



Demographic Research a free, expedited, online journal
of peer-reviewed research and commentary
in the population sciences published by the
Max Planck Institute for Demographic Research
Doberaner Strasse 114 · D-18057 Rostock · GERMANY
www.demographic-research.org

DEMOGRAPHIC RESEARCH
VOLUME 6, ARTICLE 15
PUBLISHED 28 MAY 2002
www.demographic-research.org

**Why population forecasts should be
probabilistic - illustrated by the case of
Norway**

Nico Keilman

Dinh Quang Pham

Arve Hetland

© 2002 Max-Planck-Gesellschaft.

Table of Contents

1	The need for probabilistic population forecasts	410
2	Three methods for probabilistic population forecasts	412
3	Application: A probabilistic population forecast for Norway 1996-2050	416
3.1	Method	417
3.1.1	<i>Fertility</i>	419
3.1.2	<i>Mortality</i>	424
3.1.3	<i>International migration</i>	430
3.2	Probabilistic population forecasts for Norway	431
3.2.1	<i>Total population</i>	431
3.2.2	<i>Age pyramids</i>	432
3.2.3	<i>Old age dependency ratio</i>	438
3.2.4	<i>Errors in historical age structure forecasts</i>	439
3.3	Sensitivity analysis: The importance of various types of variance	440
4	The use of stochastic population forecasts	442
5	Conclusions	445
6	Acknowledgements	446
	Notes	447
	References	449

Why population forecasts should be probabilistic - illustrated by the case of Norway

Nico Keilman¹

Dinh Quang Pham²

Arve Hetland³

Abstract

Deterministic population forecasts do not give an appropriate indication of forecast uncertainty. Forecasts should be probabilistic, so that their expected accuracy can be assessed. We review three main methods to compute probabilistic forecasts, namely time series extrapolation, analysis of historical forecast errors, and expert judgement. We illustrate, by the case of Norway up to 2050, how elements of these three methods can be combined when computing prediction intervals for a population's future size and age-sex composition. We show the relative importance for prediction intervals of various sources of variance, and compare our results with those of the official population forecast computed by Statistics Norway.

¹ Department of Economics, University of Oslo, P.O. Box 1095 Blindern, N-0317 Oslo, Norway. Phone (+47) 22 85 51 28 or 22 85 51 27 fax (+47) 22 85 50 35. E-mail: n.w.keilman@econ.uio.no. Also Statistics Norway, P.O. Box 8131 Dep, N-0033 Oslo, Norway.

² Statistics Norway, P.O. Box 8131 Dep, N-0033 Oslo, Norway. Phone (+47) 21 09 46 86. E-mail: dqp@ssb.no

³ Statistics Norway, P.O. Box 8131 Dep, N-0033 Oslo, Norway. Phone (+47) 21 09 44 37. E-mail: ahe@ssb.no.

1. The need for probabilistic population forecasts

The demographic future of any country is uncertain. There is not just one possible future, but many. Some of these are more probable than others. Therefore, an exploration of a country's demographic future, for instance its population size in 2020, should include two elements: first, a *range* of possible outcomes, and second, a *probability* attached to that range. Together, these two constitute a prediction interval for the variable concerned. Such a prediction interval expresses the expected accuracy of the population forecast. In other words, it quantifies forecast uncertainty. Why is this important?

Statistical agencies traditionally deal with the uncertainty of forecasting population variables by producing two or more forecasts of fertility or mortality (or both), and then calculating a range of forecasts. For instance, Statistics Norway expects the number of children aged 6–12 in Norway in 2010 to be between 401,000 and 436,000, depending on whether fertility is low or high — that is, on whether women will have an average of 1.5 or 2.1 children, respectively, in 2010 (Statistics Norway 1999).

There are two drawbacks connected to this traditional approach. First, no probability is attached to the intervals. Yet, those who are planning provisions for education will find it useful to know whether the likelihood of this scenario (between 401,000 and 436,000 children aged 6-12 in Norway in 2010) is roughly 30%, 60%, or even 90%. A 30 per cent chance implies that the user should be prepared for surprises. Therefore, he should include much more flexibility in the planning process than a 90 per cent chance would imply. Second, the use of high and low variants is unrealistic and inconsistent from a statistical point of view (Lee 1999, Alho 1998). In the high variant, fertility is assumed to be high in *every* year of the forecast period. Similarly, when fertility is low in one year, it is 100 per cent certain that it will be low in the following years, too. Things are even worse when two or more mortality variants are formulated, in addition to the fertility variants. In that case, a forecast variant with high population growth results from combining high fertility with low mortality (high life expectancy), and vice versa for low population growth. This means that in the high growth variant, *any* year in which fertility is high, life expectancy is high as well. In other words, one assumes perfect correlation between fertility and mortality, and perfect serial correlation for each of the two components. Assumptions of this kind are unrealistic, and, moreover, they cause inconsistencies: two variants that are extreme for one variable need not be extreme for another variable. To illustrate this with Statistics Norway's most recent population forecast, consider the future number of elderly in Norway, and the associated Old Age Dependency Ratio (OADR), that is the number of

elderly as a ratio of the number in working ages. The legal pensionable age is 67 years in Norway. In 2050, the population aged 67 and over will number between 911,000 and 1,244,000, depending on low or high population growth (Statistics Norway 1999). However, the corresponding OADR-values are 0.364 for low population growth, and 0.360 for high population growth. While there is a considerable gap between the absolute numbers of elderly in the two variants, the relative numbers, as a proportion of the population aged 20-66, are almost indistinguishable. The interval for the absolute number thus reflects uncertainty in some sense, but the OADR-interval for the same variant pair suggests practically no uncertainty. The reason for this inconsistency is that the population in working ages in this forecast is perfectly correlated with the number of elderly. A probabilistic forecast, which we propose as an alternative to traditional deterministic forecasting, does not necessarily assume perfect correlation between these two population groups. The example for Norway that we present in Section 3.2.3 results in a two-thirds OADR-prediction interval (that is, odds of two against one) in 2050 that stretches from 0.31 to 0.44.

In the discussion above, we assumed that the forecast variants presented by Statistics Norway are to be interpreted as uncertainty variants, i.e. variants that intend to show forecast uncertainty around a central path - the Medium variant. This is in line with Statistics Norway's own interpretation (Statistics Norway 2002, page 30). It should be noted that other agencies might have a different interpretation of forecast variants, namely that they represent alternative futures, without any uncertainty interpretation connected to them. In that case the purpose is to present meaningful demographic future developments, based upon different sociological, biomedical, and political deliberations.

In the recent past, demographers and statisticians developed methods for making probabilistic population forecasts, the aim of which is to calculate prediction intervals for every variable of interest. The tradition goes back to Leo Törnqvist (1949), who probably was the first one to integrate probabilistic thinking in population forecasting. Recent examples include population forecasts for the United States, Austria, Germany, Finland, and the Netherlands (Lee and Tuljapurkar 1994; Hanika et al 1997; Lutz and Scherbov 1998a, 1998b; Alho 1998; Alders and De Beer 1998; De Beer and Alders 1999), for major world regions (Lutz et al 1996, 2001), and for all countries in the world (NRC 2000). In this paper, we shall limit ourselves to national populations, but many of the arguments apply to multiregional forecasts as well.

We shall briefly review three main methods currently in use for making probabilistic population forecasts: time series extrapolation, analysis of historical forecast errors, and expert judgement. We illustrate, by the case of Norway up to 2050,

how elements of these three methods can be combined when computing prediction intervals for a population's future size and age-sex composition. Our contribution to the existing literature is twofold. First, we illustrate how historical forecast errors for fertility, mortality, and age structures can be used to assess whether the short-term prediction intervals for the current forecast are reasonable. Second, we show the relative importance for prediction intervals of residual variance in time series models, and of estimation variance.

2. Three methods for probabilistic population forecasts

Probabilistic population forecasting uses the cohort component method. Instead of one set of parameters for fertility, mortality, and migration, as in the traditional deterministic method (or three, when a high, a medium, and a low forecast variant are computed), a probabilistic forecast requires that one specify the joint statistical distribution of all input parameters. The large number of parameters (35 fertility rates, 200 death rates, and some 140 parameters for net migration *for each forecast year*) necessitates simplifying assumptions. (Even with age groups and time intervals equal to five years, a forecast for a period of fifty years, say, still requires that one specifies the joint distribution of $(7+40+28)*10=750$ parameters.) First, one splits up the joint distribution into a number of smaller distributions with fewer variables, assuming that the components of fertility, mortality, and migration are independent. Second, one focuses on the distribution of a few summary indicators, for instance the total fertility rate, the life expectancy at birth, and level of net-immigration. This implies that one ignores the statistical distributions of the detailed parameters (age specific rates). In this case one assumes that the base population is perfectly known, for instance from a recent census. This is a realistic assumption for most developed countries with good data. When there are doubts about the data quality, one should model the base population statistically, and explicitly consider covariances with all components of change.

In probabilistic forecasts, four types of correlations are important: correlation across components, across age, across sexes, and across time (serial correlation).

In a Western country such as Norway, there is little or no reason to assume correlation between the components of fertility, mortality, and migration. (In developing countries, disasters and catastrophes may have an impact both on mortality, fertility, and migration, and a correlation between the three components cannot be excluded. There may also be a positive correlation between the levels of immigration and childbearing in Western countries with extremely high immigration from

developing countries.) There is no empirical evidence of such correlation (Lee and Tuljapurkar 1994; Keilman 1997). Therefore, in the stochastic forecasts of the US, Austria, Germany, Finland, and the Netherlands, the three components were assumed independent (Lee and Tuljapurkar 1994; Hanika et al 1997; Lutz and Scherbov 1998a, 1998b; Alho 1998; De Beer and Alders 1999). The forecasts for the US, Austria, Germany, and Finland were based on numbers of net-migrants, so that any correlation between immigration and emigration could be ignored. For the Netherlands, emigration flows for migrants born in certain countries were linked to up to the stock of the foreign-born population already residing in the Netherlands. As a consequence, immigration and emigration were indirectly positively correlated.

Correlation across time is important for each component. Levels of fertility and mortality change only slowly over time. Thus, when fertility or mortality is high one year, a high level the next year is also likely. This implies a strong, but not perfect serial correlation for these two components. International migration is much more volatile, but economic, legal, political, and social conditions stretching over several years affect migration flows to a certain extent, and some degree of serial correlation should be expected. In the probabilistic forecasts of Lee and Tuljapurkar (1994), Alho (1998), and De Beer and Alders (1999) these correlation patterns were estimated based on time series models. Hanika et al. (1997) and Lutz et al. (1998a, 1998b) assumed perfect autocorrelation for the summary parameters (total fertility, life expectancy, and net migration). Lee (1999) states that this assumption underestimates uncertainty. While it is unclear whether this generally is the case, it is true when the real process is a first order autoregressive process, a random walk, and probably also when it is a moving average process. In recent work, Lutz, Sanderson, and Scherbov relaxed the assumption of perfect autocorrelation (Lutz et al. 2001).

Men and women display similar behaviour regarding mortality and migration. This gives rise to a positive correlation across the sexes for these two components. Lee and Tulapurkar (1994) and De Beer and Alders (1999) assume perfect positive correlation across the sexes for mortality. This overestimates uncertainty in the future number of elderly irrespective of sex. Alho (1998) used an empirical correlation of 0.80 between male and female historical mortality in Finland. For migration, the empirical correlation across sexes was 0.91. De Beer and Alders assumed perfect correlation, while Lee and Tuljapurkar modelled net immigration deterministically. As a result, prediction intervals in the forecasts for the population in migration sensitive ages (up to age 50, roughly) are too wide in the Dutch case, and too narrow for the US.

