Predictive Intervals for Age-Specific Fertility

NICO KEILMAN[1,2*] and DINH QUANG PHAM[1]
[1]*Statistics Norway;* [2]*Department of Economics, University of Oslo, P.O. Box 1095 Blindern, N-0317 Oslo, Norway (*author for correspondence, E-mail: nico.keilman@econ.uio.no)*

Keilman, Nico and Pham, Dinh Quang, 2000, Predictive Intervals for Age-Specific Fertility, European Journal of Population, **16**: 41–66.

Abstract. A multivariate ARIMA model is combined with a Gamma curve to predict confidence intervals for age-specific birth rates by one-year age groups. The method is applied to observed age-specific births in Norway between 1900 and 1995, and predictive intervals are computed for each year up to 2050. The predicted two-thirds confidence intervals for Total Fertility (TF) around 2010 agree well with TF errors in old population forecasts made by Statistics Norway. The method gives useful predictions for age-specific fertility up to the years 2020–2030. For later years, the intervals become too wide. Methods which do not take account of estimation errors in the ARIMA model coefficients underestimate the uncertainty for future TF values. The findings suggest that the margin between high and low fertility variants in official population forecasts for many Western countries are too narrow.

Keilman, Nico and Pham, Dinh Quang, 2000. Intervalles de confiance prévus pour des taux de fécondité par âge, Europe de l'Ouest, Revue Européenne de Démographie, **16**: 41–66.

Résumé. Un modèle multivarié de type ARIMA est combiné avec une courbe de type Gamma pour prévoir des intervalles de confiance pour les taux de fécondité par groupe d'âge annuel. La méthode est appliquée à des données norvégiennes sur les naissances par âge de 1900 à 1995 et les intervalles de confiance prévus sont calculés pour chaque année jusqu'en 2050. Les intervalles de confiance à 67%, pour la somme des naissances réduites, autour de l'année 2010, s'accordent bien avec les erreurs dans la somme des naissances réduites, utilisée dans les anciennes projections de population de l'Office Statistique de Norvège. La méthode donne des prévisions utiles pour les taux de fécondité par âge jusqu'aux années 2020–2030. Pour les années plus éloignées, les intervalles deviennent beaucoup trop larges. Les méthodes qui ne prennent pas en compte les erreurs d'estimation dans les coefficients du modèle ARIMA, sous-estiment l'incertitude des valeurs à venir de la somme des naissances réduites. Les résultats suggèrent que les marges entre les variantes haute et basse des projections de population officielles, faites dans de nombreux pays occidentaux, sont trop étroites.

1. The need for stochastic forecasts

The standard approach in population forecasting to deal with forecast uncertainty is to formulate two or more sets of assumptions for those key variables of which the future development is difficult to predict. Examples are Total Fertility and the Life Expectancy at Birth. This approach dates back to at least 1933, when Pascal Whelpton computed a population forecast for the United States, in which he presented several fertility variants. But the use of forecast variants is presumably much older: compare the intervals which Spengler (1935) reports for the results of a number of forecasts for the US. Nowadays, statistical agencies in 15 of the 18 member countries of the European Economic Area (EEA, that is the EU and EFTA countries, except Switzerland) produce forecasts with between two and four fertility variants (Eurostat, 1997a, Table 35). In the forecasts made in the beginning of the 1990s in those countries, the low and high fertility variants defined intervals of between 0.3 and 0.6 children per woman for a forecast duration of approximately 10 years. Moreover, the first co-ordinated population forecast for all 18 EEA countries has a difference between the high and the low fertility variant of 0.5 children per woman in the year 2035 (Eurostat, 1997b, Table 3).

This approach is unsatisfactory for at least two reasons. First, take the example of Norway's 1996-based population forecast, in which Statistics Norway (1997) assumes that the period Total Fertility in the year 2010 will be between 1.7 (low variant) and 2.1 (high variant) children per woman. No indication is given of the *probability* that the real TF in 2010 will have a value of 1.7–2.1, or perhaps fall outside that range. But for any educational planner with an interest in nursery and primary school, it must be of great importance to know whether the estimated probability of a TF between 1.7 and 2.1 in the year 2010 is 30, or 60, or perhaps even 90 per cent. In the former case, he should incorporate much more flexibility into the school planning process, than in the latter. Second, the variants do not consistently reflect uncertainty for forecast variables (Lee, 1999). For instance, the range for the Norwegian population aged 65+ in 2050 is from 0.97 to 1.36 million, or 34 per cent of the Medium Variant (1.16 million). For the population 20–64, the range is 28 per cent. Yet for the old-age dependency ratio, which is the age group 65+ divided by the age group 20–64, the range is only six per cent of the value in the Medium variant (0.42), surely many times smaller than one would expect.

Because the traditional approach is so unsatisfactory, some statistical agencies have in recent years attempted to compute stochastic population forecasts with predictive intervals. See for instance Hanika et al. (1997) and Lutz and Scherbov (1998a) for Austria, Lutz and Scherbov (1998b) for Germany, Alho (1998) for Finland, and Alders and De Beer (1998) for the Netherlands. These methods have been inspired by earlier work on stochastic forecasts by for instance Lee and Tuljapurkar (1994) and Alho (1990). The current paper is a contribution to the literature in this area. Its purpose is to generate predictive intervals for age-specific fertility rates. It

is part of a larger project of which the aim is to compute predictive intervals for the future population of Norway, broken down by age and sex.

2. The method – generalities

We assume that births are generated by a Poisson process. The intensity of that process depends strongly on the age of the mother. Thus the parameter of the Poisson process, that is, the fertility rate, varies by age.[1] For a certain year we assume that the age pattern of fertility follows a Gamma curve. This is a mathematical function which consists of a Gamma density and a scaling parameter. The Gamma curve has four parameters: the Total Fertility (TF), the Mean Age at Childbearing (MAC), the Variance in that age (VAR), and the minimum age. The four parameters are estimated on the basis of annual Norwegian data from 1900 to 1995. This results in a time series of parameter estimates. The series for three of the four parameters (TF, MAC, VAR) are modelled by means of a multivariate time series model. The minimum age is kept constant at its value as estimated for recent years. Predictions are made for the period 1995–2050 for all four parameters. The Gamma curve is used to transform the parameter predictions back into future age-specific fertility rates. Predictive intervals are obtained by means of stochastic simulation.

Since the ultimate purpose is to generate stochastic population forecasts, much attention is given to an appropriate quantification of uncertainty. In the present approach, there are four main sources of uncertainty attached to future birth rates:
 1. Sample variation in the historical age-specific birth rates
 2. Errors in the parameter estimates of the Gamma curve
 3. Residual variance in the time series model
 4. Errors in the parameter estimates of the time series model
Earlier studies of a similar nature, such as Bell (1992, 1997), De Beer (1989, 1992), Duchêne and Gillet-De Stefano (1974), Knudsen et al. (1993), and Thompson et al (1989) have ignored one or more of these sources.

3. The age-specific fertility rate: The Poisson model

Assume that a group of $Y(x)$ women aged x give birth to $B(x)$ children in a certain year. Assume further that the births are generated by a Poisson process. Maximum Likelihood estimators for the intensity of the process, and its corresponding variance are

$$F_x = B(x)/Y(x), \text{ and}$$
$$Var(F_x) = B(x)/Y^2(x) = F_x/Y(x) \tag{1}$$

The estimator F_x is the traditional age-specific birth rate. Its variance $Var(F_x)$ is small when the rate itself is small, or when the women at risk $Y(x)$ are numerous, or both.

We have used the Norwegian age-specific birth rates for the years 1900–1993 computed by Brunborg and Mamelund (1994). These were supplemented with rates for the years 1994 and 1995. The age range was from 16 to 44. When computing person years of exposure, we ignored mortality and international migration, and used the population of women at the beginning of each year, broken down in one-year age groups.

