3 Empirically based specification of forecast uncertainty

Juha M. Alho, Harri Cruijsen and Nico Keilman

Introduction

To make a conventional population forecast one needs to specify age-specific fertility rates for women, and mortality rates for women and men, for all future years of interest. These are used to generate births and deaths. The simplest way to handle migration is to specify net migration in absolute numbers that are added to population each year. Starting from a jump-off population, the so-called cohort-component bookkeeping (e.g. Shryock, Siegel and associates, 1976) is used recursively to keep track of the resulting changes in population, by age and sex. These methods were first used by Cannan (1895) for England and Wales, and since the 1920s and 1930s they have been widely used in Europe (DeGans, 1999). The early forecasters were aware that calculations based on the cohort-component method are only as reliable as the assumptions that go into making them. Alternative variants were offered from early on, but even the forecast producers themselves were uneasy about the methods that were used to prepare them (e.g. Modeen, 1934).

Stochastic (or probabilistic) cohort-component forecasts are similar, but in this case future fertility and mortality rates and net migration are considered as random variables (e.g. Alho and Spencer, 2005). Their distributions can be specified in various ways. Perhaps the simplest is to give first the location of the distribution, and then to specify the spread (or scale) around it to reflect forecast uncertainty. Under a normal (Gaussian) assumption for the rate (or, for example, its log transform), the measure of location is the mean (or median) and the measure of spread is the standard deviation, for example. An advantage of the normal model is that the dependency structure of the various random variables can be given in terms of correlations in an interpretable way.

From the perspective of random vital rates, cohort-component bookkeeping is a non-linear operation. Simulation is frequently used to carry out the propagation of uncertainty, from the rates to future population numbers. A joint distribution derived in this manner for the future demographic variables can be called a *predictive distribution*. In other words, it is the probability distribution of the future demographic variables, conditional on what is known as the jump-off time, or the time when a forecast is made.

Complementing Keilman, Cruijsen and Alho (this volume, chap. 2), who discuss the specification of the most likely future values of the vital rates, or the location of the predictive distributions, we illustrate in this chapter how the second-order characteristics (scales and correlations) of the predictive distributions can be specified. We carry out our analyses in a multi-country setting, i.e. we use information from other countries to stabilize error estimates that would otherwise be expected to be highly autocorrelated, and, as such, hard to estimate (Alho and Spencer, 2005).

The specification of locations is essentially equivalent to the preparation of the 'medium' variant of a conventional population forecast. The novel task of specifying the scales and correlations is important, however. It has been demonstrated that, in the past, official forecasts have had much larger errors than one would have anticipated by looking at the alternatives ('high' and 'low') that are often provided (e.g. Keilman, 1990; National Research Council, 2000). In applications involving the sustainability of public finances (e.g. pensions), too narrow a view of future contingencies can lead to erroneous policy decisions (e.g. Auerbach and Lee, 2001).

The interpretation of a predictive distribution depends on the way the location and scale are specified. The statistical outlook allows for a combination of modelling of vital processes, analysis of past forecast errors and judgement concerning factors that may not have fully manifested themselves yet (Alho and Spencer, 2005). We know from elementary statistics that probability statements always depend on a model (e.g. Freedman, Pisani and Purves, 1978), so we need to be able to construct a model that is capable of incorporating the various sources of information available. As a starting point we will *specify scales and correlations in such a way that, had they been used in the past, the prediction intervals of demographic variables would have had the specified level of coverage*. In this chapter we try to give enough detail so the reader can critically examine our assessment of uncertainty.

We begin by evaluating the spread provided by Eurostat's and the UN's high and low forecasts for eighteen European countries against probabilistic estimates produced by the project UPE, and others. This shows to what extent the past underestimation of uncertainty persists. We then review approaches to specifying past uncertainty and present details of the scaled model for error that was applied in the economic work considered in this volume. We consider prediction intervals for the basic demographic variables and compare them to the official high–low intervals, in order to see what might explain the differences. We conclude by discussing the implications of the findings for ageing research.

For readers who are not familiar with the stochastic approach to population forecasting we include, in an appendix, a schematic comparison of conventional and stochastic forecasts.

Uncertainty in official forecasts of total population

A practical finding from experiments with stochastic population forecasting in the 1980s was that the 'plausible ranges' formed by official high and low projections, or scenarios, for the total population were usually too narrow to give a realistic indication of the uncertainty to be expected. Stoto (1983) found, for example, that in the United States the high–low intervals were somewhat narrower than the 'one-sigma intervals' that one would expect, based on the normal distribution, to capture the future population about two-thirds of the time. This came as a surprise, since the primary motivation for the development of stochastic approaches was to enhance the logical coherence of alternative forecasts rather than the level of uncertainty *per se*. Yet, since we now know that the high–low intervals were too narrow, we also know that their use has had the potential to mislead decision-makers by encouraging a concentration on too narrow a range of alternative future paths.

Using Table 3.1 we can investigate to what extent the most recent official forecasts of the population of eighteen European countries (consisting of 'EU-15' plus Iceland, Norway and Switzerland) continue to provide narrow ranges. We consider year 2050.

