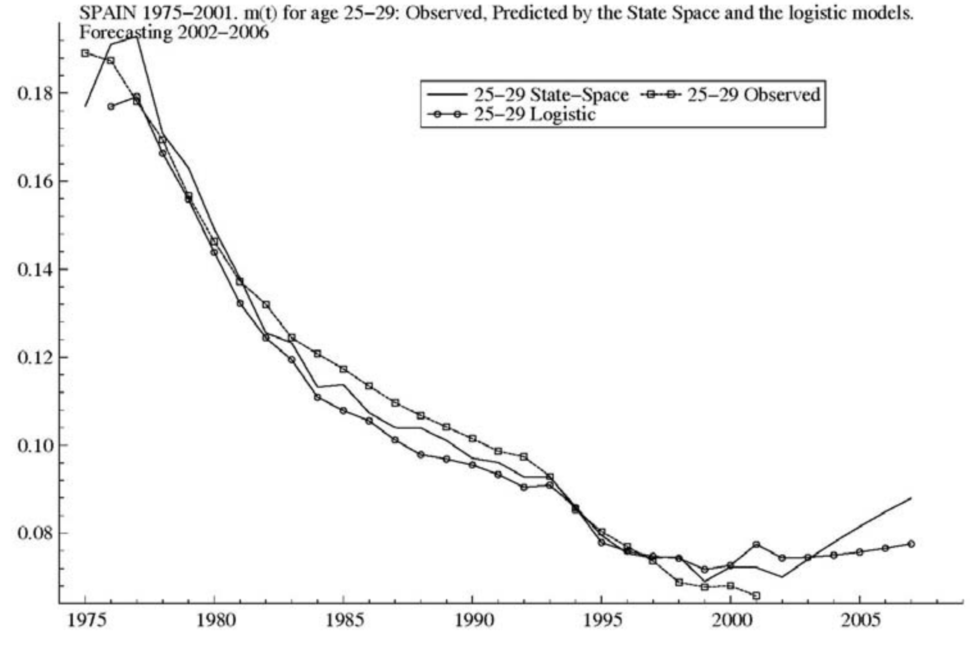


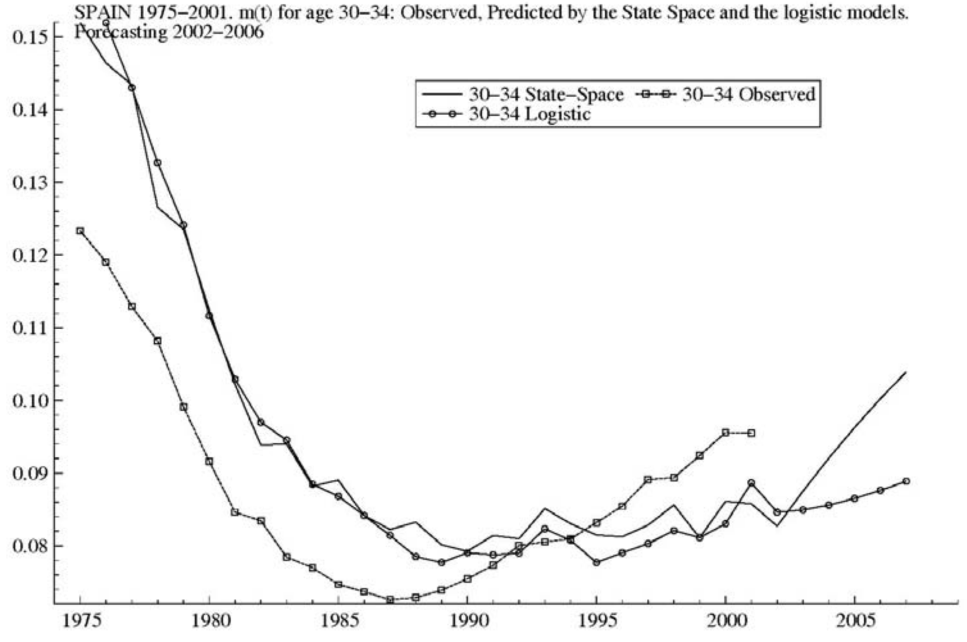
SPAIN 1975–2001. Parameter series: Adjusted logistic, Predicted by the State Space and the ARIMA models. Forecasting 2002–2006