4. The age pattern of fertility: The Gamma curve

The array of 29 rates for each year can be summarized by means of a parametric curve, which is a function of age. Thanks to the regular shape of the age pattern of fertility, such a function contains only a few parameters, usually three to five. Various curves fit the data well: normal, lognormal, double exponential, Coale-Trussell, Hadwiger, polynomial, gamma, and logistic curves, to name the most important ones. Several authors have noted the attractive properties of the Gamma curve. The fit is usually good, and the parameters can be interpreted, after an appropriate transformation, in a straightforward way. See Bell (1997), Hoem et al. (1981), and Duchêne and Gillet-De Stefano (1974). Denote the fertility intensity for age x as f_x. The Gamma curve is usually defined as

$$f_x = \frac{1}{\Gamma(\alpha_3)} \alpha_1 \alpha_2{}^{\alpha_3} (x - \alpha_4)^{\alpha_3 - 1} \exp[-\alpha_2(x - \alpha_4)], \ x \geq \alpha_4. \tag{2}$$

The four parameters α_i are to be estimated from the data. $\Gamma(.)$ is the Gamma function defined by

$$\Gamma(p) = \int_0^\infty u^{p-1} \exp(-u)du.$$

α_1 represents the TF, whereas α_4 is the minimum age at childbearing. The parameters α_2 and α_3 have no immediate demographic interpretation. However, f_x/α_1 is the Gamma density, with mean $\alpha_4 + \alpha_3/\alpha_2$ and variance $\alpha_3/\alpha_2{}^2$. Therefore it is customary to introduce the following transformation

$$
\begin{aligned}
\beta_1 &= \alpha_1 \\
\beta_2 &= \alpha_4 + \alpha_3/\alpha_2 \\
\beta_3 &= \alpha_3/\alpha_2{}^2 \\
\beta_4 &= \alpha_4.
\end{aligned}
\tag{3}
$$

Hence parameters β_1 and β_4 have the same interpretation as α_1 (TF) and α_4 (minimum age). β_2 represents the mean age at childbearing, while β_3 is the variance in that age. The β-parameters have been estimated by means of non-linear regression, by minimizing the following weighted sum of squares

$$\Sigma_x w_x (F_x - f_x)^2. \tag{4}$$

F_x is the estimated rate for age x as introduced in the previous section, f_x is the underlying theoretical intensity given by expressions (2) and (3), while w_x is the inverse value of the variance of F_x, see expression (1). The latter variance reflects the "measurement error" for the rate F_x: a small variance indicates a precise estimate for the intensity, and vice versa.[2] Hence ages for which the variance is large get less weight in the regression than those with smaller variances. Weighted least squares estimation is approximately equivalent with Maximum Likelihood estimation of the parameters β_i, with births $B(x)$ and exposure time $Y(x)$ as data (Van Imhoff 1991). *Unweighted* least squares ($w_x = 1$) would imply that one regards the birth rates F_x as data. This would give relatively much weight to imprecise rate estimates. In a similar curve fitting exercise for US fertility, Bell (1992, p. 192) used weights equal to four for ages 18–32, and one for all other ages. His aim was to give more weight to ages with high fertility. Note the difference with our approach: we give more weight to the ages where the rate variance is low, usually ages up to 22 and beyond 33 for a country like Norway.

The four β-parameters and the corresponding covariance matrix, together with their variances were estimated on the basis of Norwegian birth rates for each of the years 1900–1995, see Section 3. We used a recent update of Van Imhoff's program Profile (Van Imhoff, 1991) for small scale experiments, and the NLIN-procedure in SAS for the complete data set. In the SAS computations we selected the Marquardt algorithm for minimization of expression (4), with appropriate non-negativity constraints for the parameters. The symbolic program MAPLE computed analytical first-order derivatives. For each year, the weights w_x are U-shaped. For instance, in 1995 the minimum is 237,500 at age 28 ($F_{28} = 0.141$; $Y(28) = 33,569$). The left and right branches of the w_x-curve increase rapidly to values exceeding 1,000,000 for ages below 20 or above 38. At ages 16 and 44, the weights are 25.8E6 and 14.3E6, respectively. The consequence of these extremely high values is that the fit is very bad around the top of the curve, since the curve is determined strongly by the estimated rates at young and old ages. In order to avoid this effect, we have censored extreme weights. After some experimentation we concluded that a maximum weight value of 3,000,000 gives a good fit around the top of the curve. Therefore, all weights (for all ages and calendar years) exceeding this value were made equal to 3,000,000. For 1995 this was the case for ages 16, 17, and 42–44.[3]

Figures 1 and 2 give estimates for β_1 and β_2, with corresponding 95-per cent confidence intervals. The TF and the mean age at childbearing computed in the traditional demographic manner (i.e. moment estimators TF $= \Sigma_x F_x$ and $m = \Sigma_x x.F_x/\text{TF}$) are also given. The latter estimators are only influenced by Poisson variability, not by the fit of the Gamma curve. Details on the variance β_3 are not given here – these are contained in Keilman and Pham (1998). For the minimum age β_4 we found that estimates in recent years were invariably equal to the boundary value of zero (after initial values around 14–15 years of age in the first half of the century).[4]
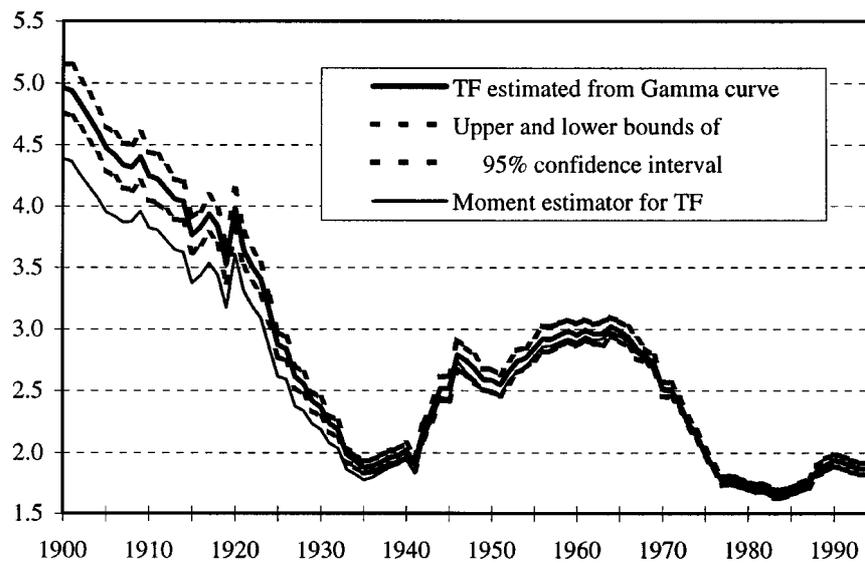
Figure 1. Total Fertility (TF) estimates and 95 per cent confidence interval.
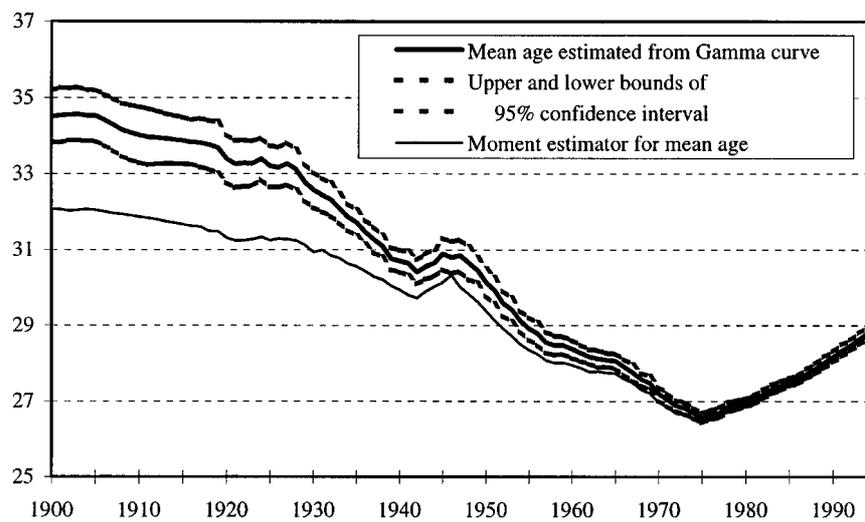

Figure 2. Estimates for mean age at childbearing and 95 per cent confidence interval.

As many other Western countries in this century, Norway had two periods with a strong fertility decrease (Figure 1). The first one, which started around 1880, ended in the 1930s, whereas the second one took place at the end of the 1960s and during the 1970s. The baby boom of the 1950s and 1960s was to a large extent the consequence of a decrease in the mean age at childbearing for women born in the years 1920–1945. This led not only to a fall in the period mean age (see Figure 2), but also to rather high period TF values (Figure 1). The period TF attained its
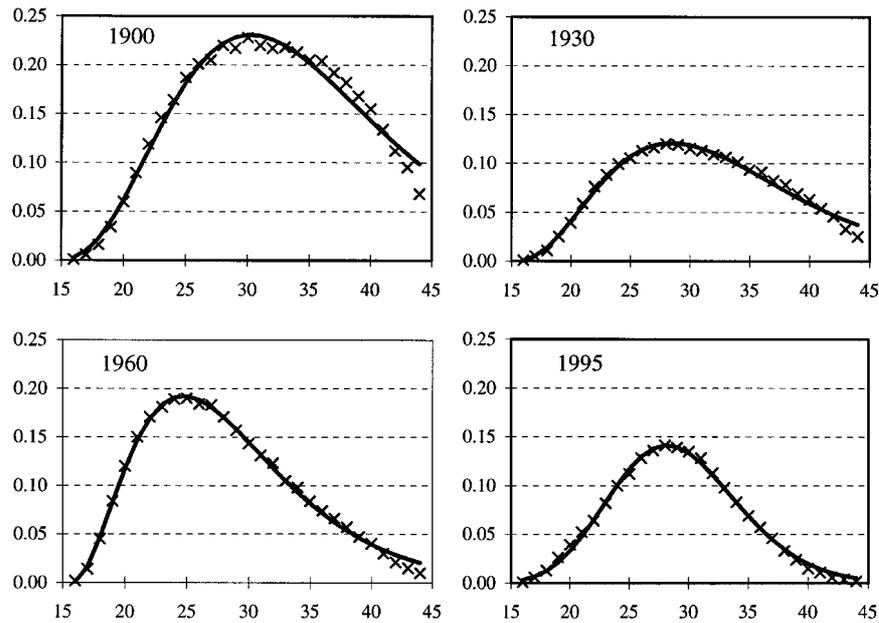
Figure 3. Age-specific birth rates, empirical values (xxx) and Gamma curve fit (solid line).

minimum in the years 1983 and 1984, when it was as low as 1.66 children per woman. After a rise towards 1.9 children per woman in 1990, the TF has been rather constant. But in recent decades, women get their children at increasingly higher ages as shown by the strong rise in the period mean age at childbearing in Figure 2. Facilitated by modern contraceptive methods, growing proportions of young Norwegian adults postponed the birth of their first child and took some form of education at the tertiary level during the 1970s and 1980s. Next they worked some years before they entered parenthood (Kravdal, 1994). Much of the fertility decrease during this century was caused by a reduction of higher-parity births, which generally take place at high ages.