Correlation across age is strong for each component. The annual age profiles of fertility (by mother's age), mortality, and migration are highly regular. This implies that

age-specific rates and numbers for these three components are strongly positively correlated in a given year. Indeed, Alho (1998) found correlations between neighbouring ages for age-specific fertility and mortality equal to 0.89 or higher. Therefore, to simplify the computations, perfect correlation has been assumed in most applications, so that only the level of the age profile was subject to stochastic variation. Yet Alho assumed an auto regressive error structure with a one-year lag (i.e. an AR(1) process) across ages, instead of perfect correlation, because the correlation between ages further apart falls rapidly. For instance, the correlation between neighbouring ages for historical fertility in Finland was equal to 0.963, which implies a correlation of $(0.963)^{10}=0.686$ between ages ten years apart.

Three main methods are in use for computing probabilistic forecasts of the summary indicators: time series extrapolation, expert judgement, and extrapolation of historical forecast errors. Time series methods and expert judgement result in the distribution of the parameter in question around its expected value. In contrast, an extrapolation of empirical errors gives the distribution centred around zero (assuming an expected error equal to zero), and the expected value of the population variable is taken from a deterministic forecast computed in the traditional manner.

Time series methods are based on the assumption that historical values of the variable of interest have been generated by means of a statistical model, which also holds for the future. A widely used method is that of Autoregressive Integrated Moving Average (ARIMA)-models. Time series models were developed for short horizons. When applied to long-run population forecasting, the point forecast and the prediction intervals may become unrealistic (Sanderson 1995). Judgmental methods (see below) can be applied to correct or constrain such unreasonable predictions (Lee 1993; Tuljapurkar 1996).

Expert judgement can be used when expected values and corresponding prediction intervals are hard to obtain by formal methods. In demographic forecasting, the method has been pioneered by Lutz and colleagues (Lutz et al. 1996; Hanika et al. 1997; Lutz and Scherbov 1998a, 1998b). A group of experts is asked to indicate the probability that a summary parameter, such as the TFR, falls within a certain pre-specified range for some target year, for instance the range determined by the high and the low variant of an independently prepared population forecast. The subjective probability distributions obtained this way from a number of experts are combined in order to reduce individual bias. A major weakness of this approach, at least based upon the experiences from other disciplines, is that experts often are too confident, i.e. that they tend to attach a too high probability to a given interval (Armstrong 1985). A second problem is that an expert would have problems with sensibly guessing whether a certain interval corresponds to

probability bounds with 90 per cent coverage versus 95 per cent or 99 per cent (Lee 1999).

Extrapolation of empirical errors requires observed errors from historical forecasts. Formal or informal methods may be used to predict the errors for the current forecast. Keyfitz (1981) and Stoto (1983) were among the first to use this approach in demographic forecasting. They assessed the accuracy of historical forecasts for population growth rates. The Panel on Population Projections of the US National Research Council (NRC 2000) elaborated further on this idea and developed a statistical model for the uncertainty around total population in UN-forecasts for all countries of the world. Others have investigated and modelled the accuracy of predicted TFR, life expectancy, immigration levels, and age structures (Keilman 1997; De Beer 1997). There are two important problems. First, time series of historical errors are usually rather short, as forecasts prepared in the 1960s or earlier generally were poorly documented. Second, extrapolation is often difficult because errors may have diminished over successive forecast rounds as a result of better forecasting methods.

The three approaches are complementary, and elements of all three are often combined. Lee and Tuljapurkar (1994) modelled the time series of the level parameter for US-fertility obtained by means of the Lee-Carter method as an ARIMA(1,0,1)-process with a constrained mean, subjectively chosen equal to 2.1. Alho (1998) compared prediction intervals for the TFR in Finland obtained by means of an ARIMA(1,1,0)-model with those that result from the errors of so-called naïve forecasts, i.e. forecasts that assume that the current TFR-level is a reasonable forecast of the future TFR. A similar method was employed for mortality. He also combined errors of naïve forecasts with time series analysis and expert judgement in his crude assessments of forecast uncertainty for twelve large world regions (Alho 1997). De Beer and Alders (1999) modelled the life expectancy of the Netherlands as a random walk with drift, and compared the resulting prediction intervals with those obtained from a time series of historical forecast errors for the life expectancy. Lutz et al. (2001) chose a certain level for the variance in the TFR in a target year. The variance was larger for regions with high fertility than for low fertility regions. As to mortality, they generally assumed that life expectancies would increase between zero and four years with 80 per cent probability. These subjectively chosen distributions were combined with a moving average time series process for the error in the TFR or the life expectancy increase. At the same time, the authors aimed at producing prediction intervals that were at least as large as those published by the NRC-panel for major world regions (NRC 2000).

Clearly, subjective choices are abundant in empirical studies of this kind. They are made not only in the expert judgement method, but also in time series analysis, in

particular when choosing a certain form of the extrapolation model, or the length of the historical series. These choices often have important consequences for the shape of the prediction intervals.

Irrespective of the method that is used to determine the prediction intervals for all future fertility, mortality and migration parameters, the next step is to apply these to the base population in order to compute prediction intervals for future population size and age pyramids. There are two common approaches to obtain such intervals: an analytical approach and a simulation approach.

The *analytical approach* is based on a stochastic cohort component model, in which the statistical distributions for the fertility, mortality, and migration parameters are transformed into statistical distributions for the size of the population and its age-sex structure. Alho and Spencer (1985) and Cohen (1986) employ such an analytical approach, but they need strong assumptions. Lee and Tuljapurkar (1994) give approximate expressions for the second moments of the distributions.

The *simulation approach* avoids the simplifying assumptions and the approximations of the analytical approach. The idea is to compute several hundreds or thousands of forecast variants ("sample paths") based on input parameter values for fertility, mortality, and migration that are randomly drawn from their respective distributions. The forecast results are stored in a database, and the α -per cent prediction interval for a certain variable ranges from the $(100-\alpha)/2$ -th percentile to the $(100+\alpha)/2$ -th percentile of that variable's values. Early contributions based on the idea of simulation are those by Keyfitz (1985), Pflaumer (1986, 1988), and Kuijsten (1988).

3. Application: A probabilistic population forecast for Norway 1996-2050

We shall illustrate the methods discussed above for the case of Norway. Our probabilistic population forecast is based on a combination of time series extrapolation, inspection of observed errors in historical forecasts, and the use of expert judgement. Most attention was given to time series extrapolation, in order to obtain a correct initial specification of covariances and autocorrelations. For mortality, this resulted in acceptable prediction intervals. However, for fertility and international migration, long-term intervals were too wide. Therefore, these intervals were reduced in an *ad hoc* manner based on subjective decisions. Observed forecast errors for fertility, mortality and the age structure were used to check the plausibility of the short-term intervals.

3.1 Method

We simulated the future population of Norway by age and sex five thousand times. Each simulation run covered the years 1996-2050. Starting point was the observed age pyramid as of 1 January 1996. Each simulation was based on random draws from prediction intervals for summary parameters of fertility (total fertility rate, mean age at childbearing, and variance in that age), mortality (life expectancy at birth for men and women), and international migration (numbers of immigrants and emigrants). We assumed stochastic age patterns for fertility and mortality, and deterministic age patterns for immigration and emigration. The prediction intervals for age-specific fertility, mortality, immigration, and emigration were derived from those for the summary parameters based on a common approach. A short general description is given below. Details can be found at [http://www.ssb.no/emner/02/03/sos105_\(Keilman et al. 2001\)](http://www.ssb.no/emner/02/03/sos105_(Keilman et al. 2001)).

We assumed that the four components of population change were independent, and accounted for correlation across the sexes for mortality (Section 3.1.2) and for migration (Section 3.1.3). A possible correlation between immigration and emigration (for instance caused by return migration), or one between migration of men and women (family migration, family reunification) was ignored. Extensive empirical checks did not result in clearly interpretable patterns. We estimated time series models for one or more log-transformed summary parameters for each component, and used Monte Carlo simulation to obtain future values of those parameters. The time series models were estimated such that they predicted expected values for the TFR, the life expectancy, the number of immigrants, and the number of emigrants, which agreed closely with those of the Medium variant of Statistics Norway's 1996-based population forecast. The parameters were assumed to follow a multivariate normal distribution in the future, with a known vector of expected values and known covariance matrix (Note 1). Multivariate normally distributed numbers were drawn from each distribution. Repeated Monte Carlo simulation resulted in five thousand independent sample paths for each age-specific parameter.

The computations consisted of the following steps.

1. We assembled time series of fertility rates by mother's age, mortality rates by age and sex, and numbers of immigrants and emigrants by age and sex for a certain base period. For each year and each age, we assumed that a Poisson process had generated the events of childbearing and death. The empirical fertility or mortality rate is the estimate of the parameter of the Poisson model (Note 2). Therefore, these empirical rates have estimation errors (sampling variance).

2. We estimated age schedules for all three components in each observation year: a Gamma curve for fertility, a Heligman-Pollard curve for mortality, and a Rogers-Castro curve for migration. We computed summary indicators for each component: TFR, mean age at childbearing, and variance in the childbearing age for fertility, life expectancy at birth (by sex) for mortality, and numbers of immigrants and emigrants (by sex) for international migration. The summary indicators for fertility and mortality have estimation errors that result from two sources: 1. the age-specific rates have estimation errors; 2. the fit of the Gamma curve and the Heligman-Pollard curve is not perfect.

3. We constructed a time series model for each of the summary indicators. The coefficients of the time series model have estimation errors, and there is a residual variance. For fertility and mortality, the input data for the time series models, i.e. the summary indicators mentioned in step 2, are not observed, but estimated. Hence the estimation errors for the summary indicators are reflected in the estimation variance of the coefficients and in the residual variance.

4. We used the time series models to extrapolate the summary indicators.

5. We broke the future summary indicators down by age to obtain the future age and sex specific parameters for the cohort component model.

In summary, this approach results in four main sources of uncertainty attached to future birth and death rates:

1. sampling variance in the historical age-specific rates;
2. estimation variance in the parameter estimates of the age pattern curves;
3. residual variance in each time series model;
4. estimation variance in the coefficient estimates of each time series model.

For migration we worked with absolute numbers, and hence only sources 2-4 were relevant. In Section 3.3 we shall analyse the relative importance of these four variance sources for prediction intervals around various forecast results.

All forecast results are computed based on a number of linked stochastic models, with specific assumptions concerning various types of variances for future fertility, mortality, and migration. The expected values of the distributions correspond to those assumed by Statistics Norway in its official population forecast. We have regarded those official results as benchmark values, and focussed on the width of prediction intervals.

3.1.1 Fertility

We assumed that age-specific fertility rates would follow a Gamma curve for each year between 1900 and 1995. The Gamma curve is a mathematical function that has proven its success in reproducing the skew pattern of age-specific fertility. Keilman and Pham (2000) show that the fit to Norwegian data was good after 1940, and excellent from 1980 onwards. The curve has four parameters: the TFR, the mean age at childbearing, the variance in that age, and the minimum reproductive age. These parameters were estimated for each year in the period 1945-1995. We used weighted least squares (WLS), giving relatively little weight to ages at which sampling variance of fertility rates was high. Next we fitted the following multivariate ARIMA (1,1,0)-model to the time series of the log-transformed estimates of the first three parameters during the years 1945-1995:

$$Z_t = \Phi Z_{t-1} + \varepsilon_t, \quad (1)$$

where Z_t is a column vector with first differences in the three parameter estimates in logarithmic form, Φ is a 3x3-matrix of coefficients, and ε_t is a multivariate normal column vector with zero expectation and constant covariance matrix Σ_ε . WLS estimation gave little weight to years in which the (co-)variances for the three parameter estimates were large. Starting from parameter estimates for 1994 and 1995, we used the ARIMA model to simulate the future values of the three parameters in question for the period 1996-2050.

Estimates for the minimum age fell below 14 in 1975, and decreased further to reach zero in 1991. We assumed that it would remain zero in the future. Predicted rates for ages lower than 15 are extremely low, in spite of this unrealistic assumption. The reason is that the TFR is low, and the mean age is high. Therefore we could ignore fertility rates for women younger than age 15. Note that Thompson et al. (1989) found a similar (even steeper) drop in the minimum age for US fertility.