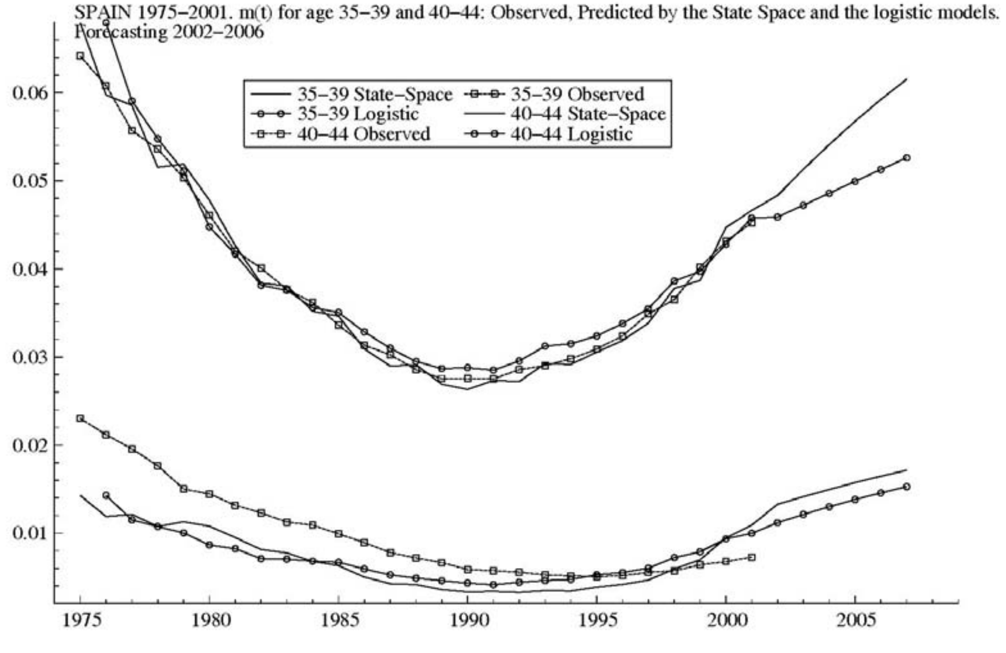


SPAIN 1975–2001. Parameter series: Adjusted logistic, Predicted by the State Space and the ARIMA models. Forecasting 2002–2006


SPAIN 1975–2001. m(t) for age 15–19 and 20–24: Observed, Predicted by the State Space and the logistic models.
Forecasting 2002–2006
15–19 State–Space
15–19 Observed
15–19 Logistic
20–24 State–Space
20–24 Observed
20–24 Logistic
0.150
0.125
0.100
0.075
0.050
0.025
1975
1980
1985
1990
1995
2000
2005


SPAIN 1975–2001. m(t) for age 25–29: Observed, Predicted by the State Space and the logistic models.
Forecasting 2002–2006
25–29 State–Space
25–29 Observed
25–29 Logistic
0.18
0.16
0.14
0.12
0.10
0.08
1975
1980
1985
1990
1995
2000
2005


SPAIN 1975–2001. m(t) for age 30–34: Observed, Predicted by the State Space and the logistic models.
Forecasting 2002–2006
30–34 State–Space
30–34 Observed
30–34 Logistic
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
1975
1980
1985
1990
1995
2000
2005


SPAIN 1975–2001. m(t) for age 35–39 and 40–44: Observed, Predicted by the State Space and the logistic models.
Forecasting 2002–2006
35–39 State–Space
35–39 Observed
35–39 Logistic
40–44 State–Space
40–44 Observed
40–44 Logistic
0.06
0.05
0.04
0.03
0.02
0.01
1975
1980
1985
1990
1995
2000
2005

## 7.- REFERENCES

Bongaarts,J. and Feeney, G. (1998). On the quantum and tempo of fertility. Population and Development Review. 24, pp 271-291.

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

Kohler, H,P. and Ortega, J,A. (2002). Tempo adjusted period parity progression measures, fertility postponement and complete cohort fertility. Demographic Research, 6, pp91-144.

Koopman, S. J., Shephard, N. and Doornik, J.A. (1999). Statistical algorithms for models in state space using SsfPack 2.2. Econometric journal, Vol 2 pp113-166.

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