The 95%-confidence intervals for the β_i are rather wide in the years 1900–25 and 1945–65, indicating a relatively bad fit. After 1980, the fit is excellent. The traditional TF (moment estimate) coincides with the estimates from the Gamma curve from 1970 onwards, and it falls within the 95 per cent confidence bounds from 1940. It is much lower in the first half of the century. However, one does not know whether the traditional TF or our Gamma-based TF are good approximations to the real (but unknown) Total Fertility, since both are only an *estimate* of the latter. Figure 3 illustrates how the fit improves over the years.

The smooth pattern of the empirical rates causes autocorrelation in the residuals in the age direction. This systematic bias can be seen very clearly for women older than 25 in the year 1900, but it is also present in the other three years. It means that the assumption of non-correlated residuals is not fulfilled. Thompson

et al. (1989) and Bell (1992, 1997) have attempted to account for the systematic deviations by including what they call a "bias adjustment" in predictions of the age-specific fertility rates. In Section 6 we will come back to this issue. Meanwhile we note that violating the assumption on zero correlation between residuals could have been avoided by using non-parametric estimation, although resulting residuals still would have been systematic.

5. A multivariate ARIMA model for the parameters of the Gamma curve

The result of the curve fitting exercise in the previous section is a series of estimates for the four parameters of the Gamma curve for each year between 1900 and 1995, and the corresponding estimated covariance matrix for each year. A multivariate ARIMA model has been used to predict three of the four parameters: β_1, β_2, and β_3. As noted earlier, the minimum age of childbearing β_4 fell from 14 in 1975 to zero in 1991, and remained at that level since. We predict that β_4 will be zero in the future, too.[5] We will briefly sketch the main features of the model. Further details are included in Keilman and Pham (1998).

Let $C_t = (\ln(\beta_{1,t}), \ln(\beta_{2,t}), \ln(\beta_{3,t}))'$ be a column vector with the Gamma curve parameters in year t in logarithmic form. First differences of C led to stationarity, and we found that a multivariate ARIMA (1,1,0) model fitted the data well. The model is of the form

$$Z_t = \phi Z_{t-1} + \epsilon_t,$$

where $Z_t = C_t - C_{t-1}$, ϕ is a fixed 3×3-matrix of coefficients, and $\epsilon_t = (\epsilon_{1,t}, \epsilon_{2,t}, \epsilon_{3,t})'$ is a multivariate normal column vector with mean 0 and constant covariance matrix Σ_ϵ. We deliberately omitted an intercept from the model, so that we avoid predicting an indefinitely increasing or decreasing pattern in β_t.[6] Estimation was done in such a way that years in which variances and covariances for the β-estimates were large, received less weight.

We have limited the time series analysis to the years 1945–1995. On the one hand, a long series is desirable on statistical grounds. On the other hand, there is little reason to believe that the childbearing behaviour of women in the first half of the century was so similar to that in more recent decades, that both can be captured by one model. Moreover, the fit of the Gamma curve was much better in the second half of this century than in the first half. However, we have also investigated the sensitivity of our predictions for choosing the shorter periods 1960–1995 and 1975–1995, see Section 7. Post-war effects in 1946 and 1947 have been removed: new estimates for β and the corresponding covariances in those years were computed by linear interpolation between 1945 and 1948.

Estimates for the elements of ϕ and corresponding standard errors are listed in Table I. Note that, because of our weighting scheme, these estimates should be interpreted as some averages over the years 1945–1995, with more emphasis on the second half than on the first half of the period.

Table I. Estimates of ϕ_{ij}

	$\hat{\phi}_{11}$	$\hat{\phi}_{12}$	$\hat{\phi}_{13}$	$\hat{\phi}_{21}$	$\hat{\phi}_{22}$	$\hat{\phi}_{23}$	$\hat{\phi}_{31}$	$\hat{\phi}_{32}$	$\hat{\phi}_{33}$
Estimate	0.6694	0	0	0	0.8852	0	0.0909	0	0.3089
Standard Error	0.1044	—	—	—	0.0735	—	0.0419	—	0.1337

Table II. Covariance estimates for non-zero elements of $\hat{\phi}^*$

	$\hat{\phi}_{11}$	$\hat{\phi}_{22}$	$\hat{\phi}_{31}$	$\hat{\phi}_{33}$
	$\times 10^{-3}$			
$\hat{\phi}_{11}$	10.185	0.011	–0.354	–0.000
$\hat{\phi}_{22}$	0.011	4.547	0.023	–0.130
$\hat{\phi}_{31}$	–0.354	0.023	1.644	–2.484
$\hat{\phi}_{33}$	–0.000	–0.130	–2.484	16.588

*Covariances between other elements of $\hat{\phi}$ are zero.

Diagonal elements are high and strongly significant. All but one ($\hat{\phi}_{31}$) of the off-diagonal elements turned out to be non-significant at the five per cent level in a trial calculation. These were set equal to zero, and the model was re-estimated with those restrictions, using Restricted Least Squares (Lütkepohl, 1993), for two reasons. First, we wanted to limit the number of estimated coefficients: the number of covariances, needed for the predictive intervals, grows quadratically with the number of estimated coefficients. Second, the predictions for β_2 based on the *unrestricted* model showed unrealistically wide 95 per cent confidence intervals, including mean ages well over 50 years already in the years 2020 and later (see Keilman and Phang 1998 for an explanation). By restricting non-significant elements of ϕ to zero, the confidence intervals for β_2, as well as those for β_1 and β_3, became realistic.

The estimated covariances for the non-zero-$\hat{\phi}_{ij}$-elements are given in Table II, whereas Table III contains estimates for the residual covariances Σ_ϵ. Much of the uncertainty, relatively speaking, concerns the TF, as witnessed by the high estimate of $\sigma_{\epsilon,11}$. This is explained by the large fluctuations in the TF since 1945 (see Figure 1), whereas those in the mean age (Figure 2) or in the variance (not shown here) were much smaller.

Predictions for the TF, the mean age, and the variance may be computed analytically, together with the confidence intervals for these three indicators. However, these intervals only reflect uncertainty around the predicted values provided that the matrix ϕ is known. But in practice, this matrix is estimated, and each element $\hat{\phi}_{ij}$ has its own distribution (except for those elements that were fixed to zero). Expressions for confidence intervals around future values of C_t that take the distribution

Table III. Estimates of $\Sigma_\epsilon = (\sigma_{\epsilon,ij})^*$

$\hat{\sigma}_{\epsilon,11}$	$\hat{\sigma}_{\epsilon,12}$	$\hat{\sigma}_{\epsilon,13}$	$\hat{\sigma}_{\epsilon,22}$	$\hat{\sigma}_{\epsilon,23}$	$\hat{\sigma}_{\epsilon,33}$
		$\times 10^{-3}$			
0.703	0.005	0.105	0.007	0.015	0.309

*Only the upper triangular part of the symmetric matrix is given.

of the ϕ-estimates into account are not known. Therefore we used simulation for the determination of predictive intervals of the elements of C_t, based on random draws of both the distribution for the ϕ-estimates *and* that for the residual vector ϵ. Predictive intervals based on the assumption that ϕ is known are reported in Section 7.1.

We simulated 10,000 sample paths for the vector β, each one from 1996 until 2050. For every sample path we drew one value for the matrix ϕ, and 55 values for the vector ϵ, one for each year. The estimates of both ϕ and ϵ follow a multivariate normal distribution. The mean and the covariance of $\hat{\phi}$ are given in Tables I and II. The mean of $\hat{\epsilon}$ is the null-vector, while its covariance estimates are contained in Table III. Multivariate normally distributed numbers were drawn from these two distributions using Cholesky decomposition of the covariance matrices (Bratley et al., 1983).