In order to produce comparable measures we calculate the following measure from the UN and Eurostat forecasts. Define M = middle forecast, H = high forecast and L = low forecast. The measures under columns 'UN' and 'ES', are of the form $100 \times (H - L)/2M$, i.e. it is the half-width of the relative difference between the high and the low, expressed in percentage terms. The column 'UPE' has been derived from summary data for a predictive distribution for each of the countries. The values used are available at www.stat.fi/tup/euupe/de112.pdf, where for each country there is a table comparable to our Table 3.1 giving the median of the predictive distribution, M', together with the upper end-point of the 80 per cent prediction interval, H', and the lower end-point of the 80 per cent prediction interval, L', up to year 2050. The entry given in Table 3.1 is calculated as $100 \times (H' - L')/2 \times 1.2816 \times M'$, where 1.2816 is the 0.9 fractile of the standard normal distribution. This approximates the coefficient of variation of the predictive distribution of the total

Table 3.1. *Relative uncertainty as expressed by high and low forecasts of the United Nations (2004) and Eurostat (2005), the relative standard deviation derived by the UPE project (UPE) and as estimated from the empirical errors of the UN forecasts in the period 1970–1990 (EMP), for the total population in 2050.*

	UN	ES	UPE	EMP
Austria	14.0	16.4	14.2	19.3
Belgium	9.8	13.1	13.2	8.8
Denmark	14.7	12.7	12.4	9.6
Finland	14.5	11.2	13.4	12.4
France	14.6	11.4	14.9	9.3
Germany	14.5	17.8	15.0	17.0
Greece	14.4	15.0	16.0	16.2
Iceland	15.1	–	15.1	9.9
Ireland	14.0	13.7	16.5	34.6
Italy	13.8	12.3	17.6	9.0
Luxembourg	9.0	16.4	16.8	45.4
Netherlands	12.0	15.2	12.8	9.3
Norway	14.3	–	13.2	10.3
Portugal	14.2	16.7	16.5	27.8
Spain	14.4	13.9	17.0	23.2
Sweden	14.3	13.2	15.3	20.0
Switzerland	15.0	–	11.8	30.8
United Kingdom	15.0	15.6	13.5	11.6
Median	14.6	13.9	15.0	14.3
Standard deviation	1.7	2.0	1.8	10.6

population in 2050.[1] Finally, the column 'EMP' is the estimated coefficient of variation (of the true population about the forecast) obtained from an empirical analysis of errors in the UN forecasts made in the period 1970–1990, as given in Table 2 of Chapter 8 in Alho and Spencer (2005).

Two findings are important. First, by using the median values across countries as a measure, we find that *all* four approaches lead to almost the same value of slightly under 15 per cent. This indicates that the high–low intervals of official forecasts for these countries are comparable to 'one-sigma intervals'; or, the probability is about two-thirds that the high–low intervals capture the total population in 2050. This agrees with Stoto's finding from the United States that there is a large probability that total population will be outside the range deemed 'plausible'.[2]

Second, the standard deviations across countries are almost the same, at about 2 percentage points, for the two official forecasts and the UPE values. As we will see below, it is likely that this is by deliberate design: forecasters, both official and those of the UPE project, have found it difficult to defend values that differ considerably across countries. The contrast with column 'EMP' is striking. The standard deviation is quite large compared to the median. This is caused by a number of countries with very large values.

We note also that although the EMP values are based on an analysis of past errors of the UN forecasts, their correlation with the UN values is negative (with P-value = 0.2 for the hypothesis of zero correlation). Their correlation with both the ES and UPE values is positive (P-values = 0.1). Although the observation that future developments have turned out to be different from what was anticipated has been a frequent motive for updating forecasts, the lack of correlation is consonant with the fact that official forecast assumptions have rarely been based on a systematic evaluation of past forecast errors.

Prudence would seem to dictate that a *higher* level of uncertainty should be specified than that given in the column UPE. As discussed by Keilman, Cruijsen and Alho (this volume, chap. 2), there have been systematic problems in the past forecasts. Eliminating those is one reason for preferring somewhat lower estimates of uncertainty. We note also that there is evidence that small countries typically have somewhat larger errors than larger countries (e.g. Alho and Spencer, 2005). No attempt was made to induce such regularity into the UPE specifications, however.

Modelling and estimating uncertainty

Problems of coherence in conventional forecasts

In the preceding section, we limited ourselves to total population. However, in many applications one is interested in sub-populations. For example, in ageing research one studies the old-age dependency ratio D. Defining W as the size of the population of working age (say, 20–64), and defining E as the size of the elderly population (65+), then $D = E/W$. Define, further, Y as the number of young (0–19) and the total population as $P = Y + W + E$.

The nature of uncertainty depends on lead time, so for the following discussion let us assume that we are looking about twenty years into the future. Then, the uncertainty in Y depends primarily on fertility, the uncertainty in W on migration and the uncertainty in E on mortality. The reasons why we make errors in forecasting different components are typically unrelated. (To be sure, higher-than-expected immigration may cause an underestimation of fertility and mortality, but the level of migration is rarely high enough, relative to variations of fertility and mortality caused by other factors, for this to matter.) Thus, it is a realistic approximation to assume the three error sources are statistically independent. It follows that errors in forecasts of P from Y, W and E *partly cancel*.

In conventional forecasts this is not taken into account. Instead, a high forecast for P, say $H(P)$, is obtained from the high forecasts of the components, $H(P) = H(Y) + H(W) + H(E)$, and similarly for the low forecasts, $L(P) = L(Y) + L(W) + L(E)$. It follows that if we view, in accordance with the medians of Table 3.1, $[L(P), H(P)]$ roughly as a two-thirds level prediction interval for P, then $[L(Y), H(Y)]$, $[L(W), H(W)]$ and $[L(E), H(E)]$ must have *lower* levels of coverage. How much lower? The precise answer depends on many details but note that if Y, W and E have the same expectation and the same coefficient of variation, say C, then the coefficient of variation of P is $C \times \sqrt{3}/3$, or 58 per cent of those of its components. Or, under a normal approximation, the intervals for the three components would each have only a 44 per cent probability of covering the true population, instead of 67 per cent.

