International Statistical Review (2004), **72,** 1, 5–18, Printed in The Netherlands

Time Series Based Errors and Empirical Errors in Fertility Forecasts in the Nordic Countries*

Nico Keilman[1,2] and Dinh Quang Pham[2]

[1] *Department of Economics, University of Oslo, Norway. E-mail: nico.keilman@econ.uio.no*
[2] *Statistics Norway, Oslo, Norway*

Summary

We use ARCH time series models to derive model based prediction intervals for the Total Fertility Rate (TFR) in Norway, Sweden, Finland, and Denmark up to 2050. For the short term (5–10 yrs), expected TFR-errors are compared with empirical forecast errors observed in historical population forecasts prepared by the statistical agencies in these countries since 1969. Medium-term and long-term (up to 50 years) errors are compared with error patterns based on so-called naïve forecasts, i.e. forecasts that assume that recently observed TFR-levels also apply for the future.

Key words: Time series; ARCH model; Stochastic population forecast; Total Fertility Rate; Empirical forecast errors; Naïve forecast; Nordic countries.

1 Introduction

Long-term population forecasts, covering a period of fifty years or more, are useful in a number of fields, two of which are analyses of the impact of population trends on contributions and expenditures for old-age pensions, and studies of demographically induced resource use and climate change. Such long-term forecasts are necessarily uncertain: for a given country, one may imagine many possible demographic futures, but some of these population developments are more probable than others. This calls for stochastic population forecasts, i.e. forecasts in terms of prediction intervals. Such prediction intervals quantify uncertainty—they express the expected probability that the future population (or age group, or number of births) falls within a certain range.

A number of recent stochastic population forecasts have used some form of time series analysis for one or more key indicators, in order to assess the expected accuracy of predicted values for these indicators. The most commonly used summary indicator for the level of fertility in a certain year t is the Total Fertility Rate (TFR), defined as the number of children a woman is expected to have over her lifetime, if the age-specific fertility of year t would hold through the woman's childbearing ages, and she would survive to the end of childbearing ages. Time series models were used to predict the TFR in stochastic forecasts prepared for the US (Lee & Tuljapurkar, 1994), Finland (Alho, 1998), the Netherlands (De Beer & Alders, 1999), and Norway (Keilman *et al.*, 2001). One attractive property of time series models is that they not only give a prediction of future values of the variable in question, but also allow us to compute prediction intervals.

*Revised version of a paper presented at the seminar "How to deal with uncertainty in population forecasting?" Vienna, 12–14 December 2002. Comments made by seminar participants are gratefully acknowledged. Research was supported by grant nr. SERD-2000-00172 under the Fifth Framework of the European Commission.

A common finding with TFR-time series in industrialized countries is that these are non-stationary. As a consequence, long run prediction intervals, when unchecked, may become extremely wide. Therefore, adjustments are necessary. For instance, Lee & Tuljapurkar (1994) introduced upper and lower bounds to the TFR by a generalized logit-transformation. This way they constrained TFR-predictions to between 0 and 4 children per woman on average. Alho (1998) found that time-series based TFR-prediction intervals 50 years ahead were 15 per cent wider than those obtained based on the volatility in the historical TFR-observations, and he decided to rely on the latter type of intervals. De Beer & Alders (1999) initially found a 95-per cent prediction interval for the TFR in 2050 equal to [0.6–2.8] based on time series models. Next, an analysis of fertility by birth order led them to suggest that an interval of [1.1–2.3] would be more appropriate. Keilman *et al.* (2001) simulated predicted TFR-values, and rejected TFR-simulations that would fall outside the interval [0.5–4] in any year up to 2050.

Checks of this kind for time series predictions involve judgement. Therefore, additional information will be useful when judging whether the subjective decisions and the adjustments are reasonable. One source of additional information is the accuracy of old TFR-predictions. Many statistical agencies have published population forecasts in the past, including TFR-predictions. When compared with actual TFR-values observed for the years after the forecast was made, one can check the observed TFR-accuracy against predicted TFR-accuracy, obtained by time series models.

In practice however, analyses of this type are restricted to forecasts up to 10 years ahead, seldom longer. This is because few statistical agencies have documented their population forecasts in sufficient detail before the 1960s, and hence the series of historical errors is rather short. However, fertility assumptions in official forecasts, in particular those on the long term, are often close to a recently observed level, which is held constant for the future. Hence, starting from a time series of observed fertility levels, one could compute so-called naïve errors, i.e. errors that result from the naïve assumption that future fertility will be the same as the current one. For a few countries, for example the Nordic countries, long time series for the TFR exist. This allows us to construct rather large data sets with naïve TFR-errors.

The purpose of this paper is to compare the results of Total Fertility Rate predictions made by time series methods with those obtained by analysing both observed forecast errors and naïve errors. The focus is on the expected accuracy of the predictions, i.e. on the width of the prediction intervals. We shall use time series models to derive model-based prediction intervals for the TFR in Norway, Denmark, Sweden, and Finland up to 2050. For the short term and medium term (up to 10 yrs), expected TFR-errors are compared with empirical forecast errors observed in historical population forecasts prepared by the statistical agencies in these countries since 1969. Long-term TFR-errors obtained from the time series models are compared with naïve errors in the TFR. The longest TFR-series we have is that for Finland, which starts in 1776. The Norwegian and Swedish series start in 1845 and 1855, respectively, while observations for Denmark are available from 1911.

2 Time Series Models

Figure 1 plots the Total Fertility Rate for the four countries. The data sources are listed in the Appendix.

The four countries show a similar pattern in the TFR, which reflects the demographic transition, followed by the effects of the economic recession in the 1930s and the baby boom in the 1950s and 1960s. Plots of first differences in the TFR (not shown here) revealed less similarity. Finland exhibits the strongest fluctuations, in particular before 1950. In the 20th century, all four countries show a tendency towards lower variability in the TFR, although this tendency is much stronger for Finland than for the other three countries, see Table 1. Major events, such as the two world wars, and the occurrence of the Spanish Influenza in 1918/1919 are clearly reflected in the series for all four countries.