The predicted β-variables are unbounded from above. Therefore, draws that resulted in too large values for the three parameters were rejected. By the middle of the next century, the childbearing behaviour of Norwegian women may be very different from today's. Medical technology may have made it possible to postpone childbearing to ages well beyond 50. But even then, a TF of, say, 11 children per woman, or a mean age at childbearing of 55 years, or a variance in the age at childbearing of 400, are clearly unrealistic. At the same time it is unrealistic to assume that teenage fertility has become so important that the mean age falls below 20. Thus we used the following, quite liberal, restrictions: $0 < \text{TF} < 10$, $20 < \text{MAC} < 50$, and $0 < \text{VAR} < 250$. In addition, a restriction was imposed on the elements of the matrix ϕ: each of those was required to lie between minus one and one, in order to ensure stationarity. Almost 9 per cent of the simulations (967 of 10,967 runs) had to be rejected on the basis of the restrictions for the elements of ϕ, or for the three parameters. New sample paths were generated until we had obtained 10,000 paths with admissible values. This resulted in 10,000 values for each of the three β_i-parameters for every year between 1996 and 2050. These were ordered by size, and the lowest and highest 250, 1000, and 1666 values were taken as the lower and upper bounds of the 95, 80, and 67 per cent confidence intervals, respectively. Figures 4 and 5 illustrate the confidence bounds for the TF and the mean age. Information on the variance in the childbearing age is contained in Keilman and Pham (1998).
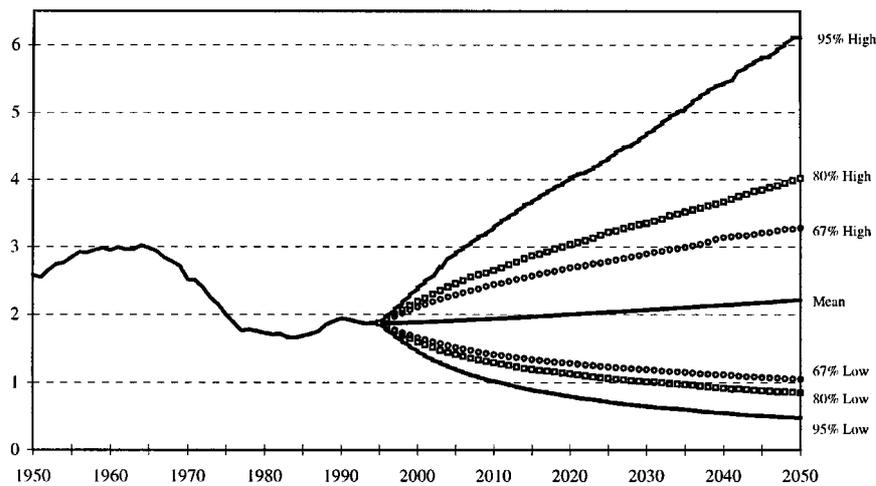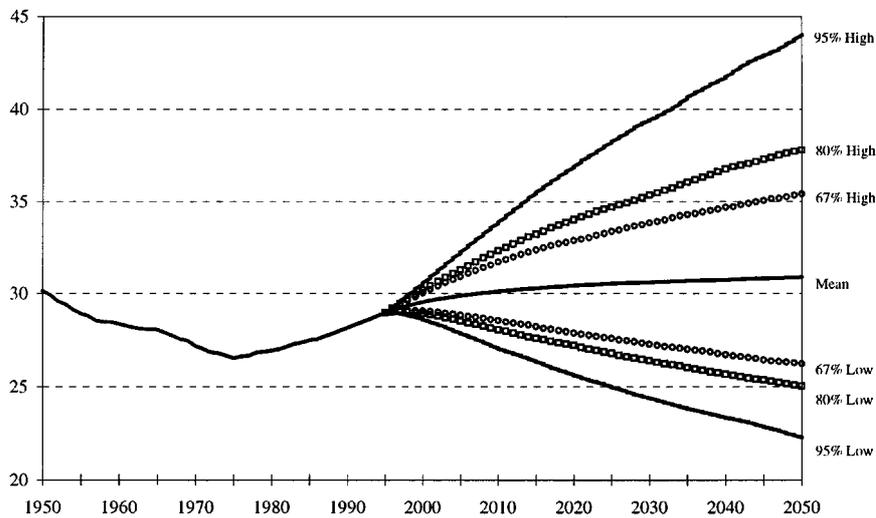
Figure 4. Total Fertility.

Figure 5. Mean age at childbearing.

The average values for the parameters in 2050 are 2.21 (TF) and 30.9 (mean age), while the medians are 1.86 children per woman and 30.2 years, respectively. The odds are two against one that the TF in 2050 will lie between 1.1 and 3.3 children per woman, while the 95 per cent confidence interval is (0.5, 6.1) in that year. These rather wide intervals do not come as a surprise: because the original time series was differenced, prediction errors cumulate over time, with no wearing off.

The expected probability that the TF will exceed 6.1 children per woman in 2050 is only 2.5 per cent. Yet it is difficult to imagine such extremely high levels of
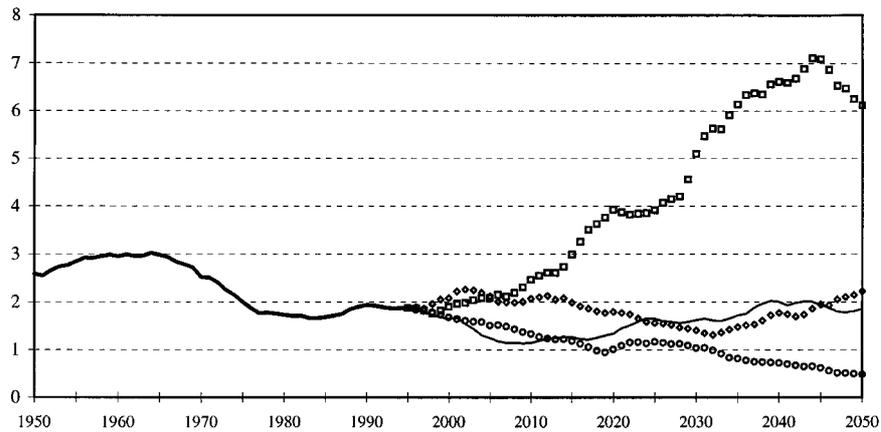
Figure 6. Four sample paths for Total Fertility.

childbearing in a country like Norway. The highest TF ever recorded for Norway is (the moment estimate of) 4.8 in 1879, the highest value since 1845, see Brunborg and Mamelund (1994). With a TF of 6.1 children per woman or more, fertility would exceed the historical maximum by more than one child, and it would exceed the current level in many less developed countries. Clearly, the model predictions for the middle of the next century cannot be considered as realistic. If we somewhat subjectively assume that a fertility level of more than four children per woman on the medium and long term should be rejected (even when the probability of such a level is only a few per cent), we see that the model gives reasonable results up to around 2020 or perhaps 2030. Beyond that, confidence intervals are too wide. One has to take recourse to other methods when predictions so far ahead are required. The easiest one is to assume that in 2030, say, uncertainty is already so large that it will not increase any more (e.g. Alho and Spencer, 1997). In that case confidence intervals are constant after 2030. A more sophisticated one is to assume that there is an upper bound to fertility levels in Norway. Considerations of this sort led Lee (1993) to use a logit transformation and to include an upper (and lower) bound in his univariate ARIMA-based TF-predictions for the United States, following a suggestion first made by Alho (1990). However, Alho and Spencer (1997) demonstrated later that when the logit-transformed β_t follows a random walk process, then its predictions will eventually be "absorbed" close to the upper and lower bound for large enough t. This anomaly also showed up in our case, an ARIMA (1,1,0)-model, see Keilman and Pham (1998) for details. The conclusion is that the logit transformation cannot be used for constraining confidence bounds. In Section 9 we describe an informal method for imposing pre-specified bounds upon the predicted values.

Figure 6 illustrates the volatility in the simulated TF. It shows four sample paths, namely the two paths that came closest to the upper and lower bounds of the 95 per

Table IV. Fertility assumptions in Statistics Norway's 1996-based population forecast

	Total Fertility (children per woman)			Mean age at childbearing (yrs)		
	Low variant	Medium variant	High variant	Low variant	Medium variant	High variant
1995*	1.86	1.86	1.86	28.79	28.79	28.79
2000	1.79	1.86	1.95	29.34	29.20	29.03
2010 and beyond	1.68	1.86	2.10	30.50	30.00	29.50

*Values computed from birth statistics.
Source: Statistics Norway 1997.

cent interval in 2050, and the paths which hit the average and the median TF-values in 2050.

We experimented also with 5,000 and 1,000 simulations, instead of 10,000, but in those cases the upper bounds of the 95 per cent intervals looked a bit ragged, in particular those for the TF.

The predictions in Figures 4 and 5 are to be compared with the fertility assumptions used by Statistics Norway in their 1996-based population forecast. In that forecast, the year 2010 was chosen as the so-called target year for fertility, that is, the year after which no change was assumed for fertility parameters. Table IV shows that the TF and the mean age in the Medium Variant agree very well with our median predictions. The constant TF is explained, at least qualitatively, by cohort developments. The cohort TF in Norway decreased rapidly for women born between 1935 (2.57 children per woman) and 1954 (2.05). After some fluctuations in the TF among women born in the second half of the 1950s, this indicator is assumed to fall slightly for women born after 1960, from 2.10 (generation 1960) to 1.97–2.05 children per woman (generation 1971). The decrease is the result of two opposite forces. On the one hand, childlessness increases slightly, from 12 per cent for women born in the 1950s, to an assumed 15–20 per cent for the generations born 30 years later. This presses the cohort TF downwards. At the same time there is an increased propensity among mothers with two children to have a third one. Therefore Statistics Norway decided to opt for a constant period TF in the Medium Variant of a little less than 1.9. But the upward slope in the mean age (26.8 years in 1980, and 28.8 in 1995) was extrapolated, so that a level of 30 years was reached in 2010.