For each of the 5,000 simulations, one value of the $\hat{\Phi}$ -matrix, and 55 values (one for each year) of the ε -vector were drawn from their respective multivariate normal distributions. This resulted in reasonable medians for the three parameters in 2050, but the expected TFR in 2050 was rather high: 2.21 children per woman (the median TFR was 1.86). This is much higher than the fertility assumption in the Medium variant of Statistics Norway's population forecast of 1999: 1.8 from 2010 onwards. Concerning the prediction intervals, we checked both the short-term and the long-term patterns.

3.1.1.1 Long-term prediction intervals

The long-term prediction intervals were excessively wide, in particular that for the TFR. The odds were two against one that the TFR in 2050 would lay between 1.1 and 3.3 children per woman, while the 95 per cent prediction interval would range from 0.6 to 6.1 children per woman in that year. The reason for these wide intervals is that the model is not able to explain the baby boom of the 1950s and early 1960s. We also experimented with other estimation periods, namely 1960-1995 and 1975-1995 (see Keilman and Pham (2000) for details). We concluded that opting for the period 1945-1995 strikes a good balance between a high residual variance (1960-1995), and an imprecise estimate of the autoregressive coefficient (1975-1995), so we will proceed with the 1945-1995-based estimates.

Although the expected probability that the TFR will exceed 6.1 children per woman in 2050 was only 2.5 per cent, this cannot be considered as realistic. The model produces reasonable results (an upper 95%-bound lower than four children, say) up to the years 2020-2030, but not for later years. We have therefore assumed upper and lower limits to future TFR-levels in Norway.

To find realistic limits, we carried out a number of experimental simulations with the ARIMA model. We defined upper limits for the period TFR of 2, 3, 4, and 5 children per woman, and lower limits equal to 0, 0.25, 0.5, 0.75, and 1 child per woman on average. In case a predicted TFR, in any year, fell outside the range defined by the upper and lower limits, the entire sample path was rejected, and new simulations were generated until we had obtained 5,000 sample paths with admissible values. An alternative model is to reject not the whole sample path, but only the TFR-value that crosses the limit, and redraw the TFR until an admissible value has been found. While we do not have any clear preference, this alternative method requires many more Monte Carlo attempts, because extreme TFR-values result more often from extreme $\hat{\Phi}$ -draws than from extreme $\hat{\mathcal{E}}$ -draws.

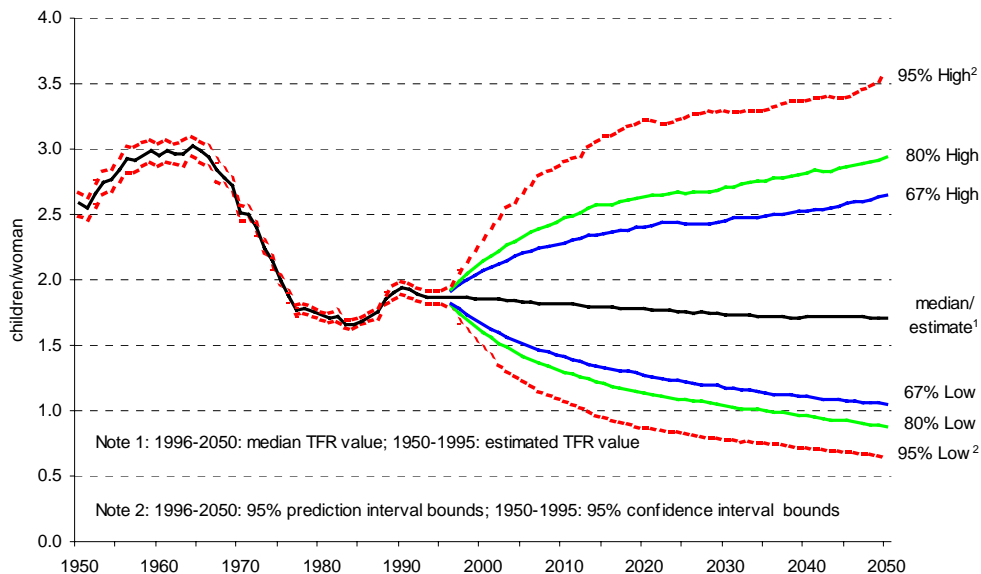
As could be expected, the lower limit had very little effect on the TFR-developments. The upper limit had strong impact on the long run, in particular for the median values and for the bounds of the 95 per cent prediction intervals. For instance, with a maximum TFR set to 2 children per woman (and minimum TFR equal to zero), both the upper and the lower bound of the 95 per cent prediction interval and the median value fell almost continuously over the period 1998-2045. With a maximum TFR equal to 3 children per woman, the upper 95 per cent bound was almost constant at 2.6 children per woman in the period 2013-2037, while the median fell continuously. Only when choosing an upper limit equal to 4 children per woman, we obtained

reasonable patterns, i.e. an expected value and a median in 2050 close to the official value of 1.8, and non-decreasing values for the upper 95 per cent bound (Note 3).

Based on these experiments, we decided to restrict the period TFR to values between 0.5 and 4.0 children per woman. We also introduced restrictions on other parameters. 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. Even then, a mean age at childbearing higher than 50 years, or a variance in the age at childbearing of 400, is clearly unrealistic. At the same time it is unrealistic to assume that teenage fertility has become so important that the mean age will fall below 20. Thus, draws that resulted in TFR values outside the interval [0.5, 4], mean age values outside [20, 50], or age variances outside [10, 250] were rejected, and new values were drawn until 5,000 sets with admissible values were obtained. We also restricted the elements of the ARIMA coefficient matrix to the interval [-1, +1]. Figure 1 gives the results for the TFR. The median TFR-predictions fall slightly (1.86 children per woman in 1995, 1.71 in 2050), as did the expected value (1.86 in 2010 and 1.83 in 2050). The model predicts a small increase in the mean age (to 30.2 years in 2050, up from 28.8 years in 1995). These trends are smooth extrapolations of current fertility developments in Norway, and they are in accordance with the fertility assumptions in the official population forecasts (Statistics Norway 1999).

It should be stressed that our approach in which we reject sample paths that fall outside a certain pre-specified range, differs from the approach pioneered by Lutz, Sanderson, and Scherbov that we labelled as "expert judgement" in Section 2. In the latter, a certain TFR-range (for example between 1 and 3 children per woman) is assumed to cover a certain part of the predictive distribution (for instance 80 per cent), such that the extreme values correspond with appropriate lower and upper percentiles (for example the 10th and the 90th percentile). In our approach, we simply limit the whole predictive interval to a certain range (between 0.5 and 4 children per woman). This can be interpreted as a "100 per cent prediction interval", so to speak. Although still difficult and quite arbitrary, we find it easier to select TFR-bounds that define completely unrealistic, or impossible, TFR ranges (larger than 4, or below 0.5), than bounds for which there still may be a 10 per cent probability that the future TFR will cross them (larger than 3, below 1). Clearly, this is merely a practical matter. Generally speaking, our method of introducing bounds and impossible ranges is not principally different from the "expert judgement" approach.

Figure 1: Estimates, confidence intervals, and prediction intervals for the TFR, 1950-2050. TFR-predictions restricted to [0.5, 4.0]



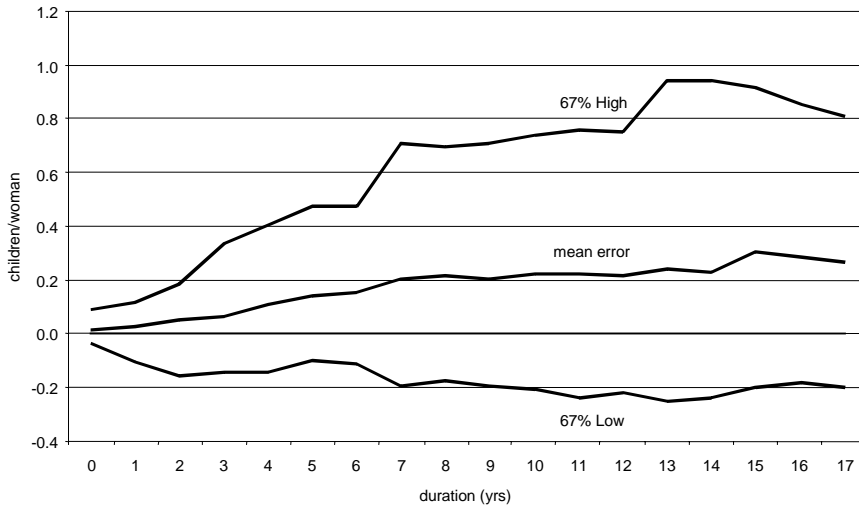
3.1.1.2 Short-term check

We have used the historical errors in twelve TFR-forecasts that Statistics Norway has published between 1969 and 1996 as an independent check against the expected short-term errors in Figure 1. For each historical forecast, we compared assumed TFR-values from the jump-off year until 1999 with observed values, updating the data originally assembled by Texmon (1992). The error in the TFR was defined as the assumed minus the observed value. Many of the twelve forecasts contained two or more variants. These variants got equal weight in the error analysis, mainly because Statistics Norway did not recommend one specific variant as being the more realistic one; see Keilman and Pham (2000) for a discussion. This resulted in 28 series of TFR errors, and each series was ordered by forecast duration (the jump-off year was defined as duration 0). For each duration, the errors were ordered from low to high. Finally, we selected (by linear interpolation, if necessary) two error values, such that one-sixth of the errors were

lower, and one-sixth were higher than these values. The result is interpreted as an empirical duration-specific 67 per cent interval.

Figure 2 shows the 67 per cent interval for the TFR errors, together with the mean error. The lower bound is approximately -0.2 , close to zero. The error was larger than -0.2 in five out of every six cases; most often, it was positive. 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 distribution of the error is skewed to the left, indicating that large errors were more frequent than small ones. In the jump-off year the 67 per cent interval is 0.13 children wide, and it is 1.1 children wide at a forecast duration of 15 years. This is only slightly wider than the 67 per cent prediction interval in Figure 1 after 15 years (0.9 children per woman). The general agreement between widths of the two intervals is rather good.

Figure 2: Empirical errors in historical TFR forecasts. Base years 1969-1996, period 1969-1999



3.1.1.3 Comparison with official population forecast

The High-Low range for the TFR in Statistics Norway’s 1999-based population forecast is [1.5, 2.1] children per woman for the years 2010 and later (Statistics Norway 1999). A comparison with Figure 1 tells us that the estimated coverage probability for that range is only 46 per cent in 2010, falling to 31 per cent in 2030 and 24 per cent in 2050. Thus, assuming that our fertility model is correct, the official TFR-forecasts have limited predictive validity.

3.1.2 Mortality

For mortality, the predictions consisted of three steps.

- First, life tables for the period 1945-1995 resulted in annual values of the life expectancy at birth for men and women (e_t^M, e_t^F). The two time series of life expectancies were modelled as a multivariate ARIMA (2,0,0) model:

$$\begin{bmatrix} \ln(e_t^M) \\ \ln(e_t^F) \end{bmatrix} = \begin{bmatrix} K^M \\ K^F \end{bmatrix} + \begin{bmatrix} \phi_1^M & 0 \\ 0 & \phi_1^F \end{bmatrix} \begin{bmatrix} \ln(e_{t-1}^M) \\ \ln(e_{t-1}^F) \end{bmatrix} + \begin{bmatrix} \phi_2^M & 0 \\ 0 & \phi_2^F \end{bmatrix} \begin{bmatrix} \ln(e_{t-2}^M) \\ \ln(e_{t-2}^F) \end{bmatrix} + \begin{bmatrix} \varepsilon_t^M \\ \varepsilon_t^F \end{bmatrix} \quad (2)$$

We estimated the model by OLS, assuming that the residual vector follows a multivariate normal distribution with zero mean and covariance matrix Σ_ε (Note 4). Monte Carlo simulation of the coefficient matrix and the residual vector resulted in prediction intervals for male and female life expectancies for the period 1996-2050 (see Figures 3 and 4 to be discussed later). The constant vector K was adjusted in such a way that the expected life expectancy values in 2050 coincided with target values assumed by Statistics Norway in its official population forecast issued in 1999 (80 years for men, 84.5 years for women).