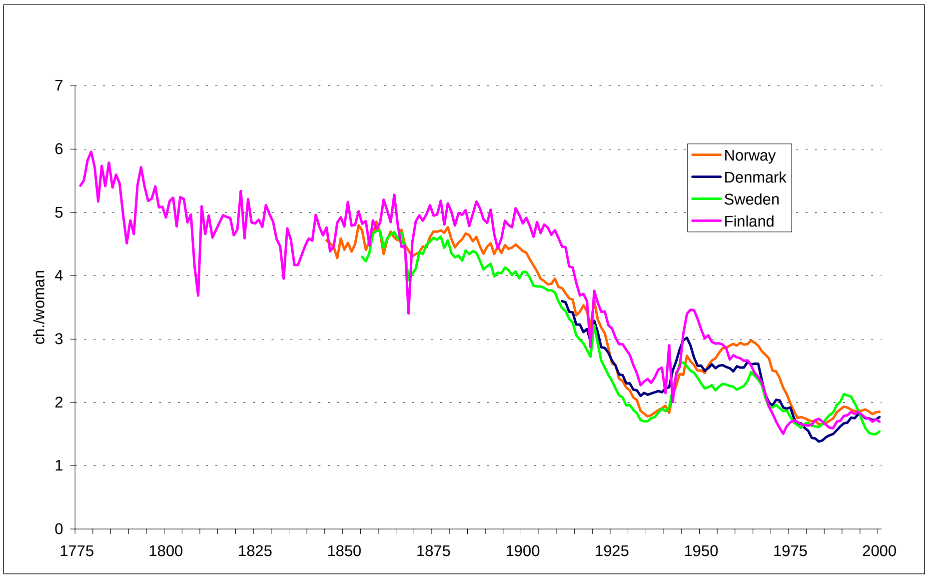

Figure 1. *Total fertility rate.*

Table 1
Standard deviations of the first differences in the TFR for various sub-periods.

	Norway	Denmark	Sweden	Finland
1776–1850	–	–	–	0.346
1851–1900	0.135	–	0.142	0.319
1901–1950	0.138	0.123	0.126	0.288
1951–2000	0.065	0.074	0.073	0.071

An important question is how much of the data should be used in the modelling. Several issues are at stake here. First, Box & Jenkins (1970, p.18) suggest at least 50 observations for ARIMA-type of time series models, although annual models (in contrast to monthly time series) probably need somewhat shorter series. Second, the quality of the data is better for the 20th century than for earlier years. This is particularly true for the denominators of the fertility rates, i.e. the annual numbers of women by single years of age. Third, one may question the relevance of data as long back as the mid-1800s. Current childbearing behaviour is very different from that of women in the 19th century. Fourth, our ultimate goal is to compute long-term predictions of some 50 years ahead, which necessitates a long series.

The ultimate choice is necessarily a subjective one, which includes a good deal of judgement and arbitrariness. We believe that we strike a reasonable balance between conflicting goals by selecting the 20th century as the basis for our models. An analysis based on the last 50 years, say, would be unfortunate: it would include the baby boom of the 1950s and early 1960s, but not the low fertility of the 1930s, to which the boom was a reaction, at least partly. A base period stretching back into the 19th century would be hampered by problems of data quality, and it would also unrealistically assume that the demographic behaviour over such a long period could be captured by one and the same model. In a sensitivity analysis we also experimented with base period 1945–2000. For Norway and Finland we found 95 per cent prediction intervals that were smaller (by 1.4 and 0.5 children per woman on average, respectively) than those that we have accepted for further analysis (see Table 4). For Denmark and Sweden they were larger (by 0.8 and 1.2 children per woman, respectively). To increase comparability across countries, we used a TFR time series starting in 1900 for Denmark as well. Danish TFR-values for the years 1900–1910 were estimated on the basis of observed Crude Birth rate values; see the Appendix.

Traditional time series models of the ARIMA type assume homoscedasticity, i.e. constant residual variance. Given the tendency towards less variability in the TFR in recent decades, such traditional models could not be used. The Autoregressive Conditional Heteroscedastic (ARCH) model introduced in Engle (1982) combines time-varying variance levels with an autoregressive process. This model, and its generalizations (generalized, integrated, and exponential ARCH models, to name a few) have gained popularity in recent years (Bollerslev, 1986). The model has already proven useful in analysing economic phenomena such as inflation rates, volatility in macroeconomic variables, and foreign exchange markets; see Bollerslev (1986) for a review. Application to demographic time series is less widespread. Yet, given the varying levels of volatility in the TFR during the 20th century, an ARCH-type of model is an obvious candidate.

Let Z_t be the logarithm of the TFR in year t. Then the model is

$$\begin{aligned} Z_t &= C + \phi Z_{t-1} + \nu_t + \eta_1 U_{1,t} + \eta_2 U_{2,t} + \eta_3 U_{3,t} + \eta_4 U_{4,t} \\ \nu_t &= \psi_1 \nu_{t-1} + \psi_2 \nu_{t-2} + \cdots + \psi_m \nu_{t-m} + \varepsilon_t \\ \varepsilon_t &= \left(\sqrt{h_t}\right) e_t \\ h_t &= \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 \end{aligned} \tag{1}$$

where $e_t \sim N(0, 1)$. This is the AR(m)–ARCH(q) model. The outliers caused by the two world wars and by the Spanish Influenza are handled by between two (Denmark and Sweden) and four (Norway, Finland) dummy variables $U_{i,t}$. In addition we have $\omega > 0$ and $\alpha_i \geq 0$.

The maximum number of terms m included in the autoregressive expression of ν_t was set equal to 10, but few of the ψ-estimates turned out to be significantly different from zero. Similarly, estimates for α_i suggested that the order (q) of the CH-part of the model could be one or two, not higher. This led to the specifications for the AR(m)–ARCH(q) model with dummy variables for the four countries as given in Table 2.

Table 2

ARCH models with dummy variables for the TFR in Norway, Denmark, Sweden, and Finland, 1900–2000.