The official forecast's High-Low gap in the TF in 2010 is 0.42 child per woman, and that for the mean age is 1 year. These intervals are rather narrow, compared to the confidence intervals in Figures 4 and 5. For instance, the two-thirds confidence interval for the TF in 2010 is 1.0 children per woman wide, and that for the mean age 3.4 years. Given the normal distribution in our model of log(TF) in 2010, we can conclude that the probability that the real TF in 2010 will lie between the values assumed in the Low and the High Variant of the 1996-based forecast is only 29 per

cent. For the mean age the expected probability is no more than 23 per cent. In the concluding section we shall comment these low probabilities.

6. Predictive intervals for age-specific fertility

Figure 7 gives confidence bounds for predicted age-specific fertility rates in the years 2010, 2030 and 2050. They are based on the 10,000 simulations described in the previous section, assuming a minimum age at childbearing equal to zero (compare footnote 5) and an age pattern that follows a Gamma curve. Clearly, the 95 per cent confidence interval is so wide in 2030, that it is not very informative. For the year 2050 it is useless to work with 95 per cent bounds. We have deliberately chosen not to censor rates beyond age 50. Recent medical advances have led to an increased demand for Assisted Reproductive Technology (such as *in vitro* fertilization) after age 30 in Western countries. We cannot exclude the possibility that in the middle of the next century childbearing will be an option for women older than 50. It has been suggested that frozen ova may be taken from the woman at age 22, say, and that these may be fertilized and implanted later (Beets, 1996). The predicted rates, however, are small. At the other end of the age scale, there are some very low rates at ages below 15. These would be ignored in a cohort-component projection.

The Gamma curve fit in Section 4 resulted in deviations for women above age 25. Since these deviations are systematic, they could be included in the predictions. Thompson et al. (1989) and Bell (1992, 1997) describe such a bias adjustment. They took the residuals from the Gamma curve in the last year of data, and extrapolated them forward as constant deviations of future age-specific rates from the forecasted Gamma curves. They report improved accuracy for the first few forecast years, but diminishing effects as the forecast horizon increased. The explanation is that the bias adjustments were dwarfed by the errors in predicting the Total Fertility, the mean age, and the variance. We have not applied any form of bias adjustment, for two reasons. First, the focus in our analysis is on long-term uncertainty, and second, the fit of the Gamma curve was relatively good in recent years.

7. Sensitivity analysis

The analysis in the previous sections differs from earlier studies in three respects. First, as Van Imhoff (1991) has noted, curve-fitting exercises almost invariably ignore the fact that a birth rate is not an observed quantity, but an estimate of the parameter of an underlying model. See, for example, Bell (1992, 1997), De Beer (1992), Duchêne and Gillet-De Stefano (1974), Knudsen et al. (1993), Thompson et al (1989), and the references they contain. (Note, however, Hoem (1976) and Hoem et al. (1981), who make a similar point, and establish an interesting link between weighted least squares estimation for the intensities F_x and minimum chi-square estimation for the counts of births B_x.) Second, ARIMA models for
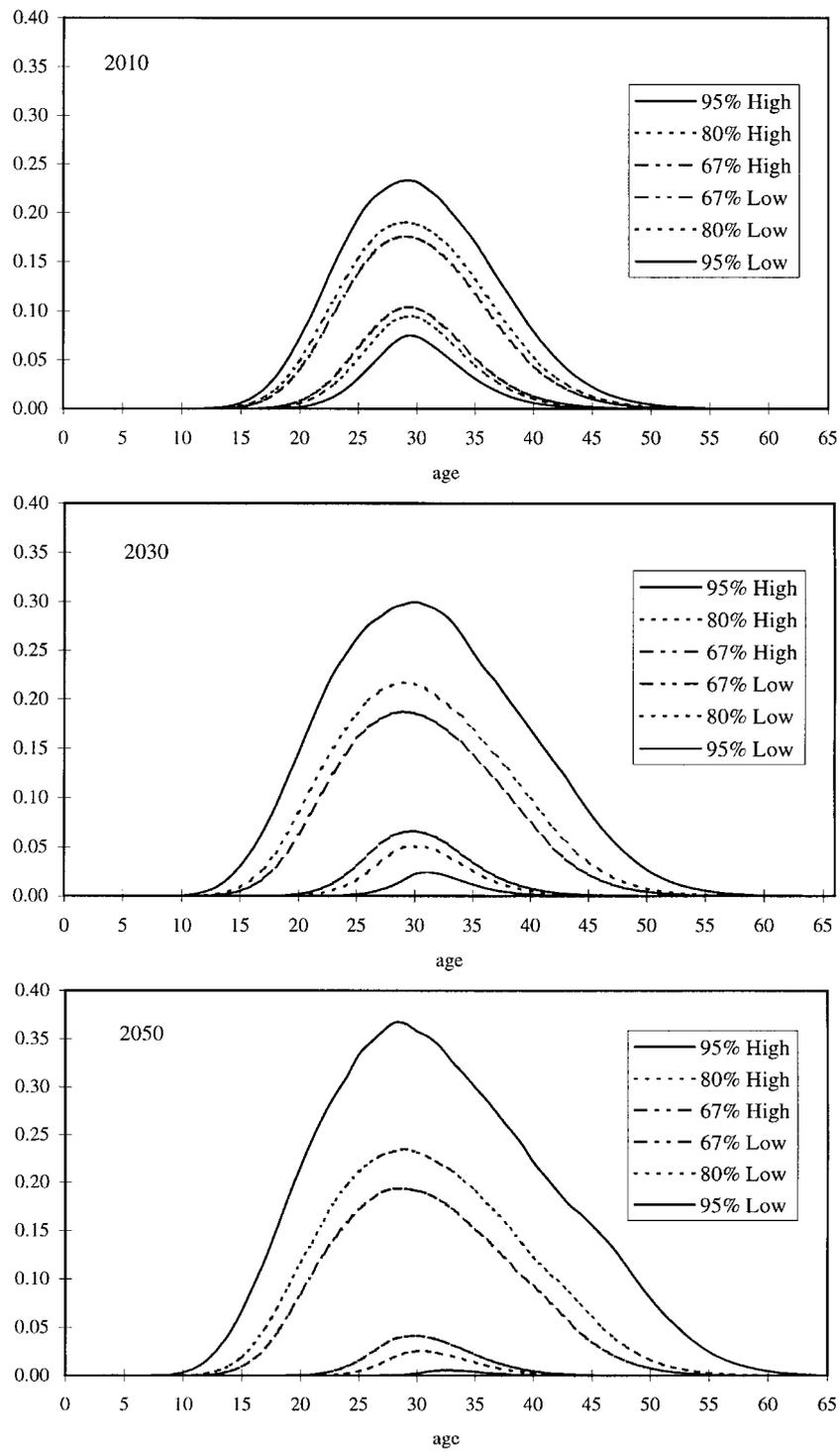
Figure 7. Predictive intervals for age-specific birth rates, 2010, 2030, 2050.

Table V. Predictions and predictive intervals for the Total Fertility (TF) and the mean age, assuming known ϕ-matrix*

	TF (children per woman)				Mean age (yrs)			
	1995	2010	2030	2050	1995	2010	2030	2050
Prediction	1.87	1.87	1.87	1.87	28.97	30.06	30.25	30.27
67% Low	—	1.42	1.19	1.05	—	28.50	27.13	26.10
67% High	—	2.46	2.94	3.33	—	31.70	33.74	35.10
95% Low	—	1.09	0.77	0.60	—	27.08	24.43	22.64
95% High	—	3.21	4.53	5.79	—	33.36	37.47	40.47

*Values for 1995 are Gamma curve estimates.

the TF or other summary indicators show the same defect, compare, for example, Knudsen et al. (1993), Bell (1997), and Lee (1993). Finally, predictions on the basis of such time-series models assume that the parameters of the model are given, whereas in reality these are only estimates, with corresponding confidence intervals. How serious are these omissions and assumptions? To what extent do they lead to smaller confidence intervals and predictive intervals? We have opted for an empirical analysis of this question, and compared the results reported in the previous sections with corresponding results based on traditional assumptions. In the next three sections we look at predictions with known ϕ-matrix, at unweighted Gamma curve estimates, and at unweighted ARIMA model estimates. The final Section 7.4 reports predictions by a model that was estimated for relatively short periods, i.e. 1960–1995 and 1975–1995. As before, we have focused on the TF and the mean age.

7.1. PREDICTIONS WITH KNOWN ϕ-MATRIX

When the ϕ-matrix is known, simulation is unnecessary, and predictions for the TF, the mean age, and the variance and corresponding confidence intervals can be computed analytically. Table V shows that the TF levels off to 1.87, and the mean age to 30.3 years in 2050, close to the medians in Figures 4 and 5. The 95 per cent predictive intervals in 2050 are somewhat narrower: by 0.5 child for the TF, and by 3.9 years for the mean age. Thus assuming the ϕ-matrix as given leads to 95 per cent bounds for the TF and the mean age in 2050 that are too narrow by 8 per cent and 18 per cent, respectively.