- Second, we assumed that for each year the age pattern of mortality could be described by means of a Heligman-Pollard (H-P) curve. We used the following version, see Heligman and Pollard (1980):

$$q_x = A^{(x+B)^C} + D \exp\{-E(\log x - \log F)^2\} + \frac{GH^x}{1 + GH^x}, \quad x = 1, 2, 3, \dots, 97$$

where q_x is the one-year probability of dying at age x , and the eight parameters $A-H$ are to be estimated from the data. (Ages 0, 98, and 99+ were treated differently. See <http://www.ssb.no/emner/02/03/sos105>.) We used data on Norwegian deaths for men and women in one-year age groups for each year in the period from 1945 to 1995 and estimated the parameters by means of Relative Least Squares. The fit for women at high ages was unsatisfactory. In many years, predicted rates for ages 80 or higher were lower than empirical rates. Therefore the third part of the curve was replaced by a pure Gompertz curve (GH^x). The estimation resulted in two multivariate time series of parameter estimates, one for each sex. The (log-transformed) time series for the years 1945-1995 were modelled by means of a multivariate ARIMA (1,1,0) model. Monte Carlo simulation resulted in five thousand multivariate sample paths for the parameters for the period 1996-2050. The H-P curve was used to transform the parameter predictions back into future age-specific death probabilities. For each year in the prediction period, a life table calculation summarized those probabilities into the life expectancy at birth. These life expectancies were assembled in a look-up table, together with the underlying life table.

- In the third and final step, life tables from the look-up table were assigned to life expectancy values for each simulation run as predicted in the first step, by matching life expectancy values from the first and the second step, controlling for calendar year and sex. The result was an age pattern for male and female mortality for each year in the future and each simulation run.

This three-step procedure does not guarantee that the historical autocorrelation for age-specific mortality is reproduced in the probabilistic forecasts. We checked the one step ahead autocorrelation in the death probabilities at selected ages, and found that the historical correlations for men were rather well preserved in the simulations, while the autocorrelation values for females generally were low compared to historical values. This means that simulated death probabilities for women vary somewhat stronger over time than historical probabilities did. But the historical autocorrelation structure of the *life expectancy* was preserved by the use of model (2).

3.1.2.1 Long-term prediction intervals

Prediction intervals for the life expectancy at birth by sex are shown in Figures 3 and 4. These are obtained based on 5,000 simulations, in which the two coefficient matrices $\hat{\Phi}_1$ and $\hat{\Phi}_2$ of expression (2) and the $\hat{\epsilon}$ – vector were treated as multivariate normal variables with known means and covariances.

The 95 per cent intervals were 10.9 (men) and 12.2 (women) years wide in 2050. The annual correlation between male and female life expectancy in the 5,000 simulations turned out to be 0.50 in 2050, decreasing from 0.63 in 1996, to 0.56 in 2010, and 0.51 in 2030. The average cross-sex correlation for the period 1945-1995, estimated from the residual covariance matrix Σ_{ϵ} , is 0.65. These values are somewhat lower than the 80 per cent correlation between male and female age-specific death rates found by Alho (1998) for mortality in Finland in the period 1900-1994, see Section 2.

Figure 3: Male life expectancy at birth. Observed 1950-1995. Predicted 1996-2050

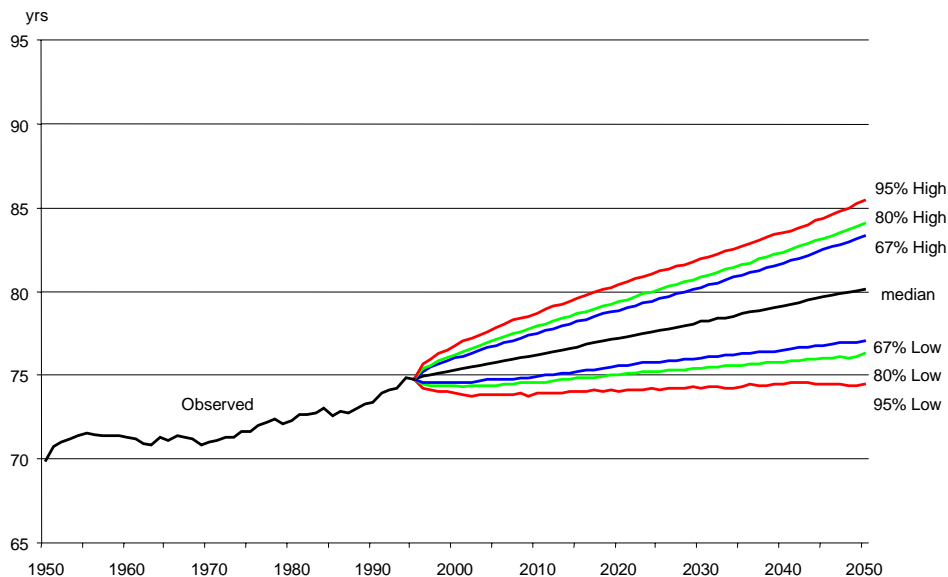
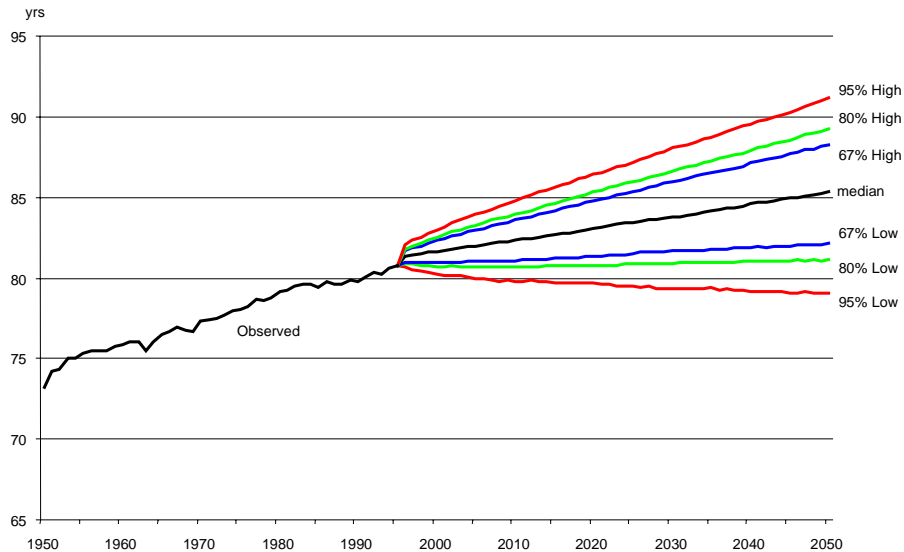


Figure 4: Female life expectancy at birth. Observed 1950-1995. Predicted 1996-2050



Our results indicate similar long-term uncertainty for the life expectancy as that estimated by Alho (2001) for Finland and De Beer and Alders (1999) for the Netherlands. The Finnish 95 per cent prediction intervals are 15.3 (men) and 10.2 (women) years wide in 2050 (Note 5). De Beer and Alders found a 95 per cent interval of 12 years in 2050, both for men and for women.

Intervals presented by Tuljapurkar et al (2000) for the G7 countries (Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States) are much smaller than ours are. Their 95 per cent intervals of combined-sex life expectancy at birth in 2050 range from a minimum of 3.3 years for Canada to a maximum of 8.9 years for the UK (Note 6). These intervals result from a random walk with drift model for the single parameter of a Lee-Carter model for age-specific mortality for the two sexes combined. The authors used an abridged life table with five-year age classes up to 80-84. Ages 85 years and higher were lumped into one age class (except for Japan). The

age and sex aggregation, which reduces random fluctuations, may have caused these relatively narrow intervals (Note 7).

3.1.2.2 Short-term check

Figures 5 and 6 summarize empirical errors for extrapolated life expectancy of men and women. Eleven forecasts with base years between 1969 and 1996 were analysed in a way similar to that described in Section 3.1.1 for the TFR. The figures show that mean errors in the life expectancy at birth become increasingly more negative for longer forecast durations. Hence Norwegian forecasters have been too pessimistic, in that they assumed too low life expectancy values in past forecasts. After 15 years, life expectancies were too low by two years on average for men and somewhat more for women. The unexpectedly strong improvement in female mortality since 1969 causes the empirical 67% interval for women to be relatively wide. The intervals do not widen with forecast duration. This is explained by the extrapolation method for life expectancy. For forecasts made between 1969 and 1982, no improvement in mortality was foreseen, and life expectancy was kept constant at its most recently observed value (Note 8). In reality, life expectancy increased more or less regularly. As a consequence, life expectancy errors show a time path for subsequent forecasts, which runs parallel to that of the mean error.

The intervals are somewhat irregular, because we had a small data set. At durations of 10-15 years, the 67 per cent interval is approximately 0.7 years wide for men and roughly 1.8 years for women. Thus prediction intervals in Figures 3 and 4 are wider than historical intervals, and this would imply that one should restrict the life expectancy predictions to narrower bounds. On the other hand, these bounds should have been wider, because (for reasons explained in the technical documentation) we decided to ignore the following error sources: estimation errors for the constant vector of the ARIMA model, sample variation in the historical age-specific rates, and estimation errors in H-P parameters. For these reasons we accepted the life expectancy predictions in Figures 3 and 4. An independent check of the prediction intervals for the age pyramid of the elderly population in 2010 against corresponding observed errors 15 years ahead in historical forecasts supports that decision, see Section 3.2.2 below.

Figure 5: Empirical errors in historical life expectancy forecasts for men. Base years 1969-1996, period 1969-1999

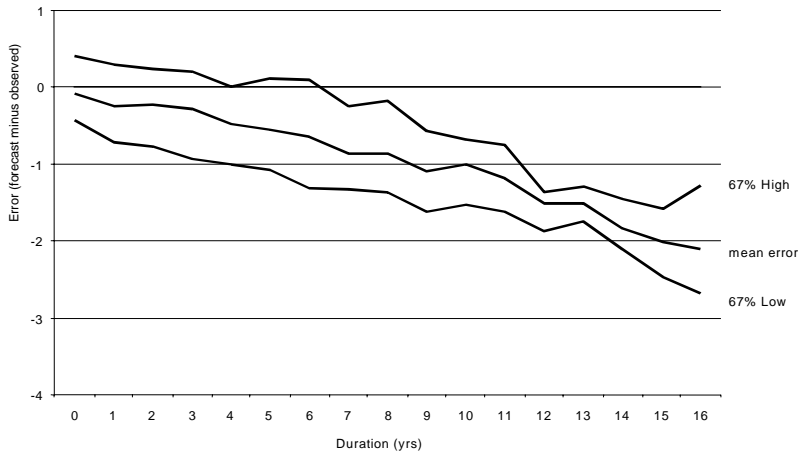
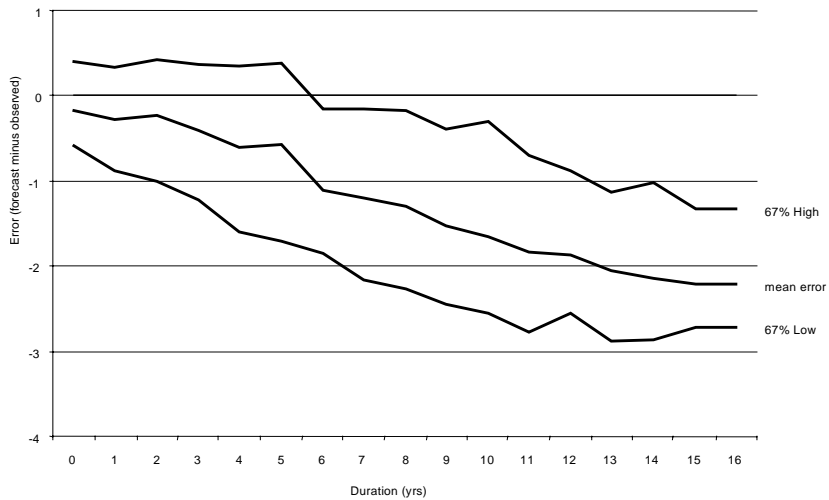


Figure 6: Empirical errors in historical life expectancy forecasts for women. Base years 1969-1996, period 1969-1999



Excess mortality of men in Norway, compared to women, is likely to decrease further in the future. In the period 1991-1995, the life expectancy at birth for men was 6 years lower than that for women, a slight reduction compared to the difference of 6.9 years that was observed in the mid-1980s. In 2010, the gap is reduced to an expected 5.7 years (the 95 per cent prediction interval for this difference equals (3.3, 8.0)), and it drops further to 5.2 years in 2030 (1.1, 9.4) and 4.7 years in 2050 (-1.3, 10.7). The lower bound of the 95 per cent interval falls below zero in 2040, implying a 2.5 per cent chance for higher female than male mortality.