	Dummy variables	AR:$\nu_t =$	CH:$h_t =$
Norway	1915, 1919, 1920, 1946	$\psi_2 \nu_{t-2} + \varepsilon_t$	$\omega + \alpha_1 \varepsilon_{t-1}^2$
Denmark	1920, 1942	$\psi_1 \nu_{t-1} + \varepsilon_t$	$\omega + \alpha_1 \varepsilon_{t-1}^2$
Sweden	1920, [1942–45][1]	$\psi_1 \nu_{t-1} + \varepsilon_t$	$\omega + \alpha_1 \varepsilon_{t-1}^2$
Finland	1919, 1920, 1940, 1941	$\psi_2 \nu_{t-2} + \varepsilon_t$	$\omega + \alpha_1 \varepsilon_{t-1}^2$

Note 1: one common dummy variable for each year in the period 1942–45.

Table 3 lists the estimation results, with t-values in parentheses. None of the C-estimates were significantly different from zero. Yet the constant was retained in the model: trial calculations omitting the constant resulted in implausibly low point predictions for the TFR in 2050, ranging from a high 1.38 children per woman in Norway, to a low 1.21 children per woman in Sweden. The predictions for 2050 based on the model with constant are in the interval from 1.71 (Denmark) to 1.46 (Finland).

Note the high α_1-estimate for Finland. It reflects the large variability in the Finnish data. The Swedish α_1-estimate is close to zero—however, omitting the ARCH(1) part from the model (which essentially boils down to assuming constant variance) would lead to rejecting the null hypothesis of constant variance in the residuals, as trial calculations showed.

Table 3

Parameter estimation results for the models in Table 2; t-values in parentheses.

	η_1	η_2	η_3	η_4	C	ϕ	ψ_1	ψ_2	ω	α_1
Norway	-0.0680	-0.0531	0.1830	0.1016	0.0114	0.9754		0.4771	0.0005	0.3626
	(-2.34)	(-2.41)	(8.78)	(2.28)	(0.61)	(48.07)		(5.41)	(5.18)	(2.45)
Denmark	0.1493	0.0901			0.0157	0.9701		0.3249	0.00009	0.2721
	(2.06)	(2.76)			(0.86)	(48.23)		(3.35)	(7.18)	(1.95)
Sweden	0.2258	0.0753			0.0215	0.9588	0.5788		0.0007	0.0000
	(10.40)	(4.51)			(1.15)	(48.69)	(6.53)		(6.01)	(0.00)
Finland	-0.2085	0.2909	-0.1095	0.1563	0.0053	0.9812		0.2135	0.0007	0.7080
	(-2.14)	(1.72)	(-6.43)	(7.46)	(0.47)	(98.03)		(2.38)	(4.24)	(3.48)

The overall impression is that model (1) is a useful device to capture the TFR-trends in the four Nordic countries during the past century. We used the model to compute prediction intervals for the future TFR up to 2050. Since we cannot be certain that the estimated coefficients are equal to the real ones, we used simulation to obtain these intervals. In each of the 5000 simulation runs, parameter values were drawn from a multivariate normal distribution, with expectation equal to the parameter estimates in Table 3, and with corresponding covariance matrix as estimated earlier. The possibility that a pandemic as bad as the Spanish Flu, or a war with consequences as catastrophic as WWI or WWII could occur during the prediction period, was included in these simulations. For each of the two dummy variables, we first drew a random number from the binomial distribution with a probability of "catastrophe" equal to 1/101. Next, the starting year for the catastrophe was determined on the basis of a random draw from the uniform distribution on the interval [2001, 2050]. Finally, the appropriate η-value was drawn from its estimated distribution.

The estimates for the AR-coefficient ϕ are all between 0.96 and 0.98. Although those estimates were extremely sharp (cfr. the corresponding t-values), the simulations resulted in a few ϕ-values equal to or larger than one. These were rejected, and ϕ was redrawn until 5000 admissible values had been obtained for each country.

Table 4 lists point forecasts (i.e. expected values) and prediction intervals for selected years up to 2050. Long-range 95 per cent prediction intervals are roughly two children per woman wide. Ten years ahead 95 per cent intervals are 1.1 to 1.2 children per woman wide.

The time-series model predicts an expected TFR-value up to 2050 around 1.7 children per woman, except for Finland, where the prediction is a low 1.5. Although this is not an impossible value, it is much lower than that in official population forecasts for Finland, for instance Statistics Finland (1.75, see Council of Europe, 2001) or the United Nations (1.85, see `http : //www.un.org/esa/population/publications/wpp2002` /wpp2002annextables.PDF). In this paper, we focus primarily on the width of the predictive distribution, much less on its central tendency. In an actual fertility prediction, all long-term point predictions would have to be examined critically, for example by inspecting the mean number of children born to women in successive birth generations.

Are the long-term prediction intervals in Table 4 reasonable? One may object that the interval

Table 4

Simulated TFR-point forecasts and prediction intervals.

	forecast	67% interval lower bound	67% interval upper bound	67% interval interval width	95% interval lower bound	95% interval upper bound	95% interval interval width
				Norway			
2000	1.85						
2005	1.86	1.72	2.02	0.30	1.55	2.21	0.76
2010	1.83	1.61	2.10	0.49	1.37	2.48	1.11
2015	1.80	1.53	2.16	0.63	1.23	2.63	1.40
2035	1.72	1.33	2.23	0.90	0.90	2.84	1.94
2050	1.68	1.24	2.25	1.01	0.77	2.96	2.19
				Denmark			
2000	1.77						
2005	1.78	1.63	1.97	0.34	1.45	2.18	0.73
2010	1.77	1.54	2.06	0.52	1.32	2.40	1.08
2015	1.76	1.48	2.10	0.62	1.21	2.53	1.32
2035	1.73	1.34	2.19	0.85	0.94	2.73	1.79
2050	1.71	1.27	2.21	0.94	0.82	2.82	2.00
				Sweden			
2000	1.54						
2005	1.61	1.44	1.79	0.35	1.29	2.03	0.74
2010	1.62	1.37	1.92	0.55	1.12	2.30	1.18
2015	1.64	1.32	2.00	0.68	1.02	2.49	1.47
2035	1.66	1.19	2.16	0.97	0.76	2.80	2.04
2050	1.67	1.16	2.18	1.12	0.69	2.83	2.14
				Finland			
2000	1.70						
2005	1.67	1.53	1.83	0.30	1.34	2.07	0.73
2010	1.63	1.42	1.89	0.47	1.17	2.28	1.11
2015	1.60	1.33	1.93	0.60	1.03	2.45	1.42
2035	1.51	1.12	2.02	0.90	0.73	2.82	2.09
2050	1.46	1.02	2.03	1.01	0.60	2.83	2.23