Things get worse with the old-age dependency ratio D. How could one get an upper and lower interval for it at all? There is no guarantee that $H(E)/H(W)$ is bigger than $L(E)/L(W)$. It could be smaller! It is certainly true that $H(E)/L(W) > L(E)/H(W)$, but forming an interval from these two is not 'plausible', since it only makes sense if there were a perfect negative association between W and E. Of course, one could separately produce upper and lower values for E and W such that the resulting interval for D would be 'reasonable'. Unfortunately, such values would typically be unusable for P.

Compounding the logical difficulties, there are even more basic problems. For example, if $[L(E), H(E)]$ is specified, based on mortality, to be of a 'plausible' width, how can one guarantee that the width of $[L(W), H(W)]$ is of a similar 'plausibility' when it is based on migration? If one of these has a coverage probability of 30 per cent and the other 70 per cent, both are 'plausible' values, but combining them does not seem sensible. The same problem for determining comparable scales arises even for a single vital process, when different lead times, different ages or the two sexes are considered jointly.

Scaled model for uncertainty

In the program PEP[3] that was used to carry out the demographic analyses in this volume, the variances of the logarithms of the future age-specific fertility and mortality rates are represented in terms of the following model (Alho and Spencer, 1997, 2005; Alho, 1998).

Suppose the true age-specific rate for age j during forecast year ($t > 0$) is of the form $R(j, t) = F(j, t)\exp(X(j, t))$, where $F(j, t)$ is the point forecast and $X(j, t)$ is the relative error. Suppose that the error processes are of the form $X(j, t) = \varepsilon(j, 1) + \ldots + \varepsilon(j, t)$, where the error increments are of the form

$$\varepsilon(j, t) = S(j, t)(\eta_j + \delta(j, t)). \tag{1}$$

Here, the $S(j, t)(> 0)$ are (non-random) scales to be specified, whence the name *scaled model*. The model assumes that for each j, the random variables $\delta(j, t)$ are independent over time $t = 1, 2, \ldots$ In addition, the variables $\{\delta(j, t) | j = 1, \ldots, \mathcal{J}; t = 1, 2, \ldots\}$ are independent of the random variables $\{\eta_j | j = 1, \ldots, \mathcal{J}\}$, and

$$\eta_j \sim N(0, \kappa), \delta(i, t) \sim N(0, 1 - \kappa), \tag{2}$$

where $0 < \kappa < 1$ is a parameter to be specified. We see that $\mathrm{Var}(\varepsilon(j, t)) = S(j, t)^2$.

In the UPE applications it was assumed that $\mathrm{Corr}(\eta_i, \eta_j) = \rho^{|i-j|}$, or $\mathrm{Corr}(\eta_i, \eta_j) = \mathrm{Corr}(\delta(i, t), \delta(j, t)) = \rho^{|i-j|}$, for some $0 < \rho < 1$. This allows for less than perfect correlation in age-specific mortality and fertility, across age.

Since $\kappa = \mathrm{Corr}(\varepsilon(j, t), \varepsilon(j, t + h))$ for all $h \neq 0$, we can interpret κ as a constant autocorrelation between the error increments. Under a random walk model the error increments would be uncorrelated with $\kappa = 0$, for example. Together, the autocorrelation κ and the scales determine the variance of the relative error $X(j, t)$.

Example 3.1. Consider the special case of constant scales, or $S(j, t) = S(j)$ for all t. It follows that

$$\mathrm{Var}(X(j, t)) = a(j)t + b(j)t^2, \tag{3}$$

where $a(j) = S(j)^2(1 - \kappa)$ and $b(j) = S(j)^2\kappa$. If $a(j)(> 0)$ and $b(j)(\geq 0)$ can be estimated from the data, then the corresponding values of $S(j)^2$ and κ can be deduced as $S(j)^2 = a(j) + b(j)$ and $\kappa = b(j)/(a(j) + b(j))$. For intuition, note that the model with constant scales can be interpreted as a random walk with a random drift. The relative importance of the two components is determined by κ. □

The key properties of the scaled model are:

- Since the choice of the scales $S(j, t)$ is unrestricted, any sequence of non-decreasing error variances can be matched. In particular, heteroscedasticity can be allowed.

- Any sequence of cross-correlations over ages can be majorized using the AR(1) models of correlation.
- Any sequence of autocorrelations for the error increments can be majorized. This means that we can always find a conservative approximation to any covariance structure using the model.

As discussed by Alho and Spencer (2005) the scaled model can be used to replicate prediction intervals of the mortality forecasts with the method of Lee and Carter (1992), for example.

The representation of error in net migration in PEP is done in absolute terms, using variables of the same type as $\varepsilon(j,t)$, above, but now j takes only two values and refers to *sex*. Dependence on age is not stochastic, and assumed to be deterministic, and it is given by a fixed distribution $g(j,x)$ over age x for each $j(=1,2)$, i.e. the error of net migration in age x, for sex j, during year $t(>0)$, is additive and of the form

$$Y(j,x,t) = S(j,t)g(j,x)(\eta_j + \delta(j,t)). \quad (4)$$

The assumption of perfect dependence is not motivated by a belief that there would not be any cancellation of error across age in migration. Surely there is. Instead, the quality of migration data is too poor in most countries to merit a more refined approach.