3.1.2.3 Comparison with official population forecast

Statistics Norway assumed in its 1999-based population forecast that the life expectancy of men would increase from 75.5 in 1999 to between 77 and 83 years in 2050, and that of women from 81.2 to 81.5-87.5 years. Assuming that our model is correct, we predict that the expected accuracy of the official life expectancy forecasts in 2050 is 64 per cent for men and 62 per cent for women. For both sexes, it is much lower in the beginning of this century (in 2010, 36 per cent for men and 42 per cent for women) but it increases steadily to the 2050-values.

3.1.3 International migration

For international migration, we adopted a rather simple approach. We distinguished between immigration and emigration flows, and modelled log-transformed annual numbers for each flow, observed for the period 1967-1997, as univariate time series models: ARMA (1,1) including a constant term for immigration, and a random walk with drift for emigration. Residuals for both models were assumed normal. Stochastic simulation resulted in five thousand sample paths with annual numbers of immigrants and emigrants for the period 1996-2050. The constant terms were adjusted such that the models predicted target levels of annual immigration and emigration in 2000 as assumed by Statistics Norway. To avoid excessively wide prediction intervals, predicted migration levels were kept constant beginning in 2000. (This admittedly *ad hoc* adjustment corresponds to the migration approach by Statistics Norway, in which extrapolated migration levels are kept constant after five years.) Predicted numbers of immigrants and emigrants were broken down by sex based on randomly drawn shares for men and women. (We used a normal approximation to the binomial distribution,

with parameters p and σ for men equal to the observed values in the period 1967-1997.) This resulted in four flows: immigration and emigration, both for men and for women. Age-specific numbers for each of the four flows were obtained based on a Rogers-Castro (R-C) curve with six parameters (Rogers and Castro 1986). The retirement peak in observed immigration and emigration turned out to be negligible in the Norwegian data. The R-C curve was fitted to age-specific shares for each of the four flows in each calendar year. The resulting time series of R-C parameters were predicted into the future by means of simple extrapolation procedures. Predicted R-C parameters were used to transform predicted flows back into predicted numbers of immigrants and emigrants by sex and age. The age pattern for each flow was the same across simulation runs.

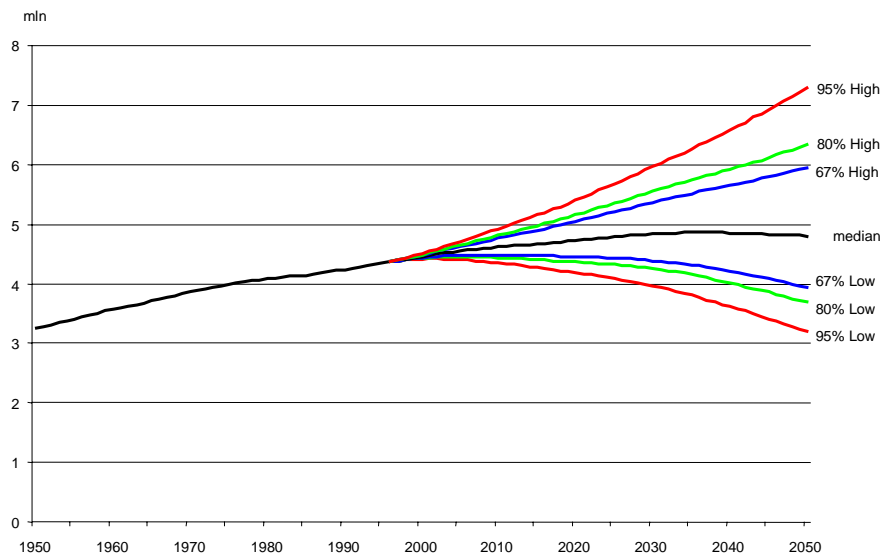
3.2 Probabilistic population forecasts for Norway

3.2.1 Total population

Figure 7 shows prediction intervals for total population size. The intervals widen rapidly when we look further into the future. The odds are two against one that the Norwegian population, now 4.5 million, will number between 3.9 and 6 million in 2050. Compared to the median forecast of 4.8 million in 2050, this two-thirds prediction interval is 43 per cent wide. There is a clear trade-off between greater accuracy (larger odds) and higher precision (narrower intervals). For instance, odds of 19 to one (95 per cent probability) are attached to an interval between 3.2 and 7.2 million in 2050. This interval is twice as wide as the 67 per cent interval: 88 per cent, compared to the median forecast.

Statistics Norway's most recent population forecast was published in 1999, and predicted between 4.2 and 6.3 inhabitants in Norway in 2050, depending on low or high population growth (Statistics Norway 1999). According to our simulations, the expected probability that this will be the case is 60 per cent. The expected short-run accuracy of the official forecast cannot be assessed this way, because the base years of the official forecast and our probabilistic one are different (1999 and 1996). However, Statistics Norway's previous forecast had 1996 as its base year (Statistics Norway 1997). For that forecast the expected accuracy of the high-low interval was 77 per cent in 2000, 66 per cent in 2010, and 56 per cent in 2050. Thus, the expected accuracy of the official forecast of total population size is somewhat higher than two-thirds on the short run, and a little lower than that on the long run.

Figure 7: Population size. Observed 1950-1996. Predicted 1997-2050



3.2.2 Age pyramids

Figures 8-11 present age pyramids for the years 1996, 2010, 2030, and 2050. The age pattern of uncertainty is very marked: prediction intervals are wide for young age groups, and narrow for the elderly. This reflects the fact that fertility and mortality have very different impact on the age structure, with international migration taking an intermediate position. Note also that the interval for the age group 0-4 grows rapidly between 2010 and 2030, because most of the parents of the youngest age group in 2030 themselves were born after 1996. As a result, intervals in 2030 under the age of 20 are so wide that the forecast is not very informative. In 2050 this is the case for virtually all age groups.