bounds are rather wide. The Norwegian model predicts a 16 per cent chance of a TFR in 2050 of at least 2.2 children per woman, which is much higher than the European average of 1.4, but close to the current fertility level of developing countries such as Brazil, Lebanon, or Tunisia. The lower 95 per cent bound in Finland in 2050 is 0.6 children per woman. Both upper and lower bounds are incompatible with current demographic theories for industrialized countries, one may argue. However, these objections are unconvincing, in our view, for three reasons. First, fertility theories have poor predictive performance. Keyfitz (1982) assessed various theories, and concluded that they have limited predictive validity in space and time, are strongly conditional, or cannot be applied without the difficult prediction of non-demographic factors. Fertility theories developed in more recent years do not fare much better (Lee, 1997, p.52; Van de Kaa, 1996, p.390). This situation is not specific for demography. Any generalizations about human behaviour are bound to be narrowly restricted to specific institutional settings or particular epochs (Nagel, 1961; Boudon, 1986; Henry, 1987). Second, research on forecasting in other fields shows that subject matter experts often are too confident, in the sense that they tend to give prediction intervals that are too narrow (Armstrong, 1985). We have no reasons to believe that the situation in demographic forecasting is very different. Third, the interval *bounds* should not be taken as *probable* fertility levels. On the contrary, the 95 per cent bounds merely indicate the *outliers* of predicted fertility paths. Even the 67 per cent bounds reflect TFR-intervals with chances equal to only one in six.

3 Other Time Series Models

In order to assess the robustness of the prediction intervals obtained in Section 2, we have experimented with several other time series models for Z_t:

- a pure AR(m)-model
- an AR(m)–CH(1) model
- an AR(m)-model with dummy variables.

The results can be summarized as follows.

Fitting an AR(m)-model or an AR(m)–CH(1) model

A purely autoregressive model for Z_t (with maximum lags equal to 5 for Norway, 2 for Denmark, 1 for Sweden, and 2 for Finland) indeed indicated non-constant variance: using a Portmanteau Q-test and a Lagrange Multiplier (LM-) test, a hypothesis of homoscedastic residuals had to be rejected at the five per cent level for Norway, Sweden, and Finland. For Denmark, such a hypothesis was not rejected at the ten per cent level. When we introduced a CH(1)-part to the model in order to account for heteroscedastic residuals, the situation improved considerably for Denmark and Sweden, but for Norway and Finland, there were still some signs of heteroscedasticity at lag 1. A Kolmogorov–Smirnov test hypothesis for normality could only be accepted for Norway and Denmark (at the five per cent level), not for the other two countries.

Fitting an AR(m)-model with dummy variables

Can the non-constant residual variance be captured by introducing dummy variables to the AR(m)-model? This is the case for three countries: Finland was the exception. For the other three countries, dummy variables for the periods around 1918, 1944, and 1970 were introduced. A hypothesis of homoscedastic residuals at all lags could not be rejected (5%) for all three countries. The 95% prediction intervals in 2050 turned out to be 2.7 (Norway), 3.0 (Denmark), and 3.5 (Sweden) children per woman wide—much wider than the intervals in Table 4.

Based on these sensitivity tests we conclude that the ARCH-model in expression (1) gives a useful and reliable description of the development in the TFR in the four countries in the previous century. Next it remains to be seen whether the prediction intervals of this model are reasonable.

4 Errors in Historical TFR Forecasts

Prediction intervals determine the *expected* errors in the *current* forecast. Investigating *observed* errors in *historical* forecasts could provide an independent check of the expected errors. We have analysed the errors in historical TFR-forecasts prepared by statistical agencies in the four countries. Bibliographic details are given in the appendix. The help of Timo Nikander and Ossi Honkanen in Helsinki, Anna Qvist in Copenhagen, Jan Qvist in Stockholm, and Åke Nilsson in Ørebro is gratefully acknowledged.

For Norway, we have TFR-forecasts starting in 1969, for Denmark the series starts in 1974, and for Sweden and Finland the starting year is 1972. Assumed TFR-values for each forecast from the jump-off year until 2000 were compared with observed values (Council of Europe, 2001). Most forecasts had more than one fertility variant, often two, or three. In that case, we included all variants in the data, because very few of the forecast reports contained a clear advice as to which of the variants the statistical agency considered as the most probable one at the time of publication. Hence, it was left to the user to select one of the variants. We may assume that all variants have been used, although the middle one probably more often than the high or the low one (in case there were three variants). Below we shall use the standard deviation of the observed TFR-error. For Norway, the error patterns in this indicator based on all forecast variants were very close to those based on main variants only. For Sweden, the all-variants standard deviations were approximately 10 per cent higher than those based on main variants. We have 31 series of TFR-forecasts for Norway, 51 for Denmark, 23 for Sweden, and 34 for Finland. Each series was ordered by forecast duration, with the jump-off year defined as duration 0.

Since TFR-forecasts may be higher or lower than observed TFR-values, we have used the *signed error* of the TFR-forecasts, defined as the assumed minus the observed value. Therefore, a positive (negative) error indicates a value that is too high (too low). The purpose of the current analysis is to check the uncertainty in model-based TFR-predictions, in other words, the width of the model-based prediction intervals. Thus, for each forecast duration, we computed the standard deviation of the signed TFR-errors, as a measure of uncertainty. The empirical standard deviation reflects uncertainty in the future TFR appropriately, provided that the expected value of the TFR is predicted correctly. The latter assumption may be relaxed by inspecting the Root Mean Squared Error (RMSE) of the TFR, which adds a bias component to the standard deviation (Maddala, 1977). The empirical RMSE's for the four countries diverged only slowly from the standard deviations, and hence for the cases of Norway, Denmark, and Sweden they declined for long durations, too (although the decline set in a few years later than that in the standard deviation). Figure 2 plots these standard deviations.