An advantage we had over earlier national analyses in this field was that we had eighteen countries under scrutiny simultaneously. This allowed us to discount idiosyncrasies (evident, for example, in column 'EMP' of Table 3.1) that could dominate the forecast results of an individual country. This is one of the traditional methods of 'borrowing strength' from similar units of observation that is widely used in small-domain estimation (e.g. Rao, 2003).

Approaches to assessing forecast uncertainty

The presence of autocorrelation distinguishes time-series analysis from the more standard statistical theory that assumes statistical independence among observations. For example, even though one can estimate the mean of a stationary sequence simply by averaging the observations, the calculation of the corresponding standard error presumes that the autocorrelation structure has been identified and estimated. In general, positive autocorrelation leads to *higher* standard errors than one would expect under independent sampling of otherwise similar data (e.g. Alho and Spencer, 2005). Intuitively, autocorrelation reduces the amount of information which a sample of given size contains.

Even if one abstracts from technicalities, it is clear that in a time-series context one is typically interested in forecasting a future value (i.e. a random variable) rather than a parameter of the process, so additional complexities arise that are related to the handling of possible nonstationarity, and to other aspects of the identification of model type.

Nevertheless, for a wide range of models (such as linear stochastic processes (Box and Jenkins, 1976; Alho and Spencer, 2005)) the results are well known. Thus, if demographic forecasts are carried out using a formal statistical model, both optimal predictions and prediction intervals can typically be produced once the model has been identified and its parameters have been estimated. Accounting for uncertainty related to the standard errors of the parameters has long been neglected in ARIMA modelling for example, but Markov chain Monte Carlo techniques have overcome many of the technical obstacles (Chib and Greenberg, 1994; Alho and Spencer 2005). These can be used 'off the shelf', but the question remains of how the choice of models and their validity over time can be included as part of the assessment of uncertainty.

Errors in past forecasts provide an obvious source of information on the level of error one might expect in the future (Keilman, 1990, 1997). It is not necessary to believe that future errors will be similar to those in the past, but if one does not believe that they will be, it is necessary to provide argumentation as to why the future is expected to be different from the past.

A difficulty in the use of past errors as a guide is that of obtaining estimates that are statistically stable. We have seen in Table 3.1 that empirical forecast errors can vary considerably from one country to another. Other analyses indicate that there can also be large variations over time (Alho and Spencer, 2005). One would expect this to be the case given the high autocorrelation of forecast errors.

Another problem in the use of past errors as a guide is that the number of observations (i.e. the number of available past forecasts) diminishes rapidly when lead time is increased. There is no country in the world for which there would be more than a handful of forecasts with lead time exceeding fifty years, whose error can be assessed. Yet, for pension problems, even longer lead times must be considered (Lee and Tuljapurkar, 1998).

As a way out, it was proposed by Alho (1990) that one should resort to so-called naïve, or baseline, forecasts. It had been noted that official forecasts of fertility, in industrialized countries, typically assume little change from the current level (Lee, 1974). Thus, one can approximate past forecasts by assuming no change. This forecast is easily computed for any point in the past, even if no forecast was actually carried out at that time. Its empirical error can similarly be computed with ease, as long into the future as there are data points. In the case of mortality an assumption of a constant rate of change has been shown to be competitive with official forecasts (Alho, 1990), so it can serve as a baseline forecast.

Example 3.2. Suppose we have data for years $t = 1, 2, \ldots, T$. Suppose a baseline forecast, made at t with lead time $k = 1, 2, \ldots$ for $R(j, t + k)$ is denoted by $F(j, t, k)$. The absolute value of its relative error is then $e(j, t, k) = |\log(F(j, t, k)) - \log(R(j, t + k))|$. It follows that we have a collection of values $Z(j, k) = \{e(j, t, k) \mid t = 1, 2, \ldots, t - k\}$ available. These can be used to estimate the parameters $a(j)$ and $b(j)$ of (3) in various ways. In order to discount the values of outliers caused, for example, by wars, a robust procedure is to determine first the medians $M^*(j, k) =$ median over t of the set of values $Z(j, k)$, for every $k = 1, 2, \ldots$ of interest. Then, since the 0.75 fractile of a standard normal distribution is 0.6745, we can find a standard deviation for a normal distribution whose absolute value has the same median, as $M(j, k) = M^*(j, k)/0.6745$. These $M(j, k)$ values can serve as our basic data. Returning to the case of Example 3.1, we note that by minimizing the sum of squares,

$$\Sigma_t (M(j, t)^2 - a(j)t - b(j)t^2)^2, \qquad (5)$$

we can find values $a(j)$ and $b(j)$ that fit as closely as possible. If the values satisfy the necessary positiveness conditions, we can deduce the parameters of the scaled model. □

The example given above is by no means the only approach available. First, in the case of fertility and mortality we consider *relative* error $X(j, t)$. Absolute error could also be used, but given the large variations in the level of the processes, relative error is a more comparable measure. Second, in assessing the magnitude of error we do not subtract the mean. This means that *we are including the possible forecast bias* in the error estimate. This is motivated by the fact that we do not believe that biases can be avoided in the future either. Third, the added twist of using the medians rather than *averages* typically reduces the estimated uncertainty. This is relevant if the intended use of the predictive distribution is to give an indication of how much variability one should expect under normal, peacetime conditions. For analyses with other background assumptions, one might resort directly to root-mean-squared error or other measures of spread.