7.2. UNWEIGHTED GAMMA CURVE ESTIMATES

In the weighted case, one takes account of the fact that low birth rates at young and old ages have small variances, and thus these rates get more weight than high rates at intermediate ages. In the unweighted case, rate variances are ignored, and

Table VI. Comparison of parameter estimates in the unweighted and the weighted case. Years in which differences were largest or smallest

	Year	Unweighted estimate (1)	Weighted estimate (2)	Difference (1)–(2)	Difference (1)–(2) as a % of (2)
TF-largest	1952	2.73	2.66	0.07	2.6
TF-smallest	1985	1.69	1.68	0.01	0.4
MAC-largest	1956	29.24	28.79	0.45	1.6
MAC-smallest	1979	26.90	26.90	0.00	0.0

all weights w_x in expression (4) are chosen equal to one. Then it is assumed that birth rates are observed, instead of parameter estimates for the Poisson model of Section 3. The consequence is that the estimated TF becomes higher compared to the weighted case, since the fitted curve follows the high rates more closely. To what extent the mean age is influenced, is an empirical matter.

We fitted the Gamma curve using unweighted least squares to the data for the years 1945–1995, and found that estimates for all three parameters are higher than in the weighted case in almost every year. The difference is very small after 1970, but larger between 1945 and 1970, when the fit of the Gamma curve was somewhat less (see Section 4). Table VI compares weighted and unweighted parameter estimates in those years in which differences were smallest and largest. The results indicate that weighting has only had minor impact on the estimates for the TF and the mean age. However, the estimated parameter variances for the two parameters are relatively small in the unweighted case. This leads to narrow confidence intervals for many predicted age-specific rates (see for instance Figure 8 for the year 1995). During the ages of high childbearing, the confidence intervals around the predicted age-specific rates are nearly half as large in the unweighted case compared to the weighted case. For younger and older ages the differences are much smaller – at certain ages the confidence intervals become even *larger* when one does not weight.

7.3. UNWEIGHTED ARIMA-MODEL ESTIMATES

Most authors ignore the fact that not only the birth rates, but also summary parameters such as the Total Fertility or the mean age are estimates, each with their own variance. Table VII gives the predictive intervals for an unrestricted multivariate ARIMA (1,1,0) model that was estimated for the three log-transformed parameters of interest. These parameters were computed on the basis of unweighted birth rates, and the ARIMA-model was estimated giving equal weights to each year. We call this the unweighted ARIMA-model. The intervals are computed analytically,
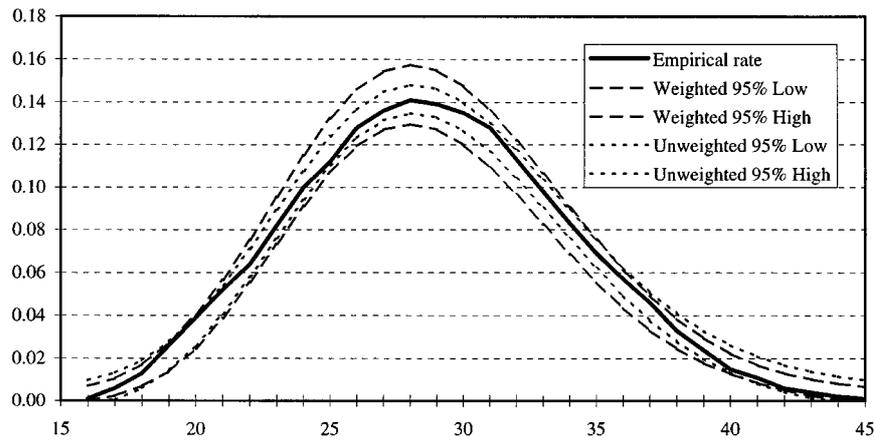
Figure 8. Age-specific birth rates, 1995.

Table VII. Predictions and predictive intervals for the Total Fertility (TF) and the mean age, unweighted ARIMA-model*

	TF (children per woman)				Mean age (yrs)			
	1995	2010	2030	2050	1995	2010	2030	2050
Prediction	1.87	1.88	1.88	1.88	28.97	30.86	31.30	31.36
67% Low	—	1.50	1.30	1.18	—	29.34	28.14	27.05
67% High	–	2.35	2.70	2.99	—	32.45	34.82	36.36
95% Low	—	1.21	0.92	0.75	—	27.96	25.40	23.47
95% High	—	2.91	3.84	4.68	—	34.05	37.58	41.90

*Values for 1995 are unweighted Gamma curve estimates.

assuming that the estimated ϕ-matrix is the real one. Therefore these intervals should be compared with those in Section 7.1.

Whereas the predicted TF in 2050 is almost the same as that in the weighted case, the mean age increases to a higher equilibrium level. This is explained by the high estimates for ϕ_{22} (0.986) that we obtained. The 95 per cent interval for the TF is much narrower than in the weighted case (by 1.7 child in 2050), and that for the mean age a little so (by almost 4 years).[7] The consequence of ignoring weights thus is that we are too optimistic about the future TF, in the sense that the predictive intervals are too narrow. If we would have ignored the sample variation in the historical age-specific birth rates in this particular empirical application, and also ignored the errors in the parameter estimates of the Gamma curve, we would have put too much confidence in subsequent births predictions.

7.4. ESTIMATION PERIODS 1960–1995 AND 1975–1995

Figures 1 and 2 show that Norwegian fertility was far from stable during the post-war period, when both the mean age and the TF showed major trend shifts. The model estimates in Tables I–III were obtained on the basis of data for the years 1945–1995, and hence those estimates as well as the predictions in Figures 4 and 5 reflect these rather turbulent years. We have re-estimated the model on the basis of two shorter periods, namely the years 1960–1995 and 1975–1995. Would a model estimated for these periods lead to different predictions? The predicted levels for the TF, the mean age, and the variance are probably hardly affected, since the ARIMA-model tends to pick up the trends for the most recent period. Whether the confidence intervals become wider or smaller is an empirical issue. On the one hand, fertility is less volatile during the shorter periods than in the years 1945–1995. This would lead to narrower intervals. On the other hand, when the estimation period is reduced, the estimated residual variance increases (other things remaining the same), and the intervals widen.

We found that the predicted *levels* for the TF, the mean age, and the variance are not significantly affected when we base the extrapolations on the shorter periods 1960–1995 or 1975–1995, see Keilman and Pham (1998) for details. As to the *confidence intervals*, the picture is somewhat mixed. In three of the six cases the intervals are comparable with those based on the period 1945–1995. The TF-intervals are sensitive for the choice of estimation period. It seems as if opting for the period 1945–1995 strikes a good balance between a high residual variance (1960–1995), and an imprecise estimate for the autoregressive coefficient (1975–1995).

8. Comparison with historical TF errors

Confidence intervals determine the *expected* errors in the *current* forecast. Investigating *observed* errors in *historical* forecasts can provide an independent check of the expected errors. We have looked at the errors in the TF-forecasts that Statistics Norway has published between 1969 and 1993. (Errors in the mean age or in the variance have not been computed, because the assumed values for these indicators in historical forecasts have not been documented.) There have been twelve of such forecasts, with jump-off years 1969, 1970, 1972, 1975, 1977, 1979, 1982, 1985, 1987, 1990, 1993, and 1996. We have omitted the 1996 forecast because it is of too recent date. For the other eleven we have compared assumed TF values for each forecast from the jump-off year until 1995 with "observed" values (moment estimators). We have updated the data originally assembled by Texmon (1992), who collected, among others, TF errors for the forecasts of 1969–1987 during the years 1969–1989. Most forecasts had more than one fertility variant, often two or three. In that case we included all variants in the data, because none of the forecasts, except one, contained a clear advice as to which of the variants was considered by Statistics Norway as the most probable one at the time of publication. Hence it
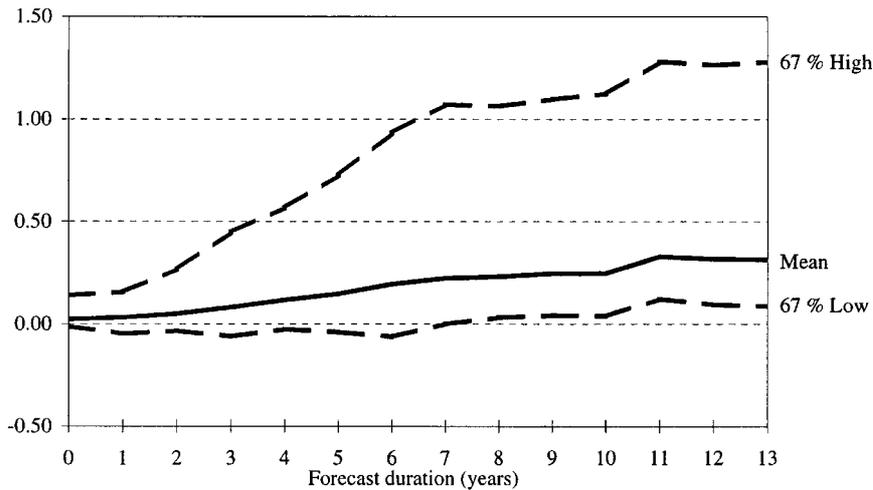
Figure 9. Mean error in historical TF-forecasts and 67 per cent interval for those errors, by forecast duration.

was left to the user to pick one of the variants. The exception was the 1993-based forecast, for which it was clearly indicated that the Medium fertility variant was considered as more probable than the High or the Low variant (Statistics Norway, 1994, p. 14). However, users were also advised to investigate the consequences of choosing the other two variants as an input to their own plan or analysis. Only if the user's conclusions depended little on the choice for the two extreme variants, was the user advised to employ the Medium variant. Thus we may assume that all variants have been used for the forecasts published between 1969 and 1993, although the middle one probably more often than the high or the low one (in case there were three variants).