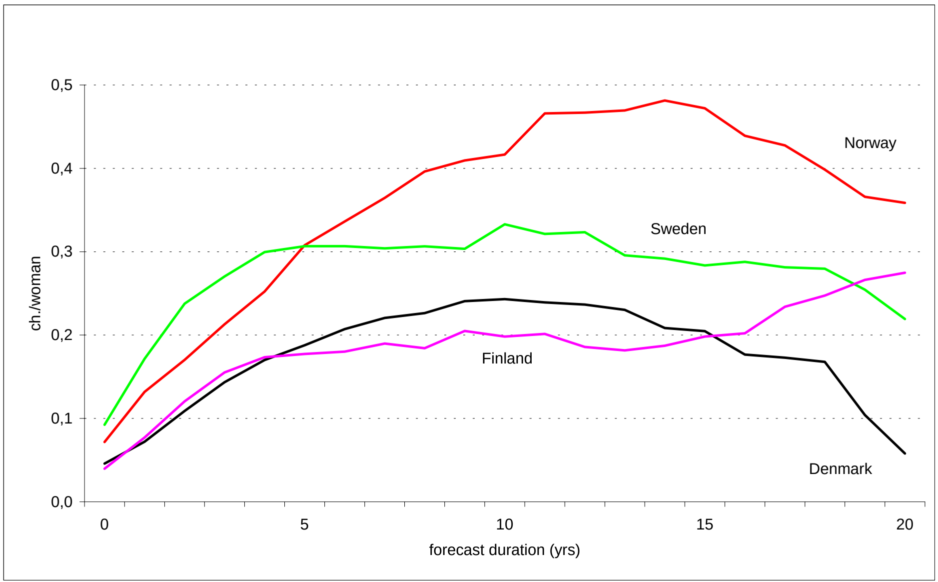

Figure 2. *Standard deviations of signed TFR-errors.*

Figure 2 shows that for the first years of the forecast, the standard deviations increase regularly, from less than 0.1 children per woman in the first forecast year (duration 0), to 0.2–0.3 children per woman five years ahead. Next, the patterns tend to stabilize (quite soon for Sweden, but at 10–15 years ahead also for Denmark and Norway). This stabilization, however, should not be interpreted as an indication that the uncertainty in TFR-predictions prepared since 1969 was roughly constant for durations longer than five years ahead. The stable pattern for durations beyond five years ahead in Figure 2 is a consequence of the fact that the data set is right-censored. The year 2000 was the last year for which TFR-errors were computed. Hence, there are very few errors for long forecast durations. For instance, 15 years ahead we have only 18 observations for Norway, 13 for Denmark, 11 for Sweden, and 16 for Finland. Therefore, the long-term standard deviations in Figure 2 are not very reliable. At the same time, the errors at long durations all apply to the relatively stable period of the last ten years—see Figure 1.

Assuming that the errors for a given duration and a particular country are approximately normally distributed, we can construct a confidence interval for the estimated standard deviation of that distribution. Table 5 gives these intervals, which were obtained on the basis of a χ^2-distribution with degrees of freedom equal to the number of observations minus 1.

Table 5

Estimated standard deviations in TFR-errors for forecast durations of 5 and 10 years, with corresponding 95-per cent confidence intervals in parentheses.

	5 years ahead	10 years ahead
Norway	0.308 (0.240–0.428)	0.417 (0.320–0.595)
Denmark	0.188 (0.152–0.245)	0.243 (0.187–0.348)
Sweden	0.307 (0.237–0.434)	0.333 (0.248–0.507)
Finland	0.177 (0.142–0.236)	0.198 (0.150–0.293)

We note that for forecasts five years ahead, errors for Norway and Sweden have considerably higher standard deviations than those for Denmark and Finland, even when estimation uncertainty is taken into account. This is probably due to the steep fall in the Norwegian TFR in the early 1970s, and the strong fluctuations in the Swedish TFR around 1990, see Figure 1. These factors made it rather difficult to prepare an accurate TFR-forecast, at least more difficult than for Denmark and Finland. The estimates and confidence intervals 10 years ahead still reflect this to a certain extent, although the (unknown) theoretical standard deviations of Denmark and Sweden are probably not very different from each other.

5 Errors in Naïve TFR-forecasts

Inspection of historical official population forecasts shows that in many cases, forecasters in effect have assumed that the TFR would remain relatively close to its current level, even in the long run; see, for example, Lee (1974) for the US, Keilman (1990) for the Netherlands, and Texmon (1992) for Norway. This observation led Alho (1990) to propose inferring prediction intervals for the TFR from the errors of what he called naïve forecasts, i.e. forecasts based on the naïve assumption that current levels will persist over a long time. Clearly, in times of falling fertility, demographers will extrapolate the decline for at least a few years, but the examples mentioned above demonstrate that a constant level is assumed quite soon.