In general, error estimates based on naïve forecasts should be conservative in the sense of providing intervals that are potentially too wide. Indeed, it is possible to find countries and periods during which naïve forecasts have been very bad. However, this is also true of official forecasts, and as discussed in Alho (1990), official US forecasts of mortality have not been better than naïve forecasts in the post-World War II period. Note also that like error estimates based on past forecasts, error estimates based on naïve forecasts are 'self-validating' in that they are correct if volatility does not change. In this respect they are superior to purely model-based estimates.

Practical specification of uncertainty

Variance of total fertility and age-specific fertility

Based on empirical estimates from six countries with long data series (Denmark, Finland, Iceland, the Netherlands, Norway and Sweden) one could determine that the (logarithm of the) total fertility rate behaves essentially like a process of independent increments, so we took $\kappa = 0$ for all countries. The same analysis led to an approximate formulation with a scale appropriate for the total fertility 0.06 (Alho and Spencer, 2005). However, during the most recent decade, fertility has been less volatile than in the past, on average. To account for the recent decreases in volatility we estimated a starting level of scale for each country based on data from 1990–2000. These were assumed to increase linearly to 0.06 in twenty years time.

The initial scales determined in this manner were scaled up by a factor of 1.25 (or, we took $S(j, t) = 1.25 \times 0.06$ for $t > 20$) to account for the fact that the correlation of the relative error over age was not perfect, but 0.95. This cross-correlation has an empirical basis in the work of Alho (1998) in Finland. The same scale was used in all ages.

The relative standard deviations of error for the total fertility rate are given in Table 3.2.[4]

For comparison, by defining M = middle forecast, H = high forecast, and L = low forecast, the measures under columns 'UN' and 'ES', are of the form $100 \times (H - L)/2M$. For the United Nations forecasts, it is assumed that all countries except Greece have $M = 1.85$, $H = 2.35$, and $L = 1.35$, but that Greece has $M = 1.78$. In contrast, Eurostat has both different point forecasts for different countries and, in some cases, assymmetric high–low intervals. The difference between the UN and Eurostat assumptions is striking. The Eurostat values are almost exactly one-half of the UN ones. Presumably both attempt to cover a 'reasonable range', but the message given to the forecast users is quite different.

The UPE values are based on an empirical analysis, and a UPE–UN comparison shows that UN values are approximately three-quarters of the empirically estimated UPE values that yield the 'one-sigma intervals'; or, the probability content of the UN high–low intervals is about 55 per cent. The implication is that the high–low intervals for the young (Y) have a

Table 3.2. *The standard deviation of the relative error of total fertility during first and last forecast years as specified by the UPE project, and relative uncertainty as expressed by high and low forecasts of the United Nations (2004) and Eurostat (2005).*

Source		UPE	UN	ES
Year	2003	2049	2050	2050
Austria	2.2	36.6	27.0	17.2
Belgium	2.6	36.9	27.0	13.2
Denmark	1.9	36.3	27.0	13.9
Finland	2.0	36.3	27.0	13.9
France	2.3	36.7	27.0	13.5
Germany	2.9	37.2	27.0	17.2
Greece	1.6	36.0	28.1	18.3
Iceland	2.8	37.2	27.0	–
Ireland	2.2	36.6	27.0	11.1
Italy	2.5	36.8	27.0	17.9
Luxembourg	2.6	37.0	27.0	13.9
Netherlands	1.9	36.3	27.0	14.3
Norway	2.2	35.8	27.0	–
Portugal	2.2	36.6	27.0	15.6
Spain	2.6	36.9	27.0	17.9
Sweden	3.9	38.4	27.0	13.5
Switzerland	1.5	36.0	27.0	–
United Kingdom	1.5	35.9	27.0	14.3
Median	2.2	36.6	27.0	14.3
Standard deviation	0.6	0.6	0.3	2.2

clearly lower level of coverage than the resulting intervals for the total population (P).

Variance and autocorrelation of age-specific mortality

In the case of mortality the UPE project had available direct estimates of relative uncertainty in naïve forecasts of age-specific mortality. However, we were also urged by the expert we interviewed[5] not to underestimate the potential effect of unforeseen developments in medical technology on mortality. This is especially relevant in the oldest ages for which past improvements had the least effect. Also, given that conditions influencing future mortality may be quite different for young ages than for older ones it was decided to make estimates of uncertainty by age. Thus, for mortality, the scales we used depended on age. On the other hand, we

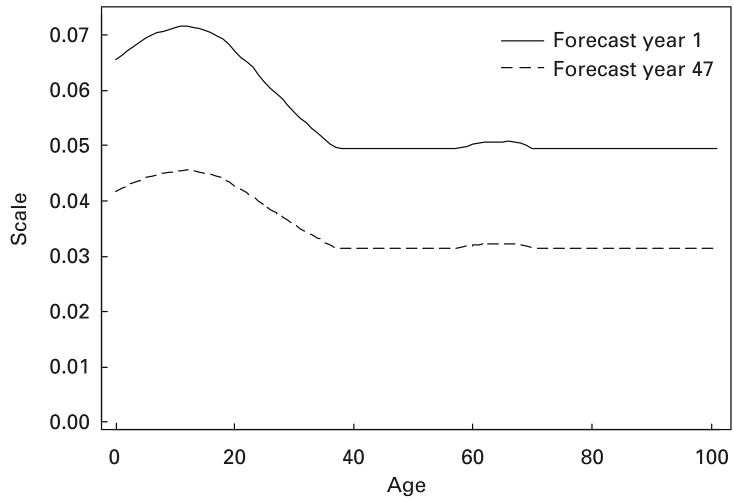

Figure 3.1. Scales for age-specific mortality in forecast year 1 and forecast year 47.

could not see a good reason to make separate estimates for males and females, or for the different countries. This enhances statistical stability. Finally, as we were also urged to consider the possibility of changing trends in the future, we used a value $\kappa = 0.05$ for the autocorrelation of the error increments. Purely empirical estimates from the eighteen countries would have led to an average value close to zero, whereas an empirical analysis of Finland alone (Alho, 1998) gave as high a value as 0.15. Judgement was brought to bear.