   The error in the TF was simply defined as the assumed minus the "observed" value. Hence a positive or negative error indicates a value that is too high or too low. We have 25 series of TF errors, with a length of between three (the 1993 forecast) and 26 years (the 1970-forecast). The 1969-forecast contained results up to the year 1990 only. Next all errors were ordered by forecast duration, where the jump-off year was defined as duration 0. Hence we had 25 errors for each of the durations 0, 1, and 2 years; 22 for the durations 3–5 years, 19 for durations 6–8, 18 for durations 9–10, and 15 for durations 11–13 years. Errors for longer durations were so few that these were not analysed. For each duration, the errors were ordered from low (including negative values) to high. Finally we selected, by linear interpolation, if necessary, two error values, such that one-sixth of the errors for each duration were lower, and one-sixth were higher than these values. Hence these two values can be interpreted as the bounds of an empirical 67 per cent "confidence" interval. Two out of three errors are within these bounds, and one-sixth of the errors is higher or lower.

Figure 9 shows the bounds of the 67 per cent interval for the TF errors, together with the mean error. The lower bound is close to zero: the error was positive in five out of every six cases. This reflects the fact that the strong fertility decline in the 1970s (see Figure 1) came as a surprise for Norwegian population forecasters, as was the case for demographers in many other Western countries. The distance between the mean and the upper bound is much larger than that between mean and lower bound, indicating that large errors were much more frequent than small ones in the historical forecasts. The width of the 67 per cent interval grows from 0.15 children per woman in the jump-off year to 1.13 at duration 13. The historical errors in Figure 9 increase somewhat faster than the expected ones do. The 67 per cent confidence interval in Figure 4 is 1.13 children wide after a duration of 17 years, instead of 13 years, but the agreement between the two types of errors is striking. Thus the analysis in this section supports the main findings concerning the width of the predictive TF intervals in Section 5.[8]

9. Conclusion and discussion

We have shown how statistical techniques can be used to quantify uncertainty connected to age-specific fertility in the future. Stochastic simulation was used to generate 10,000 future time paths for fertility, based on a multivariate ARIMA model which describes the development over time in key parameters of the annual age schedule. In turn, these multiple time paths may be used as inputs to the traditional cohort component model. Together with similar time paths for mortality and migration, this facilitates the forecaster to simulate confidence intervals for future population sizes and age distributions.

Unlike traditional analyses, the method used in this paper takes due account of (1) sampling variability in the birth rates, (2) errors in the parameter estimates of the Gamma curve used for the fertility age pattern, (3) errors in the parameter estimates of the time series model for the prediction of fertility, and (4) residual variance in the time series model. Previous studies have dealt almost exclusively with the fourth source of uncertainty. The method was applied to data for Norway during the period 1945–1995. We found that when error source 3 is ignored, the 95 per cent predictive intervals for the TF and the mean age at childbearing in 2050 become too narrow by 8 per cent and 18 per cent, respectively. Where the first error source is not taken into account when fitting the Gamma curve, confidence bounds around predicted birth rates at ages where fertility is high, may be half as wide as they ought to be.[9] (But for forecasts of total births in a year one can expect cancellation of errors across ages, and this source of uncertainty is negligible.) When both the first and the second error sources were ignored, the 95 per cent predictive interval for the TF in 2050 became 1.7 child per woman too narrow.

The predictive intervals for the TF, the mean age at childbearing, and the age-specific fertility rates in this paper appear rather wide. For instance, around the year 2040 the 95 per cent interval for the TF is between 0.6 and 5.4 children per woman

on average. The interval for the mean age at childbearing ranges from 23.3 to 41.7 years. These intervals are so wide, that the model gives no useful information any longer. The method produces reasonable results up to the year 2020, or perhaps 2030. In the latter year, the 95 per cent intervals for the TF and the mean age are (0.7, 4.7) children per woman, and (24.4, 39.4) years, respectively. For the more distant future, the method cannot be used.

Yet it is reasonable to assume that information not included in the statistical models will imply narrower intervals. For instance, the adoption of modern forms of contraception during the 1960s is not explicitly contained in the model. There is no reason to assume a reversal of that trend in the future. An informal way of restricting the intervals is the following. If we are 100 per cent certain that Total Fertility in Norway in 2050 will not fall outside a range of, say, between 0.5 and 5 children per woman, this restriction can be included in the simulations, instead of the more liberal restriction of 0–10 children used in Section 5.[10]

The computed 67 per cent interval for the TF in 2010 is 1.0 children per woman wide, and that for the mean age is 3.2 years wide. In contrast, the gaps between TF and mean age values in 2010 in the High and the Low variant of Statistics Norway's 1996-based population forecast are only 0.4 children per woman and 1.0 years, respectively. On the basis of our model we must conclude that the chances are no more than approximately one in four that the real TF and the real mean age in 2010 will lie between the high and the low values assumed in the official forecast. This means that Norwegian forecasters have been too optimistic when they assumed such narrow bounds between their High and Low fertility variants. Our conclusion on wide bounds is supported, at least for the TF on the medium term, by an independent analysis of historical errors in TF forecasts since 1969.

The optimism (or self-confidence) among Norwegian population forecasters is not unique, we believe, for two reasons. First, forecasters in *other Western countries* do not have a much better record concerning the accuracy of their birth forecasts during the last thirty years (Keilman, 1997). Fertility trends in those countries appear to follow a more or less similar pattern. This pattern is characterized by high birth rates in the 1950s and in the first half of the 1960s, a steep fall in the 1970s, followed by smaller fluctuations during the last fifteen years when postponement and catching up processes played their part. The high-low gap between TF variants in those countries often amounts to 0.3–0.6 children per woman on the medium term (approximately 10 years, see Cruijsen and Keilman, 1994). These two facts together make it rather improbable that the real future TF will lie between the levels assumed for the High and the Low variant. Second, forecast evaluations for *other disciplines* have shown that experts often are too confident (Armstrong, 1985, p. 143). Forecasters, who state that they have confidence in their forecast, do not predict more accurately than those who say they are uncertain. Furthermore, self-reported confidence increases when a forecasting task is done more often. On the other hand, when forecasters are informed about the accuracy of earlier predictions, they become less confident.

The overoptimism among population forecasters is strikingly illustrated by two recent stochastic forecasts, one for Austria (Hanika et al., 1997; Lutz and Scherbov, 1998a), and the other one for Germany (Lutz and Scherbov, 1998b). The assumption in the Austrian forecast is that the TF in the year 2020 will lie with 90 per cent probability between 1.2 and 1.8 children per woman. This assumption is based on expert opinion, not on a statistical analysis. On the basis of the experiences reported in this paper, we expect that the probability for an interval of only 0.6 children per woman wide in 2020 should be in the order of magnitude of 50 per cent, instead of the assumed 90 per cent. The Austrian experts are far too optimistic, in our view. The German experts are a bit more cautious than their Austrian colleagues, but also in this case the assumed confidence intervals are too narrow. In the German study it is assumed that the TF in the year 2030 will lie with 90 per cent probability between 1.0 and 1.9 children per woman. We clearly advocate the construction and use of stochastic forecasts, but these should not be based on expert judgement alone: in our view, a *combination* of time-series analysis, analysis of historical errors, and expert opinion is appropriate. This paper has shown how these elements can be combined. Expert judgement has been implicit in many phases: choice of ARIMA model, length of data series, predicted levels, etc. It may be introduced even more explicitly by adopting upper and lower bounds ("100 per cent predictive intervals", loosely speaking) for the predicted Total Fertility levels, and rejecting simulations that exceed those bounds.

Acknowledgements

Notes

[1] To assume that fertility only depends on age, and not on other covariates (parity, birth interval, education, labour force status, income) of the woman or her partner is clearly a strong simplification of reality. It is unclear to what extent this affects the findings of this paper. De Beer (1989) finds that including parity in ARIMA models for fertility in the Netherlands improves their predictive accuracy somewhat.

[2] Since in the end we are concerned with errors in the projected numbers of births, we consider the *absolute* size of the measurement error, not the *relative* size, which would amount to $Var(F_x)/F_x$. Thus for some ages the measurement error is small simply because the intensity is low.