Figure 8: Observed population by age and sex. 1996

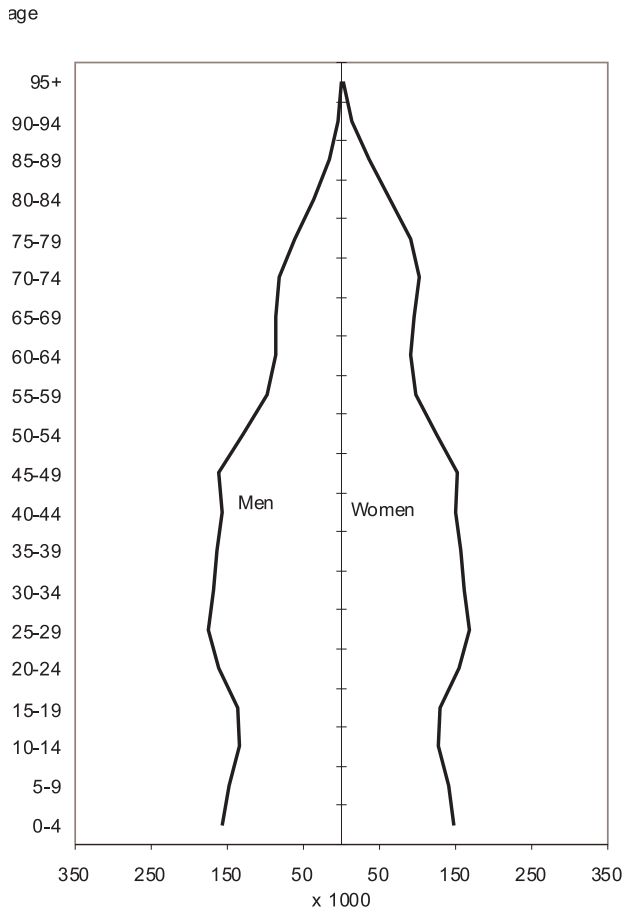


Figure 9: Prediction intervals for population by age and sex. 2010

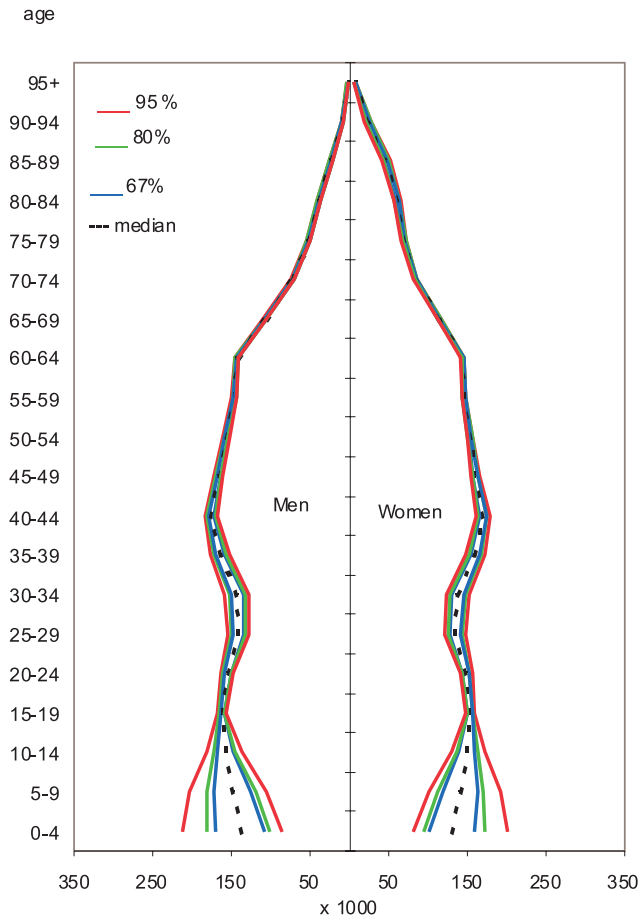


Figure 10: Prediction intervals for population by age and sex. 2030

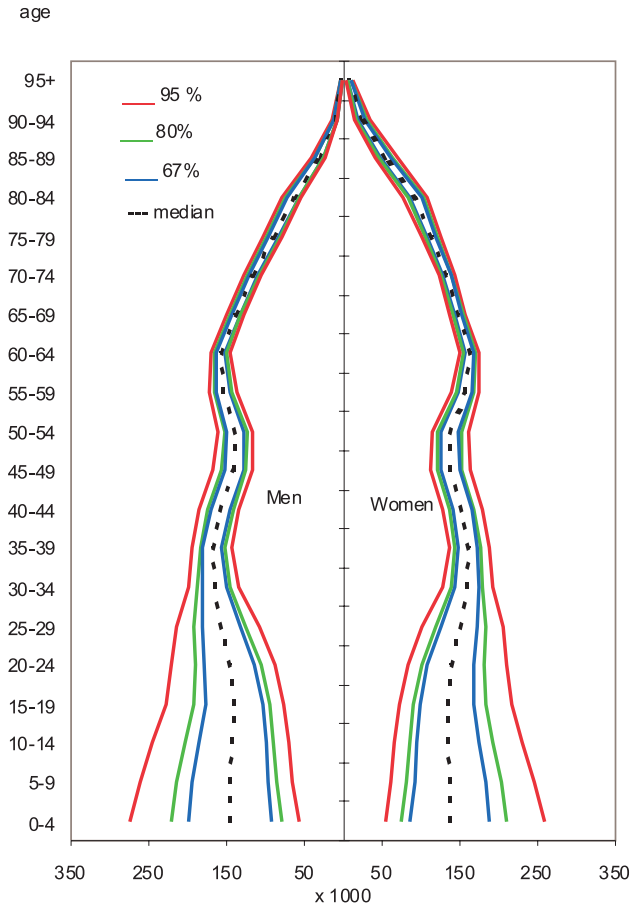
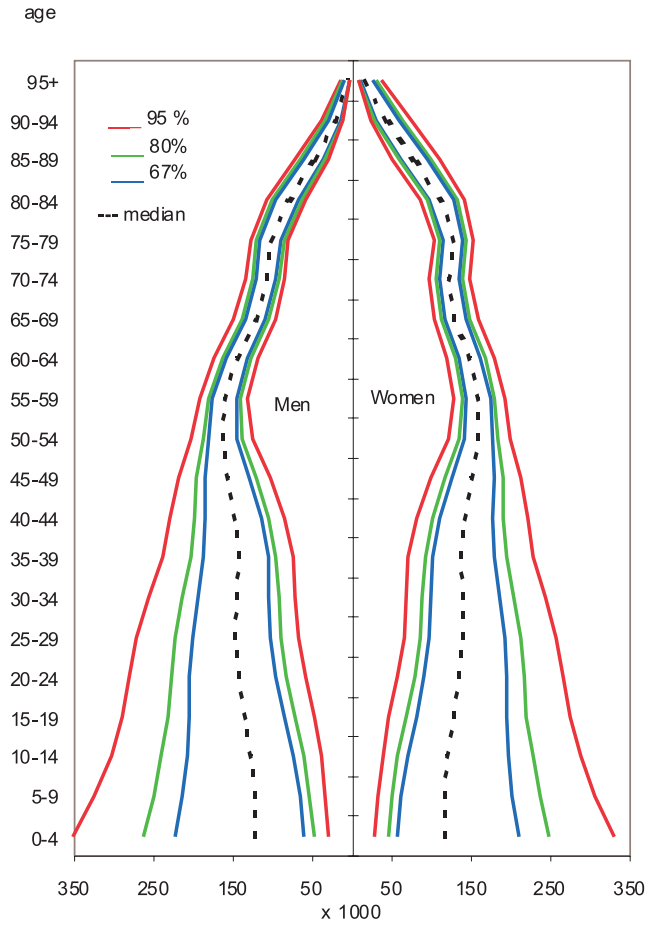


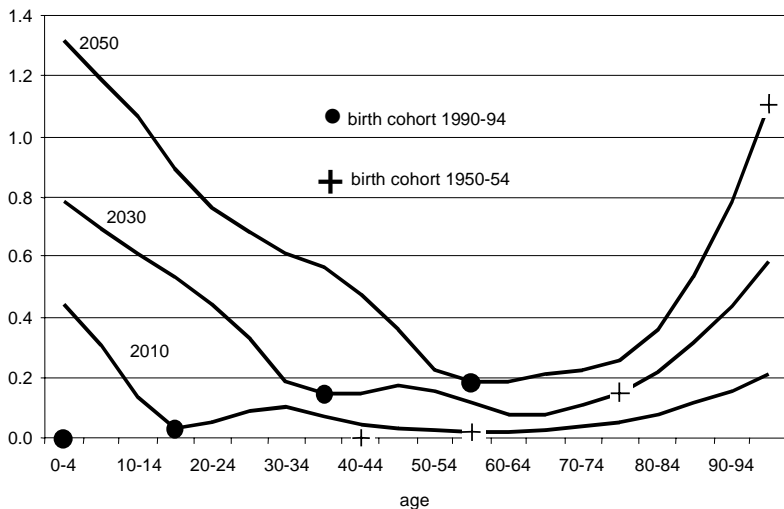
Figure 11: Prediction intervals for population by age and sex. 2050



For purposes of comparison, it is instructive to inspect the width of the relative intervals, i.e. the intervals as a ratio of the median. Figure 12 illustrates that for men over 95, relative uncertainty is almost as large as it is for the youngest age groups. The pattern for women is similar, the largest differences occurring for elderly women. For instance, at age 95+, women have relative 67 per cent prediction intervals that are 0.27, 0.61, and 0.97 per cent times the corresponding median values in 2010, 2030, and 2050.

The lines for 2010, 2030, and 2050 indicate relative uncertainty *cross-sectionally*. They suggest that uncertainty first decreases from birth to middle ages (up to an age equal to the forecast duration), and that it increases thereafter. However, the cross-sectional patterns do not reflect uncertainty over the life course. The relative intervals for the birth cohort 1950-1954 illustrate that the age gradient for the elderly is much steeper than what the cross-sectional pattern indicates. The plot for birth cohort 1990-1994 shows that uncertainty increases during childhood as well.

Figure 12: Relative width of 67% prediction interval, men

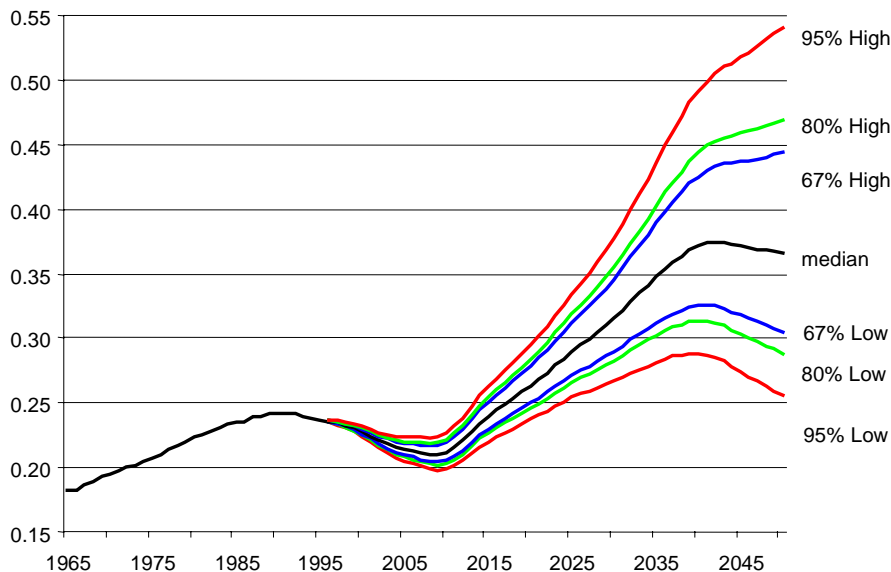


3.2.3 Old age dependency ratio

Continuous ageing is almost certain, at least until around 2040. In that year, the odds are two against one that the Norwegian old age dependency ratio (OADR) will be between 0.33 and 0.43, i.e. at least 10 points higher than today's value of 0.23, see Figure 13. The probability of a ratio in 2040 that is lower than today's is close to zero.

As mentioned in Section 1, the OADR-range in the official high growth and low growth variants by Statistics Norway stretches from 0.360 to 0.364 in 2050, which is only 1.7 per cent of the official medium forecast. This narrow interval suggests much less uncertainty for the OADR than a true probabilistic forecast does. For instance, Figure 13 shows relative prediction intervals in 2050 that are 38 (67%), 49 (80%), and 78 (95%) per cent wide, relative to the median OADR-forecast.

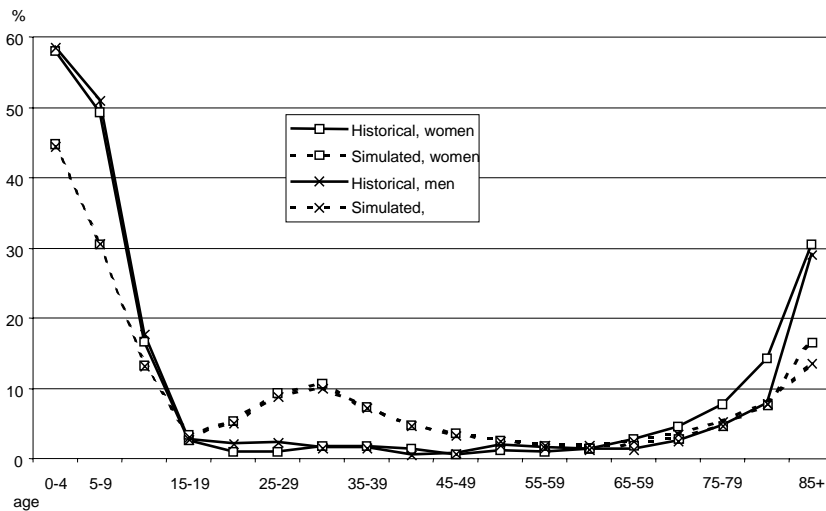
Figure 13: Old age dependency ratio



3.2.4 Errors in historical age structure forecasts

How do the prediction intervals around the age structure compare with empirical intervals computed on the basis of historical forecasts? We have assembled data on observed forecast errors in the age structure (five-year age groups, by sex) in the forecasts published by Statistics Norway with base years between 1969 and 1985, extending Texmon's database mentioned in Section 3.1.1. We have restricted ourselves to a forecast duration of 15 years. Forecast variants were given equal weight. For each age group up to 80-84 we had fifteen error values, for the open age group 85 and over twelve. Figure 14 plots the width of the historical two-thirds intervals for men and women, and compares these with the simulated prediction intervals for the year 2010.

Figure 14: *Relative width of 67 per cent interval. Forecast duration = 15 years*



There is rather close agreement between the historical pattern and the simulated one. In both cases, errors are high for young ages, and moderate for the elderly. Historical variation is relatively large for ages 0-9 and 80 and over. For the youngest age groups

this is explained by the sharp fall in the TFR in the 1970s and the modest increase at the end of the 1980s. Historical variation is relatively large for the elderly because of the strong increase in survival chances since 1970, in particular for women. In the simulated fertility and mortality trajectories up to 2010, such sudden developments are less likely (but not excluded). For ages 25-39, prediction intervals are wider than one would expect based on historical forecast errors. This reflects the fact that the variation in year-to-year immigration has increased since the beginning of the 1990s. It is unlikely that the smooth developments that were observed until the mid-1980s will repeat themselves in the period up to 2010.

3.3 Sensitivity analysis: The importance of various types of variance

The ultimate purpose was to generate probabilistic population forecasts. Therefore, much attention was given to an appropriate quantification of uncertainty. In Section 3.1, we mentioned four main sources of uncertainty attached to future birth and death rates: 1. sampling variance in the historical age-specific rates; 2. estimation variance in the parameter estimates of the age pattern curves; 3. residual variance in each time series model; and 4. estimation variance in the coefficient estimates of each time series model. In principle, all four should be taken into account. We were able to accomplish this for fertility. For the life expectancy at birth, we only included sources 3 and 4. Source 1 was ignored because the variances involved were low; see Section 3.1.2. For age specific mortality, we omitted sources 1 and 2 when we estimated the Heligman-Pollard (H-P) curve and the multivariate time series model for the H-P parameters, respectively. Experimental calculations resulted in extremely unstable estimates when sources 1 and 2 were included. For migration, source 1 was irrelevant, and we ignored source 2 for reasons of simplicity.

Thus, sources 3 and 4 have been included in the simulations for all four components. Table 1 shows how the prediction intervals of selected variables change when only residual variance is accounted for in the time series prediction, assuming that the estimated coefficients are the real ones, and thus ignoring source 4. Note that the effects for population size and OADR result from ignoring estimation variance in time series models for all four components (fertility, mortality, immigration, and emigration) jointly.

Table 1: *Width of 95 per cent prediction interval in 2050 for selected variables, and width reduction when parameter estimation variance in time series models is ignored*

	Interval width	Width reduction (%)
Total population size	4.2 million	4
Old age dependency ratio	0.285	10
Total fertility rate ¹	5.6 children/woman	8
Mean age at childbearing ¹	21.7 years	18
Life expectancy at birth, men	10.9 years	13
Life expectancy at birth, women	12.2 years	24
Number of immigrants	21,900	28
Number of emigrants	19,100	1

Note 1: Unconstrained fertility predictions. For constrained predictions, the reduction was negligible.