Contemporaneous correlation of (the logarithm of) age-specific mortality, in a given age, was taken to be 0.85 between males and females. In Alho (1998) an average correlation of 0.80 was obtained. This was increased, on judgemental grounds, to maintain a smaller male–female gap.

Figure 3.1 gives the scales in forecast years 1 and 47. We see that in the past, volatility has been highest in the youngest ages. For those ages where the curves are flat, empirical estimates were adjusted upwards to the level of the overall average on judgemental grounds.

Figure 3.2 summarizes the effect of scales and autocorrelation into a standard deviation of the relative forecast error. The curves are all parallel, and the one shown in Figure 3.2 represents the lowest level of uncertainty that obtains in the middle and oldest ages (the flat regions in Figure 3.1). The shape of the curve is based on an analysis of errors of naïve forecasts

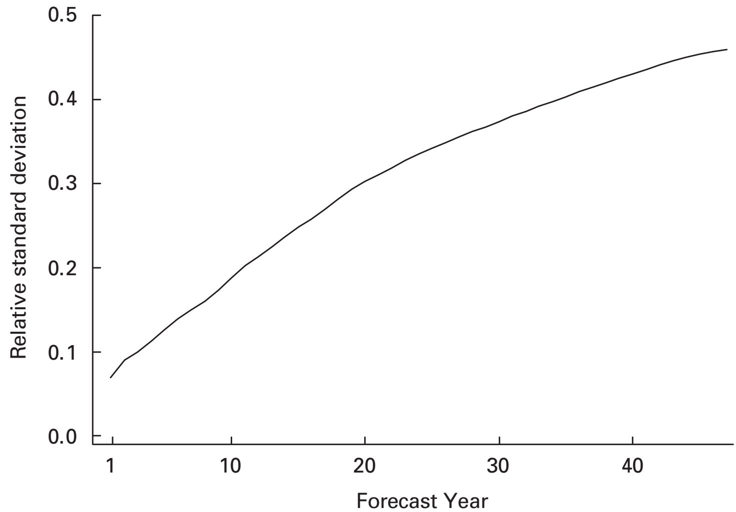

Figure 3.2. Relative standard deviation of age-specific mortality for ages 40+.

from Austria, Denmark, France, Italy, the Netherlands, Norway, Sweden, Switzerland and the UK. The scales were deduced from these estimates. The concavity explains why the later curve is lower than the earlier one in Figure 3.1.

Table 3.3 presents a comparison between the one-sigma intervals of the UPE project and the half-width of Eurostat (2005) high–low intervals for life expectancy at birth, both expressed in years. The UPE intervals are based on Table 8 (p. 64) of the final report that is available at www.stat.fi/tup/euupe/upe_final_report.pdf. The width of the 80 per cent intervals given there has been divided by 2 × 1.2816 (as in the case of Table 3.1) to arrive at the values of Table 3.3. Remarkably, the United Nations (2004) provides one mortality variant only, and is therefore omitted. The finding is that uncertainty in life expectancy as expressed by Eurostat has a width of two-fifths of a one-sigma interval. Under a normal assumption such an interval would have a probability of less than 31 per cent of covering the true value.

In substantive terms this means that should a user of the Eurostat forecast take the high–low intervals for the elderly as a ‘reasonable range’ (or, *a fortiori*, the lack of any range on the part of the UN), then the true uncertainty to be expected would be grossly underestimated.

Table 3.3. *The half-width of one-sigma intervals for life expectancy at birth as specified by the UPE project for the year 2049, and the half-width of the high–low interval of Eurostat (2005) for the year 2050.*

Source	UPE		ES	
Sex	Females	Males	Females	Males
Austria	2.9	3.3	1.2	1.4
Belgium	3.4	3.8	1.3	1.5
Denmark	3.9	3.9	1.4	1.5
Finland	3.3	3.7	1.2	1.4
France	3.4	3.9	1.3	1.5
Germany	3.6	4.2	1.4	1.6
Greece	3.1	3.5	1.3	1.6
Iceland	4.1	3.3	–	–
Ireland	3.7	3.7	1.6	1.8
Italy	3.3	3.5	1.2	1.4
Luxembourg	4.1	4.6	1.5	1.5
Netherlands	3.4	3.5	1.4	1.4
Norway	3.3	3.5	–	–
Portugal	3.6	4.1	1.6	1.9
Spain	3.5	4.0	1.1	1.6
Sweden	3.9	3.6	1.2	1.2
Switzerland	3.2	3.3	–	–
United Kingdom	3.5	3.7	1.8	1.2
Median	3.5	3.7	1.3	1.5
Standard deviation	0.33	0.35	0.19	0.18

Variance and autocorrelation of net migration

As discussed above, the uncertainty of net migration was formulated in terms of a model where cross-correlation over age was taken to be 1.0. The functions $g(j, x)$ are given in Figure 3.3 separately for males ($j = 1$, solid) and females ($j = 2$, dashed). As discussed by Alho and Spencer (2005) it makes sense to specify the distributions as proportional to gross migration. The estimates given were obtained from the years 1998–2002 for Sweden, Denmark and Norway, which have reliable data on gross migration.[6] The values for both sexes sum to 1.0, or $\Sigma_x g(j, x) = 1$, $j = 1, 2$.