[3] We have not investigated whether full Maximum Likelihood (instead of the approximate ML approach) or other estimation methods would have avoided this problem.

[4] Estimates for the minimum age β_4 fell below 14 in 1975 and decreased further to reach zero in 1991. During the same period, the estimates for the mean age at childbearing β_2 rose from 26.6 to

28.3 years. Together with the relatively low estimates for the Total Fertility β_1 during these years (<2), the predicted birth rates at ages below 16 are still negligible, in spite of the unrealistic estimate for the minimum age.

[5] Note that Thompson et al. (1989) find a similar (although steeper) drop in the minimum age.

[6] Preliminary tests on the basis of a univariate model for the Total Fertility showed that this was an appropriate choice.

[7] This empirical evidence is in line with theoretical results obtained by Koreisha and Fang (1999), who show that predictions generated by (univariate) ARIMA models become more uncertain as a result of measurement errors, given certain conditions for the autocovariance function of the process.

[8] One may argue that using the historical errors is too pessimistic, since one ignores improvements in forecast methodology. Indeed, average errors for the forecasts made in the 1980s are smaller than those for the forecasts from the 1970s. But this is to be expected, because the early forecasts are of longer duration, and hence more uncertain. When one wishes to estimate the average error in subsequent forecasts, controlling for this duration effect, a multivariate model is needed. An analysis of this kind carried out for Norwegian fertility forecasts showed hardly any improvement over subsequent forecast rounds (Keilman, 1997). Period effects dominate the error pattern.

[9] These bounds increase inversely proportionally with the square root of the number of women at risk. Compared to the case of Norway (4.4 million inhabitants), they are a third in countries with approximately $9 * 4.4 = 40$ million inhabitants, and only one-fourth in countries with $16 * 4.4 = 70$ million inhabitants. As one referee pointed out, although the context is different, this is consistent with Schweder's finding (1971) that random variations in fertility and mortality rates have a minor impact on forecast uncertainty. Shifts in vital processes are the dominant source of errors.

[10] When this was implemented in a trial simulation, we noted that the 67 per cent confidence interval for Total Fertility in 2010 was still 0.9 children per woman wide (between 1.43 and 2.37), which is to be compared to the 1.0 children gap in Figure 4. In the long run, the difference with Figure 4 was somewhat larger, in particular for the more extreme TF-values. For instance, the 95 per cent confidence bounds became (0.7, 4.3), instead of (0.5, 6.1) in Figure 4.

References

Alders, M. and de Beer, J., 1998. 'Kansverdeling van de bevolkingsprognose' ('Probability distribution of population forecasts'). *Maandstatistiek van de Bevolking* 46(4): 8–11.

Alho, J. M., 1990. 'Stochastic methods in population forecasting'. *International Journal of Forecasting* 6: 521–530.

Alho, J. M., 1998. A stochastic forecast of the population of Finland. Reviews 1998/4. Statistics Finland, Helsinki.

Alho, J. and Spencer, B., 1997. Statistical demography and forecasting. Unpublished manuscript.

Armstrong, J., 1985. *Long-range Forecasting: From Crystal Ball to Computer*, 2nd ed. Wiley, New York.

Beets, G., 1996. 'Does the increasing age at first birth lead to increases in involuntary childlessness?' Proceedings European Population Conference Milan 4–8 September 1995, Volume 2. FrancoAngeli, Milan, 15–30.

Bell, W., 1992. 'ARIMA and Principal Components models in forecasting age-specific fertility', in N. Keilman and H. Cruijsen (eds.), *National Population Forecasting in Industrialized Countries*. Swets & Zeitlinger, Amsterdam, 177–200.

Bell, W., 1997. 'Comparing and assessing time series methods for forecasting age-specific fertility and mortality rates'. *Journal of Official Statistics* 13(3): 279–303.

Bratley, P., Fox, B. and Schrage, L., 1983. *A Guide to Simulation*. Springer-Verlag, New York.

Brunborg, H. and Mamelund, S.-E., 1994. '*Kohort- og periodefruktbarhet i Norge 1820–1993*' ('Cohort and Period Fertility for Norway 1820–1993'), Report no. 94/27. Statistics Norway, Oslo.

Cruijsen, H. and Keilman, N., 1994. 'Population projections by statistical agencies in countries of the European Economic Area', Paper Annual Meeting of the Population Association of America, Miami, May 1994.

De Beer, J., 1989. 'Projecting age-specific fertility by using time-series models'. *European Journal of Population* 5: 315–346.

De Beer, J., 1992. 'General time-series models for forecasting fertility, in N. Keilman and H. Cruijsen (eds.), *National Population Forecasting in Industrialized Countries.* Swets & Zeitlinger, Amsterdam, 148–175.

Duchêne, J. and Gillet-De Stefano, S., 1974. 'Ajustement analytique des courbes de fécondité générale'. *Population et Famille* 32(2): 53–93.

Eurostat, 1997a. 'Latest national population forecasts for the countries of the European Economic Area and Switzerland – Reviewed by a set of tables', Background Document Meeting Working Party on Demographic Projections. Luxemburg, September 1997.

Eurostat, 1997b. 'Beyond the predictable: Demographic changes in the EU up to 2050', *Statistics in Focus, Population and Social Conditions* 1997/7.

Hanika, A., Lutz, W. and Scherbov, S., 1997. 'Ein probabilistischer Ansatz zur Bevölkerungs-vorausschätzung für Österreich'. *Statistische Nachrichten* 12: 984–988.

Hoem, J. M., 1976. 'On the optimality of modified chi-square analytic graduation'. *Scandinavian Journal of Statistics* 3: 89–92.

Hoem, J. M., Hansen, H. O., Madsen, D., Nielsen, J. Løvgreen, Olsen, E. M. and Rennermalm, B., 1981. 'Experiments in modelling recent Danish fertility curves'. *Demography* 18(2): 231–244.

Keilman, N., 1997. 'Ex-post errors in official population forecasts in industrialized countries'. *Journal of Official Statistics* 13(3): 245–277.

Keilman and Pham, D. Q., 1998. 'Predictive intervals for age-specific fertility', Documents nr. 98/19. Statistics Norway, Oslo.

Knudsen, C., McNown, R. and Rogers, A., 1993. 'Forecasting fertility: An application of time series methods of parameterized model schedules'. *Social Science Research* 22: 1–23.

Koreisha, S. G., and Fang, Y., 1999. 'The impact of measurement errors on ARMA prediction'. *Journal of Forecasting* 18(2): 95–109.

Kravdal, Ø., 1994. 'The importance of economic activity, economic potential and economic resources for the timing of first births in Norway'. *Population Studies* 48: 249–267.

Lee, R., 1993. 'Modeling and forecasting the time series of US fertility: Age distribution, range, and ultimate level'. *International Journal of Forecasting* 9: 187–202.

Lee, R., 1999. 'Probabilistic approaches to population forecasting', in W. Lutz, J. Vaupel, and D. Ahlburg (eds.), *Frontiers of Population Forecasting.* Supplement to Vol. 24 of *Population and Development Review*: 156–190.

Lee, R. and Tuljapurkar, S., 1994. 'Stochastic population forecasts for the United States: Beyond High, Medium, and Low'. *Journal of the American Statistical Association* 89(428): 1175–1189.

Lesthaeghe, R. and Surkyn, J., 1988. 'Cultural dynamics and economic theories of ferility change'. *Population and Development Review* 14(2): 3–45.

Lütkepohl, H., 1993. *Introduction to Multiple Time Series Analysis*, 2nd ed. Springer-Verlag, Berlin.

Lutz, W. and Scherbov, S., 1998a. 'An expert-based framework for probabilistic national population projections: The example of Austria'. *European Journal of Population* 14(1): 1–17.

Lutz, W. and Scherbov, S., 1998b. 'Probabilistische Bevölkerungsprognosen für Deutschland' ('Probabilistic population projections for Germany'). *Zeitschrift für Bevölkerungswissenschaft* 23(2): 83–109.

Schweder, T., 1971. 'The precision of population projections studied by multiple prediction methods'. *Demography* 8(4): 441–450.

Spengler, J., 1936. 'Population prediction in nineteenth century America'. *American Sociological Review* 1: 905–921.

Statistics Norway, 1994. *Framskriving av folkemengden 1993–2050* ("Population Projections 1993–2050"). Statistics Norway, Oslo/Kongsvinger.

Statistics Norway, 1997. *Framskriving av folkemengden 1996–2050* ("Population Projections 1996–2050"). Statistics Norway, Oslo/Kongsvinger.

Texmon, I., 1992. 'Norske befolkningsframskrivinger 1969–1990' ('Norwegian population projections 1969–1990'), in O. Ljones, B. Moen, and L. Østby (eds.), *Mennesker og Modeller.* Statistics Norway, Oslo/Kongsvinger, pp. 285–311.

Thompson, P., Bell, W., Long, J. and Miller, R., 1989. 'Multivariate time series projections of parametrized age-specific fertility rates'. *Journal of the American Statistical Association* 84: 689–699.

Van Imhoff, E., 1991. *Profile: A Program for Estimating the Coefficients of Demographic Age-intensity Profiles*. NIDI-Report no. 15. The Hague: NIDI.