We have computed errors in the naïve TFR forecasts for Norway, Sweden, and Finland for the years 1900–2000, and for Denmark for the period 1911–2000. Hence the naïve TFR forecasts for the first three countries are based on observed TFR levels for the years 1899–1999 (duration 0), 1888–1988 (duration 1), 1887–1987 (duration 2), etc. This means that there are 101 errors for forecast durations 0–50 years for Norway and Finland, and for durations 0–46 years for Sweden. For durations 47, 48, 49, and 50 years there are 100, 99, 98, and 97 naïve errors for the latter country. For Denmark, the error series are much shorter, in particular for longer forecast durations: the length declines regularly

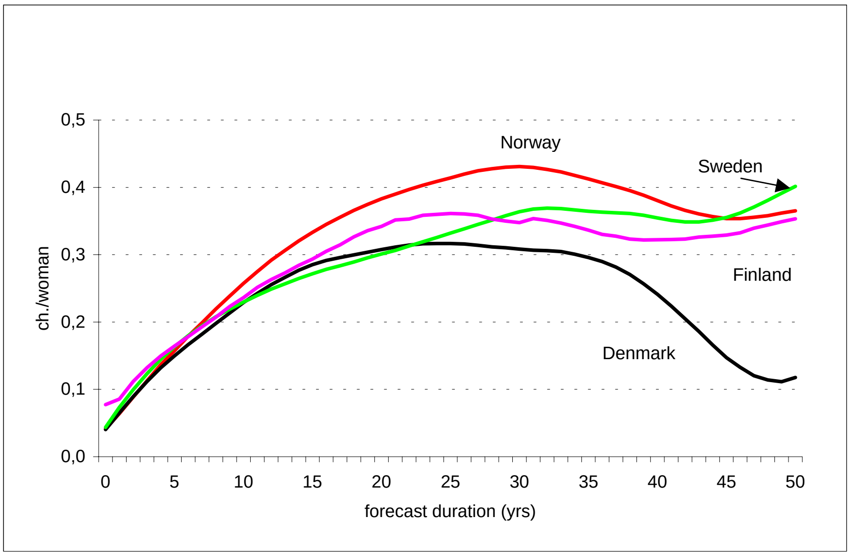

Figure 3. *Standard deviation of signed naïve TFR-errors.*

from 89 errors for duration 0 to 39 for duration 50. Figure 3 plots the standard deviations of these errors. We computed the errors in the log scale, but transformed the results back to the original scale.

The Danish standard deviation falls for forecast durations longer than 20 years, because the long-term errors apply to the period after World War II. Although this period included the baby boom and bust, the volatility in the TFR measured this way was less than that for the whole 20th century. The other three countries show tendencies of the same pattern, but much less extreme than the curve for Denmark. The four curves coincide quite well up to 20 years ahead. As to the long-term behaviour, Figure 3 suggests (ignoring the curve for Denmark) that standard deviations 50 years into the future are around 0.4 children per woman.

We tested an assumption of normally distributed naïve TFR-errors at forecast durations 35 and 50 years ahead for Norway, Sweden and Finland. This assumption had to be rejected (5%) for Norway at a duration of 50 years, but not for the other five cases. Thus for the latter cases we can infer the width of the 67 per cent prediction interval as twice the standard deviation. For the case of Norway 50 years ahead, we have to be more cautious. Not knowing the distribution, Chebyshev's inequality tells us that at least a share equal to $(1 - 1/k^2)$ of probability mass is covered by a $k\sigma$-interval (centered around the mean). Hence the width of the 67-percent prediction interval in this case is estimated as *at least* $\sqrt{3}$ times the standard deviation.

6 Comparison of Model-based and Empirical Prediction Intervals

Figure 4 facilitates a comparison of the expected width of prediction intervals from the three sources: time-series models, errors in historical forecasts, and naïve errors. We have omitted long-term naïve errors for Denmark for reasons mentioned in Section 5.

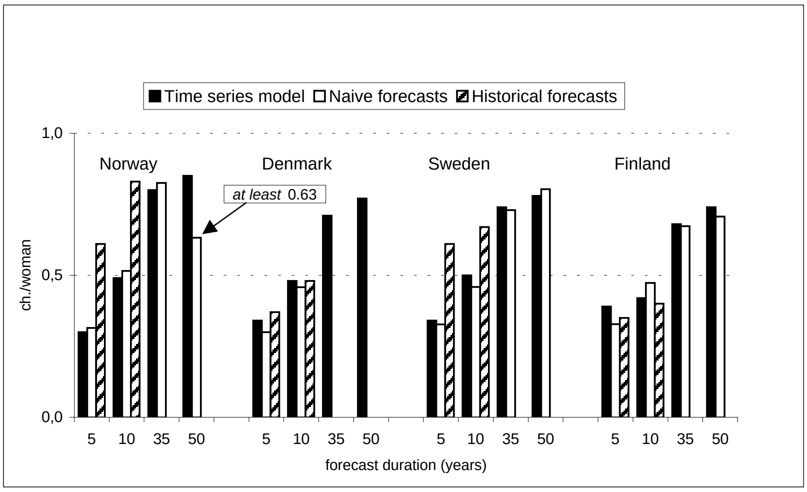

Figure 4. *Width of 67 per cent prediction intervals for the TFR.*

Short term

At a forecast duration of 5 years ahead, the model-based 67 per cent prediction intervals for Denmark and Finland agree quite well with those based on historical errors: in both countries the intervals are 0.3–0.4 children per woman wide. For Norway and Sweden the historical errors are much larger than those stemming from time-series models. One reason may be that the historical errors are not normally distributed. We have relatively few observations (23 for Sweden and 25 for Norway), and therefore the normality assumption could not be tested. In any case, one has to be cautious. The naïve intervals are very close to the model-based ones.

When we consider ten-year ahead forecasts, the number of empirically observed TFR-errors becomes even less (ranging from 17 to 22 for the four countries). With due caution one may conclude that the intervals are roughly 0.5 children per woman wide, primarily based on the time-series models and the naïve errors, but (for the cases of Denmark and Finland) supported by the historical errors.

Medium term

At the medium term, at 35 years into the future, historical errors cannot be used. We note that there is rather good agreement between model-based errors and naïve errors for Norway, Sweden, and Finland. The model-based intervals are approximately 0.7–0.8 children per woman wide for these three countries.

Long term

At 50 years into the future, the agreement between model-based intervals and naïve intervals is rather good for the three countries for which we could do the comparison. Figure 4 suggests 67 per cent intervals that are approximately 0.75–0.85 children per woman wide for Norway, Sweden, and Finland.