For some variables, such as the female life expectancy and the number of immigrants, ignoring estimation variance in the time series coefficients reduces the 95 per cent interval in 2050 by one-fourth. The impact on the life expectancy for men is less than that for women. This is explained by the fit of model (2), in particular the estimated standard error for the residual compared to that for the ϕ_1 -estimate. (The ϕ_2 -estimates are so small that their role in the sensitivity analysis is negligible.) For men, the standard error of the ϕ_1 -estimate is 20.3 times that of the residual, whereas for women the corresponding ratio is 26.9. Thus, relative to the residual's standard error, the standard error of the ϕ_1 -estimate is less important for men than for women. A similar explanation can be given for the fact that emigration is much less sensitive than immigration. For the emigration model (random walk with drift), the standard error ratio for the constant (the drift estimate) relative to the residual is 0.16, whereas it is no less than 3.8 for the immigration model (ARMA (1,1) with constant term). An additional reason is that the immigration model contains three parameters to be estimated, and the emigration model only one.

The general conclusion is that estimation variance for a certain summary indicator *may* have considerable impact on the prediction intervals. The impact depends on two factors: the fit of the time series model concerned (in particular the residual variance compared to the estimation variance for the model coefficients), and the relative importance of the summary indicator for the population variable (total population, age groups, dependency ratio, etc.). Therefore estimation variance cannot be ignored.

For fertility, we also investigated the importance of variance sources 1 and 2. The most important effect of ignoring rate variances (source 1) is that the estimated parameter variance for the TFR is reduced rather strongly (Keilman and Pham 2000). This leads to narrow intervals for many predicted age-specific rates. At prime childbearing ages, the intervals around the predicted age-specific rates are reduced by nearly one-half. For younger and older ages the differences are much smaller.

Not only the birth rates, but also the three summary parameters are estimates, each with their own variance. We estimated a multivariate ARIMA (1,1,0) model for the three log-transformed fertility parameters (TFR, mean age at childbearing, and variance in that age), ignoring variance sources 1 and 2. We found almost no change in the predicted TFR, but the 95 per cent prediction interval in 2050 became smaller by 1.7 children. This reduction is to be compared with the original interval of [0.5, 6.1] and a width of 5.6 children per woman, because we used the unrestricted time series model here. The consequence of ignoring variance sources 1 and 2 thus is that we are too optimistic about the future TFR, in the sense that the prediction intervals are too narrow. (In our case the impact would have been much less, because we restricted our long-run TFR predictions to the interval [0.5, 4.0].)

4. The use of stochastic population forecasts

Population forecasts are widely used in various planning situations, such as for schooling, health care, and pension systems. In the very short run, the uncertainty expressed by stochastic forecasts is limited (Note 9), and a user who has a five-year planning horizon, say, may safely use a deterministic forecast computed in the traditional way. In such cases, the point forecast (expected value) is likely to be of more interest than the prediction intervals. In the long run however, the expected error in the forecast results becomes increasingly important, and planners who are interested in the age structure of the population 30 years or more into the future, should take uncertainty seriously. We shall give two examples.

Health care. In a recent analysis of the future demand for hospital beds in Norway, Paulsen et al. (1999) concluded that in 2050, the population would need between 2.7 and 4.0 million person days in hospital annually. This implies an increase by 45-79 per cent compared with the current number of 1.9 million person days in hospital per year. The results were obtained on the basis of the low and the high population forecast of Statistics Norway, combined with certain assumptions regarding health parameters such as the mean duration of hospital stay for future patients, and the mean number of

patients compared to the whole population. Many of these parameters are age-specific, and the increase in the demand for hospital beds in this analysis is entirely a consequence of population ageing (the authors assumed a slight decrease in the average duration in hospital per patient). The demand interval in 2050 of 1.3 million person days is 38 per cent of the medium forecast. The majority of the patients are 65 or older. For this age group, two-thirds prediction intervals in 2050 increase sharply with age, from 21 per cent for the 65-69s, to 26-36 per cent for the 80-84s, and 65-78 per cent for the 90-94s, see Figure 11. This suggests that the accuracy of the 2.7-4.0 interval probably is lower than two-thirds, perhaps even lower than one-half. For 2010 and 2030, the authors obtain intervals of 13 and 25 per cent of the medium forecast, respectively. Thus the Paulsen et al. projection results should be treated with great caution, in particular their long-run findings.

Public pensions. Fredriksen (1998) describes a microsimulation model that is able to simulate, among others, future public pension benefits and contributions of the Norwegian population. Norwegian public pensions are of the Pay As You Go (PAYG)-type. In one set of simulations starting in 1993, the author shows that the contribution rate will rise from 16 to 23 per cent in 2030. The simulations are based upon a number of demographic and non-demographic assumptions (labour market participation, disability, earnings, retirement). Next, he analyses the consequences for the contribution rate of various alternative assumptions. The rate turns out to be between 21 and 25 per cent in 2030, depending on high or low disability risks in the population, or high or low labour force participation among women. The four-percentage point difference implies a relative interval of 16 per cent of the reference value for the contribution rate. Because of the PAYG nature of the pensions, it is relevant to inspect the Old Age Dependency Ratio. (In case a fund based pension system is considered, instead of a PAYG-system, life course uncertainty of the type illustrated in Figure 12 would be relevant.) Figure 13 shows a relative two-thirds OADR-prediction interval of 18 per cent in 2030. This means that the bandwidth defined by possible policy options is smaller than the uncertainty implied by the demographics alone. In other words, the effects of labour market changes and disability changes will likely drown in inherent population uncertainty at the horizon 2030 (Note 10).

These two examples show how important it is to take population forecast uncertainty seriously. However, an often-heard objection is that projections of the kind mentioned here are purely conditional “what-if” calculations, and that it is unnecessary to consider expected forecast errors. “If the population would show this or that trend, *what* would the consequences be for the health care system?” Any deviations from the assumed demographic path are not of primary importance. We do not agree with this

objection. First, since demographic variables are hard to predict, the policy relevance of many long-term deterministic planning studies is limited, unless one can demonstrate that demographic uncertainty is negligible. In all other cases one is forced to think in terms not of point forecasts, but of intervals - and intervals imply expected accuracy and statistical analyses. Second, also conditional “what-if” projections may usefully be couched in probabilistic terms (Alho 1997). This way one can take interventions into account in the assessment of uncertainty of forecasts that are conditional on some planned future policy being adopted. Technically speaking, there is no difference between our approach in which prediction intervals were calibrated such that expected values agreed with the medium variant of Statistics Norway’s population forecast, and an approach in which other expected levels are selected as targets.

Do the large prediction intervals for some long-term forecasts imply that we should give up trying to forecast 50 years from now? We do not think so. In certain cases (e.g. pension systems), it is necessary to plan that far ahead. Clearly, policy inputs and other changes in underlying factors may affect the forecast results, but this can be handled by flexible planning and frequent updates. The point is that the user should know which trends are more certain than others, that is, which plans should be flexible, and which can be rather more fixed.

The format in which stochastic population forecasts are made available to the users is very different from that employed for traditional deterministic forecasts. The latter type of results can be included in tables in printed reports. If results for one-year or five-year age groups are published, the user who is interested in a specific larger age group can find the corresponding number by simply adding the age-specific results. A stochastic forecast, however, is presented in the form of predictive distributions. *Each forecast result has its own distribution.* Thus, the boundaries of the prediction interval of a larger age group are generally *not* equal to the sum of the boundaries of the constituent ages. For instance, the lower and upper bounds of the 95 per cent interval of total population in 2050 are 3.20 and 7.29 million, respectively, see Figure 7. On the other hand, the sum of the corresponding bounds for men and women broken down in five-year age groups in Figure 11 are 2.77 and 8.00 million. The prediction interval is smaller than the results obtained by simple addition, because the results for the various age groups are not perfectly correlated. The consequence is that stochastic forecasts should be made available in the form of a database (Alho 1990), from which the user can construct the prediction interval of any age group in which he is interested.

5. Conclusions

In this paper we have argued that traditional deterministic population forecasts do not reflect forecast uncertainty appropriately. High-low bands around the Medium variant results of a traditional forecast do not quantify expected errors, and implicitly they assume perfect correlation between components and over time.

We showed how time series methods can be used to compute a probabilistic population forecast. In our application for Norway until 2050, we checked the short-term prediction intervals for the TFR, the life expectancy at birth, and the age structure against observed errors in old forecasts. This assessment was quite informal; yet we think it was useful, as it captured the essential patterns in the historical errors. The long-term prediction intervals of the TFR and the levels of immigration and emigration were excessively wide, and were therefore adjusted in a subjective manner. The time series models were calibrated such that they predicted expected values for fertility, mortality, and migration that corresponded with Medium variant assumptions in Statistics Norway's population forecast. We investigated the relative importance for prediction intervals around a number of forecast outcomes when various types of variance are included in the computations. Ignoring the estimation variance in time series coefficients can have a strong impact on the prediction intervals, depending on the fit of the time series models, and the type of forecast result in which one is interested.

We estimated that the odds are four against one (80 per cent chance) that Norway's population, now 4.5 million, will number between 4.3 and 5.4 million in the year 2025, and 3.7-6.4 million in 2050. The probabilistic population forecasts of the youngest and the oldest age groups show largest uncertainty, because fertility and mortality are hard to predict. As a result, prediction intervals in 2030 for the population younger than 20 years of age are so wide, that the forecast is not very informative. International migration shows large prediction intervals around expected levels, but its impact on the age structure is modest. In 2050, uncertainty has cumulated so strongly, that intervals are very large for virtually all age groups, in particular when the intervals are judged in a relative sense (compared to the median forecast).

According to our statistical model, the expected accuracy of the total population size forecast published by Statistics Norway is somewhat below two-thirds on the long run, and a little above that level on the short run. The official TFR assumptions have estimated coverage probabilities of only 46, 31, and 24 per cent in the years 2010, 2030, and 2050. The official mortality (i.e. life expectancy) assumptions have higher expected accuracy in 2050 (just over 60 per cent), but lower accuracy in the beginning of this century (just over a third in the period 2000-2010).

There is no guarantee that the statistical model that we used to compute prediction intervals is the “real” or “correct” one. Unfortunately, different statistical models may produce widely differing prediction intervals (e.g. Lee 1974, Cohen 1986, Sanderson 1995). This is also confirmed by our sensitivity tests. Therefore, we have most confidence in our short-term intervals (up to roughly 15-20 years ahead; at least for the TFR and the age structure), because these agree rather well with independently observed errors in historical forecasts. But our long run results must be interpreted with caution. We remind the reader of Joel Cohen’s warning: “Uncertainty attaches not only to the point forecasts of future population, but also to the estimates of those forecasts’ uncertainty.” (Cohen 1986).

Finally, all our prediction intervals and assessments of future errors are mere extrapolations of the variability in past data. Thus we assume that the future structure underlying the demographic dynamics will be similar to that of the past - no structural shifts are expected. Of all past futures, so many were a smooth continuation of their immediate past, that we expect this is also likely in the years to come. However, we cannot exclude surprises, and some of our sample paths do indeed contain unexpected developments. For instance, a baby boom with a TFR over 3 children per woman in any year before 2050 has a likelihood of 19 per cent according to our model, and one with a TFR higher than 3.5 children still one per cent. In contrast, the expected value of the TFR in 2050 is 1.8, virtually the same as the current value. Similarly, the life expectancy of women is expected to increase to 84.5 years in 2050, more than three years higher than its current value of 81.3. This is in line with the continuous improvement of the survival chances of women in the past century. Nevertheless, the simulations show a chance of 12 per cent that women will have a life expectancy below 81 years in 2050, i.e. *higher* mortality in 2050 than nowadays.

6. Acknowledgements

Comments on an earlier version of this paper by Øystein Kravdal, Helge Brunborg, and six anonymous reviewers are gratefully acknowledged. Kluwer Academic Publishers kindly granted us the right to include in the current text parts of our paper “Predictive intervals for age-specific fertility”, which appeared in the *European Journal of Population*, Vol. 16, no. 1 of 2000. The research was supported by grant nr. 114055/730 from the Norwegian Research Council.