The scales for total migration were assumed to be the same for both females and males. They start from an initial level for year 2003 and rise linearly to a final level ten years later. The initial level represents recent past volatility, but the final level is primarily judgemental: estimates

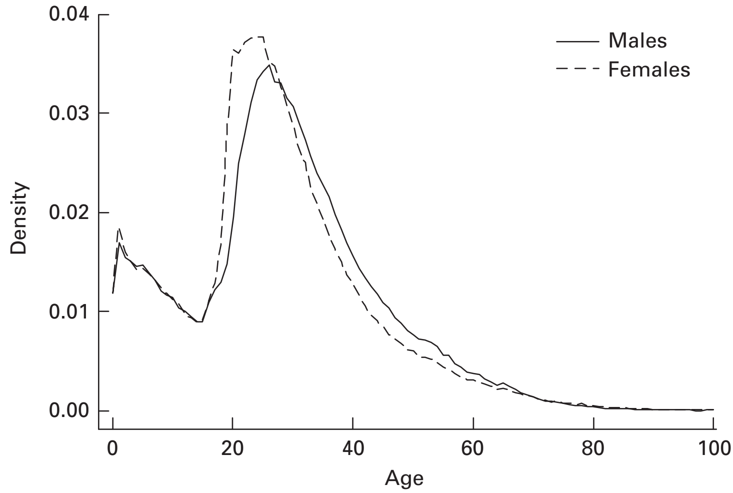

Figure 3.3. Distribution of gross migration by age: males and females.

obtained using time-series models were adjusted mostly downwards, but in some cases upwards, to reach an interpretable order of uncertainty among the countries. The cross-correlation of the sexes was assumed to be 0.90, based on empirical analyses by Alho (1998) of Finland and by Keilman, Pham and Hetland (2002) of Norway. An adjustment was made to the scales of both sexes to allow for less than perfect correlation.

Finally, the values for autocorrelations κ were determined for each country by a three-step procedure. First, time-series models estimated by Keilman and Pham (2004) were used to deduce the standard deviation of the predictive distribution of the cumulative net migration by the year 2050. Second, judgemental estimates of the same standard deviation were derived based on the assumption that the probability would be 2.5 per cent that the cumulative net migration would be negative by the year 2050. This calculation is based on the point forecasts of net migration as discussed by Keilman, Cruijsen and Alho (this volume, chap. 2); see also www.stat.fi/tup/euupe/. The former estimates were slightly higher but both were of the same order of magnitude, on average. However, there were some countries for which the time-series based estimates were implausibly high. These were first adjusted, and then the two estimates were averaged. This produced a set of candidates for standard deviations from which those values of κ could be deduced that would lead exactly to these standard deviations. We noted, however, that the estimates of κ arrived at in this way varied significantly across countries. Therefore, as the third step, we used a synthetic estimator (Rao, 2003) to pull the

Table 3.4. *Scale of total net migration in 2003 and 2049 per thousand of population in year 2000, the estimated autocorrelation κ as specified by the UPE project and the half-width of the high–low interval of Eurostat (2005), per population in year 2005.*

	Scale			ES
Year	2003	2049	κ	2050
Austria	3.2	4.0	0.21	2.5
Belgium	1.2	2.0	0.19	1.7
Denmark	1.1	2.0	0.18	1.5
Finland	0.7	2.0	0.34	0.9
France	2.1	4.0	0.13	1.0
Germany	3.4	4.0	0.20	2.7
Greece	6.0	6.0	0.17	2.6
Iceland	2.1	4.0	0.14	–
Ireland	2.9	5.0	0.22	3.6
Italy	3.3	5.0	0.26	1.4
Luxembourg	6.0	6.0	0.22	3.6
Netherlands	1.9	2.0	0.29	2.1
Norway	1.7	2.0	0.56	–
Portugal	6.0	6.0	0.17	3.2
Spain	3.3	5.0	0.28	2.1
Sweden	1.3	2.0	0.43	1.8
Switzerland	1.7	2.0	0.43	–
United Kingdom	2.7	4.0	0.17	2.2
Median	2.4	4.0	0.22	2.1
Standard deviation	1.7	1.6	0.12	0.85

national estimates towards the cross-country median of the values, by taking the average of the country-specific estimate and the median. The final values are given in column κ of Table 3.4.

In the same way as for mortality, the UN has not provided alternative variants for net migration at all. The finding from the comparison between the UPE and Eurostat assumptions (ES) is that the high–low intervals have a half-width which is slightly over one-half of the one-sigma interval specified by the UPE project. In fact, the uncertainty is only slightly lower than the UPE values corresponding to the first forecast year 2003. In this case the cumulative effect of the uncertainty specification of the Eurostat forecasts may come close to that of the UPE project because of the implicit assumption of perfect autocorrelation. Note also that in the case of migration the UPE specification was based much more on judgement than those for fertility and mortality.