7 Conclusion

In this paper we have shown how time series methods may be combined with information from historical and naïve forecasts when one wants to construct long-term prediction intervals for the Total Fertility Rate (TFR). We compared the expected accuracy of three types of TFR-predictions for Norway, Denmark, Sweden, and Finland. Time-series models were fitted to data for the period 1900–2000. These resulted in model-based prediction intervals for the TFR up to 2050. For the short term (5–10 yrs), we analysed empirical forecast errors observed in historical population forecasts prepared by the statistical agencies in these countries after 1969. We also analysed medium-term and long-term (up to 50 years) error patterns based on so-called naïve forecasts, i.e. forecasts that assume that recently observed TFR-levels apply for the future.

For the short term, model-based intervals and those derived from historical errors tend to be of the same order of magnitude, although we have to be cautious with historical errors, because our data set is rather limited. Naïve errors provide useful information both for the short and the long run. Indeed, model-based intervals 50 years ahead agree quite well with naïve errors-based intervals, except for Denmark. For the latter country, the data set did not allow us to compute reliable naïve error patterns for forecast periods beyond 20 years. In general, one may conclude that historical TFR-errors and naïve errors do not indicate that model-based prediction intervals are excessively wide. We found 67 per cent intervals that are approximately 0.5 children per woman wide for a forecast horizon of 10 years, and roughly 0.85 children per woman wide 50 years ahead.

The three types of data sets show us that we have to be modest when we try to predict the accuracy of our TFR-forecasts, at least when we assume

- the same variability in the future TFR as that in the past 100 years (ARCH-model)
- the same short-term forecastability as in the past 30 years (historical errors)
- the same long-term forecastability as in the past 90–150 years (naïve errors).

Some scholars may react that the model-based intervals are too wide and thus reflect too much uncertainty in the TFR-predictions. Our analysis does not give statistical evidence for such a conclusion. If, nonetheless, one believes (for instance, on judgemental grounds) that the expected accuracy of TFR-predictions is better than what time series models, historical forecasts, and naïve forecasts show, the results presented in this paper can serve as a useful benchmark in the discussion.

References

Alho, J. (1990). Stochastic methods in population forecasting. *Internat. J. Forecasting* **6**, 521–530.

Alho, J. (1998). *A stochastic forecast of the population of Finland*. Reviews 1998/4. Helsinki: Statistics Finland.

Armstrong, J. (1985). *Long-range forecasting: From crystal ball to computer*. New York: Wiley (2nd ed.).

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics*, **31**, 307–327.

Boudon, R. (1986). *Theories of social change: A critical appraisal*. Cambridge: Polity Press.

Box, G.E.P. & Jenkins, G.M. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden Day.

Brunborg, H. & Mamelund, S.-E. (1994). *Kohort- og periodefruktbarhet i Norge 1820–1999* (Cohort and Period Fertility for Norway 1820–1993). Report no. 94/27. Oslo: Statistics Norway.

Chesnais, J.-C. (1992). *The Demographic transition: Stages, patterns, and economic implications*. Oxford: Clarendon Press.

Council of Europe (2001). Recent demographic developments in Europe 2001. Strasbourg: Council of Europe Publishing.

De Beer, J. & Alders, M. (1999). Probabilistic population and household forecasts for the Netherlands. Paper ECE-Eurostat Work Session on Demographic Projections, Perugia, Italy, 3–7 May 1999. Internet www.unece.org/stats/documents/1999.05.projections.htm

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica*, **50**, 987–1008.

Henry, L. (1987). Perspectives et prévision. In *Les projections démographiques. Actes du VIIIe Colloque National de Démographie*. Tome 1. Paris: Presses Universitaires de France (Travaux et Documents Cahier no. 116).

Keilman, N. (1990). *Uncertainty in national population forecasting: Issues, backgrounds, analyses, recommendations*. Amsterdam: Swets and Zeitlinger.

Keilman, N., Pham, D.Q. & Hetland, A. (2001). *Norway's uncertain demographic future*. Social and Economic Studies 105. Oslo: Statistics Norway. Internet www.ssb.no/english/subjects/02/03/sos105_en.

Keyfitz, N. (1982). Can knowledge improve forecasts? *Population and Development Review*, **8**, 729–751.
Lee, R. (1974). Forecasting births in post-transition populations: Stochastic renewal with serially correlated fertility. *J. Amer. Statist. Assoc.*, **69**, 607–617.
Lee, R. (1997). History of demography in the U.S. since 1945. In *Les contours de la démographie au seuil du XXIe siècle.*, Eds. J.-C. Chasteland and L. Roussel, pp. 31–56. Paris: Institut National d'Etudes Démographiques, Presses Universitaires de France.
Lee, R. & Tuljapurkar, S. (1994). Stochastic population forecasts for the United States: Beyond high, medium, and low. *J. Amer. Statist. Assoc.*, **89**, 1175–1189.
Maddala, G. (1977). *Econometrics*. Auckland etc., McGraw-Hill International Book Company.
Nagel, E. (1961). *The structure of science: Problems in the logic of scientific explanations*. New York: Harcourt, Brace and World.
Statistics Norway (1994). *Framskriving av folkemengden 1993–2050* (Population Projections 1993–2050). Oslo/Kongsvinger: Statistics Norway.
Statistics Norway (1997). *Framskriving av folkemengden 1996–2050* (Population Projections 1996–2050). Oslo/Kongsvinger: Statistics Norway.
Statistics Norway (1999). Fortsatt befolkningsvekst. Befolkningsframskrivinger. Nasjonal og regionale tall, 1999–2050 (Population growth continues. Population forecasts: national and regional results 1999–2050). *Dagens statistikk*, 17.11.1999. Internet `http://www.ssb.no/folkfram`.
Texmon, I. (1992). Norske befolkningsframskrivinger 1969–1990 (Norwegian population projections 1969–1990). In *Mennesker og Modeller*, Eds. O. Ljones, B. Moen, and L. Østby, pp. 285–311. Oslo/Kongsvinger: Statistics Norway.
Turpeinen, O. (1979). Fertility and mortality in Finland since 1750. *Population Studies*, **33**(1), 101–114.
United Nations. Internet `http://www.un.org/esa/population/publications/wpp2002/`wpp2002annextables.PDF.
Van de Kaa, D. (1996). Anchored narratives: The story and findings of half a century of research into the determinants of fertility. *Population Studies*, **50**(3), 389–432.