Notes

1. The assumption of a normal distribution can be debated. Although it is quite usual in empirical applications so far, other options have been proposed, for instance the beta distribution, the uniform distribution (Alders 1997), and the censored normal distribution (Keilman 2001). These options may be useful when one wants to obtain bounded prediction intervals. For fertility, still another option is to model the TFR as a specific stochastic process, for instance one in which the upward *slope* in the TFR is correlated negatively with the present *level*, or alternatively one in which there are two equilibrium levels, with a very low probability for levels in-between (Bonneuil 1989). For reasons of simplicity, we did not analyse such more complicated models. At the same time we noticed, that at least for fertility, the short-term errors predicted by our model agree closely to the observed errors in historical forecasts (see Section 3.1.1). At the same time, some sample paths of future TFR implied a baby boom, but our model predicts that such a boom is not very probable. For example, 412 of our 5000 sample paths, or less than one per cent, contained a TFR value larger than 3.5 in any year; in 19 per cent of the paths the TFR exceeded 3 children per woman in any year.
2. We did not consider other processes than the simple Poisson process in which the rate only depends on age. See Note 2 in Keilman and Pham (2000) for a brief discussion.
3. The impact of the upper limit on the shape of the TFR-distribution in any future year can be investigated *analytically* when the TFR follows a random walk (Keilman 2001). In that case, a maximum TFR of 3.4 children per woman has only negligible impact on the distribution and the prediction interval for the Norwegian TFR. Although a random walk clearly is not a realistic model for the Norwegian TFR after World War II, the critical value of 3.4 gives qualitative support to the conclusion reached by simulation of the more realistic ARIMA (1,1,0) model, i.e. a critical value of 4 children per woman.
4. We ignored the sample variance of the empirical life expectancy (Chiang 1968), because estimated standard errors for life expectancies at birth were as low as 0.01-0.03.
5. Alho reports 80 per cent prediction intervals. In order to obtain 95 per cent intervals, we multiplied his 80 per cent intervals by $1.96/1.282=1.53$, assuming a normal distribution.

6. We multiplied the 90 per cent prediction intervals reported by Tuljapurkar et al (2000) by $1.96/1.645=1.19$.
7. The impact of age grouping can be illustrated as follows. For the US, the 95 per cent interval is 5.2 years wide in 2050. The same model (random walk with drift combined with Lee-Carter model) applied to *unabridged* combined-sex life tables gives a corresponding interval of 8.3 years wide (interpolated between 2035 and 2065 in Figure 2.5 of Lee and Tuljapurkar, 2001).
8. For the forecast of 1982, life expectancy was kept constant beginning in 1991.
9. For some very specific forecast results, such as the number of centenarians, uncertainty may be considerable also on the short run. The same holds true for stochastic forecast results at a low regional level, for instance small and medium-sized municipalities.
10. Fredriksen investigates also “packages” of alternative assumptions. For instance, in an “ageing alternative” he combines low population growth with high disability risks, and a “growth alternative” combines high population growth with low disability risks and high female labour force participation. The contribution rate varies between 16 and 30 per cent in 2030, when either of these packages is chosen. This implies a relative variation of 58 per cent, much larger than the prediction interval for the OADR. In this case, the sensitivity analysis has clear policy relevance. However, he implicitly assumes perfect correlation between the components: in the ageing alternative for example, *each* year when fertility is high, *also* disability risks are high, and both mortality *and* immigration are low.

References

- Alders, M. (1997). Constructing probability distributions of population forecasts. Unpublished note, Statistics Netherlands, Department of Population, 3 September 1977.
- Alders, M. and J. de Beer (1998). "Kansverdeling van de bevolkingsprognose" ("Probability distribution of population forecasts")., *Maandstatistiek van de Bevolking* **46**: 8-11.
- Alho, J. M. (1990). "Stochastic methods in population forecasting", *International Journal of Forecasting* **6**: 521-530.
- Alho, J.M. (1997). "Scenarios, uncertainty, and conditional forecasts of the world population". *Journal of the Royal Statistical Society A* **160**, Part 1, 71-85.
- Alho, J. M. (1998). A stochastic forecast of the population of Finland. Reviews 1998/4. Statistics Finland, Helsinki.
- Alho, J. M. (2001). Experiences from the forecasting of mortality in Finland. Paper National Social Insurance Board, Committee for Mortality Projections meeting, Stockholm, June 2001.
- Alho, J. and B. Spencer (1985). "Uncertain population forecasting" *Journal of the American Statistical Association* **80**: 306-314.
- Armstrong, J. (1985). *Long-range forecasting: From crystal ball to computer*. New York: Wiley (2nd ed.).
- Bonneuil, N. (1989). "Conjoncture et structure dans le comportement de fécondité". *Population* **44**(1): 135-157.
- Chiang, C.L. (1968). *Introduction to stochastic processes*. New York: Wiley.
- Cohen, J. (1986). "Population forecasts and confidence intervals for Sweden: A comparison of model-based and empirical approaches" *Demography* **23**: 105-126.
- De Beer, J. (1997). "The effect of uncertainty of migration on national population forecasts: The case of the Netherlands" *Journal of Official Statistics* **13**: 227-243.

- De Beer, J. and M. Alders (1999). "Probabilistic population and household forecasts for the Netherlands" Working Paper nr. 45, Joint ECE-Eurostat Work Session on Demographic Projections, Perugia, Italy, 3-7 May 1999.
- Fredriksen, D. (1998). *Projections of population, education, labour supply, and public pension benefits: Analyses with the dynamic microsimulation model MOSART*. Social and Economic Studies 101. Oslo: Statistics Norway. Internet www.ssb.no/emner/02/03/sos101/sos101.pdf (30 April 2002).
- Hanika, A., W. Lutz and S. Scherbov (1997). "Ein probabilistischer Ansatz zur Bevölkerungsvorausschätzung für Österreich" *Statistische Nachrichten* **12/1997**: 984-988.
- Heligman, L. and J. Pollard (1980). "The age pattern of mortality" *Journal of the Institute of Actuaries* **107**: 49-80.
- Keilman, N. (1997). "Ex-post errors in official population forecasts in industrialized countries" *Journal of Official Statistics* **13**: 245-277.
- Keilman, N. (2001). "TFR predictions and Brownian motion theory" *Yearbook of Population Research in Finland* **38**: 207-219.
- Keilman, N. and D.Q. Pham (2000). "Predictive intervals for age-specific fertility" *European Journal of Population*. **16**: 41-66.
- Keilman, N., D. Q. Pham, and A. Hetland (2001). *Norway's uncertain demographic future*. Social and Economic Studies 105. Oslo: Statistics Norway. Internet www.ssb.no/english/subjects/02/03/sos105_en (30 April 2002).
- Keyfitz, N. (1981). "The limits of population forecasting" *Population and Development Review* **7**: 579-593.
- Keyfitz, N. (1985). "A probability representation of future population" *Zeitschrift für Bevölkerungswissenschaft* **11**: 179-191.
- Kuijsten, A. (1988). "Demografische toekomstbeelden van Nederland" *Bevolking en Gezin* **1988/2**: 97-130.
- Lee, R. (1974). "Forecasting births in post-transition populations: Stochastic renewal with serially correlated fertility" *Journal of the American Statistical Association* **69**: 607-617.

- 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 S. Tuljapurkar (1994). "Stochastic population forecasts for the United States: Beyond High, Medium, and Low" *Journal of the American Statistical Association* **89**: 1175-1189.
- Lee, R. and S. Tuljapurkar (2001). Population forecasting for fiscal planning: Issues and Innovation. Pp. 7-57 in A. Auerbach and R. Lee (eds.). *Demographic change and fiscal policy*. Cambridge University Press.
- Lutz, W., W. Sanderson, and S. Scherbov (1996). "Probabilistic population projections based on expert opinion" pp. 397-428 in W. Lutz (ed.). *The future population of the world: What can we assume today?* London: Earthscan (rev. ed.).
- Lutz, W. and S. Scherbov (1998a). "An expert-based framework for probabilistic national population projections: The example of Austria" *European Journal of Population* **14**: 1-17.
- Lutz, W. and S. Scherbov (1998b). "Probabilistische Bevölkerungsprognosen für Deutschland" ("Probabilistic population projections for Germany") *Zeitschrift für Bevölkerungswissenschaft* **23**: 83-109.
- Lutz, W., W. Sanderson, and S. Scherbov (2001). "The end of world population growth", *Nature* **412**: 543-545.
- National Research Council - NRC (2000). *Beyond six billion: Forecasting the world's population*. Panel on Population Projections. John Bongaarts and Rodolfo Bulatao (eds.). Committee on Population, Commission on Behavioral and Social Sciences and Education. Washington DC: National Academy Press.
- Paulsen, B., B. Kalseth, and A. Karstensen (1999). "Eldres sykehusbruk på 90-tallet: 16 prosent av befolkningen – halvparten av sykehusforbruket". SINTEF rapport STF78 A99527. Trondheim: SINTEF Unimed.

- Pflaumer, P. (1986). "Stochastische Bevölkerungsmodelle zur Analyse der Auswirkungen demographischer Prozesse auf die Systeme der sozialen Sicherung" *Allg. Statist. Archiv* **70**: 52-74.
- Pflaumer, P. (1988). "Confidence intervals for population projections based on Monte Carlo methods" *International Journal of Forecasting* **4**: 135-142.
- Rogers, A. and L. Castro (1986). "Migration". Pp. 157-206 in A. Rogers and F. Willekens (eds.) *Migration and settlement: A multiregional comparative study*. Dordrecht: Reidel Publ. Co.
- Sanderson, W. (1995). "Predictability, complexity, and catastrophe in a collapsible model of population, development, and environmental interactions" *Mathematical Population Studies* **5**: 233-279.
- Statistics Norway (1997). *Framskrivning av folkemengden 1996-2050* ("Population projections 1996-2050"). Norges offisielle statistikk C414. Oslo: Statistisk sentralbyrå.
- Statistics Norway (1999). "Continued population growth. Population projections. National and regional figures, 1999-2050". Internet www.ssb.no/english/subjects/02/03/folkfram_en (30 April 2002).
- Statistics Norway (2001). *Framskrivning av folkemengden 1999-2050* ("Population projections 1999-2050"). Norges offisielle statistikk C693. Oslo: Statistisk sentralbyrå. Internet www.ssb.no/emner/02/03/nos_c693/nos_c693.pdf (30 April 2002).
- Stoto, M. (1983). "The accuracy of population projections" *Journal of the American Statistical Association* **78**: 13-20.
- Texmon, I. (1992). "Norske befolkningsframskrivninger 1969-1990" ("Norwegian population projections 1969-1990"). pp. 285-311 in O. Ljones, B. Moen, and L. Østby (eds.). *Mennesker og Modeller*. Oslo/Kongsvinger: Statistics Norway.
- Thompson, P., W. Bell, J. Long, and R. Miller (1989). "Multivariate time series projections of parameterised age-specific fertility rates" *Journal of the American Statistical Association* **84**: 689-699.
- Tuljapurkar, S. (1996). "Uncertainty in demographic projections: Methods and meanings" Paper Sixth Annual Conference on Applied and Business Demography. Bowling Green, 19-21 September 1996.

- Tuljapurkar, S., N. Li, and C. Boe (2000). "A universal pattern of mortality decline in the G7 countries" *Nature* **405**, 15 June 2000: 789-792.
- Törnqvist, L. (1949). "Om de synpunkter, som bestämt valet av de primäre prognos-antagandena" ("On the points of view that determined the choice of the main forecast assumptions") pp. 69-75 in J. Hyppölä, A. Tunkelo, and L. Törnqvist *Beräkningar rörande Finlands befolkning, dess reproduktion och framtida utveckling* Statistiska Meddelanden nr. 38. Helsinki: Statistiska Centralbyrån (in Swedish and Finnish).