Résumé

Nous avons construit un modèle de séries temporelles du type ARCH pour calculer des intervalles de prédiction pour l'Indice Synthétique de Fécondité (ISF) pour la Norvège, la Suède, la Finlande, et le Danemark jusqu'à l'année 2050. Pour le court terme (5–10 ans dans le futur), on compare les erreurs attendues pour l'ISF avec les erreurs calculées dans des prévisions démographiques historiques, préparées par des bureaux de statistique dans ces pays depuis 1969. Les erreurs à moyen terme et long terme (jusqu'à 50 ans dans le futur), sont comparées avec des structures d'erreur fondée sur des prévisions dites "naïves"— c'est-à-dire, des prévisions qui supposent que le niveau d'ISF observé pour une période récente est valable aussi pour le futur.

À court terme, nous trouvons que les intervalles de prédiction calculés par le modèle de séries temporelles et ceux dérivés des erreurs historiques sont du même ordre d'amplitude. Cependant, il faut être prudent, car la collecte des données de base historiques est limitée. Les erreurs "naïves" fournissent de l'information utile pour le court terme et le long terme. En effet, des intervalles de prédiction fondés sur des erreurs naïves à 50 ans dans le futur se comparent très bien avec des intervalles fondés sur le modèle de séries temporelles, sauf pour le Danemark. Pour ce pays, les données de base ne nous permettent pas de calculer des intervalles "naïfs" pour des périodes de prévision au-delà de 20 ans. En général, on peut conclure que les erreurs historiques et les erreurs naïves ne montrent pas que les intervalles de prédiction fondés sur des modèles de séries temporelles du type ARCH sont excessivement larges. Nous avons constaté que les intervalles à 67pour cent de l'ISF ont une amplitude d'environ 0.5 enfants par femme à l'horizon de 10 ans, et approximativement 0.85 enfants par femme à 50 ans.

8 Appendix: Data Sources

Observed TFR

The Council of Europe (2001) publishes annual TFR-values for Norway, Denmark, Sweden, and Finland for the years 1960–2000.
Chesnais (1992, Appendix 2) tabulates annual observed TFR-values for Sweden since 1855 and for Denmark starting in 1911.
Brunborg & Mamelund (1994, Table 1a) give annual TFR-values starting in 1845.
Turpeinen (1979, p.112) gives Finnish TFR-values for each year in the period 1776–1925. Those for later years are taken from the Statistical Yearbook of Finland.

For Denmark, Chesnais tabulates annual TFR-values starting in 1911, and isolated data points for 1903 and 1908. On the other hand, his Table A1.5 lists pre-1911 values for the Danish Crude Birth Rate (CBR). We have estimated TFR-values for the years 1900–1910 based on a linear regression between TFR and CBR. A plot of TFR and CBR showed a near linear relationship for the years 1911–1940. We have used this relationship and "backcasted" the Danish TFR for the years 1900–

1910. This way, we obtained the values and 95 per cent prediction intervals as given in Table A1. The TFR in Figure 1 and its first difference do not reveal any apparent anomalies in the predictions for the years 1900–1910.

Table A1

Backcasts and 95% prediction interval bounds for the TFR of Denmark, 1900–1910.

	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910
Prediction	4.07	4.07	3.99	3.91	3.94	3.86	3.88	3.83	3.89	3.83	3.71
L95	4.01	4.01	3.93	3.85	3.88	3.80	3.81	3.77	3.83	3.77	3.65
U95	4.12	4.12	4.04	3.96	3.99	3.91	3.93	3.88	3.94	3.88	3.77

Chesnais (1992, p.545) reports a TFR in 1908 equal to 3.84, and in 1903 equal to 4.04. Our estimate for 1908 is close to Chesnais' value, while the 1903-estimate is somewhat lower.

TFR Predictions in Official Population Forecasts

Norway

We have updated the data originally assembled by Texmon (1992), who collected, among others, TFR errors for the forecasts of 1969–1987 during the years 1969–1989. Updates were made based on Statistics Norway (1994, 1997, 1999).

Denmark

Various issues of "Befolkning og valg" published by Danmarks Statistik, viz.1983:15, 1984:20, 1985:12, 1986:16, 1988:13, 1989:12, 1990:17, 1992:2, 1992:16, 1993:13, 1994:12, 1995:15, 1996:14, 1997:12, 1998:15, 1999:11, 2000:11, 2001:11. For earlier forecasts we consulted "Statistiske Efterretninger" issues 1978 A39, 1980 A7, 1981 A40, and 1982 A33, and "Statistiske Undersøgelser" nr. 33 published in 1975.

Sweden

Jan Qvist and Åke Nilsson (personal communications) of Statistics Sweden provided us with assumed TFR-values in forecasts prepared by Statistics Sweden starting in 1972. Most of these values can also be found in the following reports published by Statistics Sweden:
"Information i prognosfrågor" issues 1972:5, 1973:6, 1974:7, 1975:6, 1976:3, 1978:5, 1983:2. Later forecasts have been published in "Demografiska rapporter" 1986, 1989:1, 1991:1, 1994:3, and 2000:1. Forecasts with base years 1997, 1998, 1999 have been published at Statistics Sweden's home page, and in Statistical Yearbooks of Sweden issues 1998, 1999, 2000.

Finland

From Ossi Honkanen of Statistics Finland (personal communication) we received assumed TFR-values in official Finnish population forecasts starting in 1972. Most of these values were annual. In some cases, we interpolated linearly between values given for distinct calendar years. Most of these values can also be found in the following reports published by Statistics Finland:
"Statistisk rapport" issues VÄ 1972:7, VÄ 1973:6, VÄ 1975:12, VÄ 1978:17, VÄ 1982:5, VÄ 1982:5, and VÄ 1985:10; "Statistiska meddelanden" 49:1972; "Statistical surveys" issues 52:1974, 64:1979, 70:1983; "Muistio" 91:1984; "SVT/OSF Population" issues 1989:3, 1992:6, 1993:10, 1995:9, and 1998:6.

[*Received December 2002, accepted November 2003*]