As we are concentrating here on national populations, we do not discuss cross-country correlations in net migration or the other demographic processes. However, cross-country correlations can be, and were, estimated (Alho, 2005).

Implications for research on ageing

In this chapter we have shown that official forecasts continue to present alternative variants to the public that can only be considered as giving a 'reasonable range' if one believes that European demographics have become considerably less volatile than they have been during the past century. Note that this is so even though we have deliberately eliminated shocks caused by wars and similar 'outliers' from this assessment, by using robust estimation procedures.

For the population as a whole the chances are about two-thirds that the future will be in the official high–low range, but for both the young and the old the chances are much less. Interestingly, the UN and Eurostat arrive at approximately the same level of uncertainty for the total population by assumptions that are in stark contrast. For the UN the uncertainty comes entirely from births, whereas Eurostat also provides variants for mortality and migration.

The fact that the uncertainty in the number of elderly is clearly underestimated by both official forecasts means that policy-makers handling ageing-related problems, who rely on official forecasts, are unlikely to appreciate the true magnitude of uncertainty. As a result they may discount the possibility of future surprises whose consequences would be borne by the European citizenship.

Dealing with uncertainty in a thoughtful manner is harder than basing policy discussions on a fixed consensus view. It is not impossible, however. Papers in Auerbach and Lee (2001) have already charted a range of issues that come up. In this book we take a step closer to actual practice by presenting a range of macro-economic models that are used in the planning and evaluation of ageing policies.

Appendix: Conventional and stochastic population forecasts – pros and cons

Population renewal can be expressed as: (population at $t + 1$) = (population at t) + (births during t) − (deaths during t) + (net migration during t). In the cohort-component method the book keeping is done by age and sex. *Both conventional and stochastic forecasts use the same book keeping*. In the cohort-component method, births and deaths are generated via vital rates. In conventional forecasts the rates are specified as *numbers*, whereas in the stochastic forecasts they are taken to be *random variables*. From a logical point of view, a *conventional forecast is a special case of a*

Table 3.5. *Comparison of properties of conventional and stochastic population forecasts.*

	Conventional population forecast	Stochastic population forecast
Transparency	Involves addition, subtraction and multiplication only. The results may be checked by direct inspection, and understood with only modest education.	Forecasts are carried out by simulation. Checking is more complex. Understanding the results requires more education (including probability and statistics).
Simplicity of presentation	Middle variant may be sufficient to convey the main message. Alternative calculations may be provided, with one leading to a higher population, another leading to a lower population etc. These can be printed back-to-back in a book. Results for aggregated age-groups are produced by addition.	A full predictive distribution is approximated via computer simulation. Still, it may not be necessary to display more than the point forecast for a lay audience. Prediction intervals can be published for individual ages, but to produce intervals for aggregates a database of simulated values must (and can) be accessed.
Ease of preparation	Only a point forecast needs to be specified.	In addition, the specification of variances and correlations is required.
Interpretation and logical coherence of intervals	High and low intervals are (usually) interpreted as providing a 'plausible range'. But, there is no way to guarantee the comparability of the ranges for different ages, different sexes, different forecast years, etc. The results for different demographic functionals are necessarily incoherent.	Intervals with desired probability content for the size of any population aggregate and for the size of any demographic functional (like age ratios) at any future time can be provided.
Ability to handle uncertainty	Cannot handle uncertainty in a coherent and interpretable manner.	Uncertainty is handled in a coherent and interpretable manner.
Use of past data	Only trend estimates can be used. Judgement is necessary for questions like which models and which data periods to use. Knowledge of intermediate correlations (i.e. different from $\pm$ 1) across demographic variables cannot be used.	Can utilize information on varying trends, lack of fit, changes of volatility, etc. The uncertainty related to the choice of models and data periods can be incorporated statistically. Knowledge of intermediate correlations can be incorporated.

Table 3.5. (*cont.*)

	Conventional population forecast	Stochastic population forecast
Conditional forecasts and scenarios	Can represent scenarios that correspond to, for example, assumed effects of policy interventions or exogenous factors.	Can, in addition, incorporate information about the uncertainty concerning the effects of interventions or exogenous factors.
Sensitivity analysis	Can vary values of fertility, mortality or migration to see what the effect is.	Can, in addition, vary second moments.
Ease of computation	A large number of computer programs are available. There is extensive experience of their use. Results are obtained in seconds.	Several computer programs are available, but with differing capabilities. There is experience of their use from the past 10–15 years only. Results are obtained in minutes.

stochastic forecast, where all variance parameters are zero. From a practical point of view the main properties are as listed in Table 3.5.

NOTES

1 For a normally distributed variable the two measures are equal.
2 Here, and in the sequel, we will ignore the fact that the point forecasts can have different biases. Each forecast is assessed in turn as if it were unbiased, and only the width of the interval is considered. Note also that while we present results for year 2050 only, for the intermediate years the coverage of the official high–low intervals would typically be lower than at the end, as they assume perfect autocorrelation of errors over time.
3 Program for Error Propagation. For a description, see http://joyx.joensuu.fi/~ek/pep/pepstart.htm.
4 Note that the standard deviation in 2049 differs across countries because the estimates of initial volatility differ. For Norway an additional adjustment was made.
5 The UPE project interviewed experts in fertility, mortality and migration, in addition to carrying out statistical analyses.
6 For increased accuracy one might consider modifying this assumption for countries such as Spain, for which migration at retirement age is important.

REFERENCES

Alho, J. M. (1990). 'Stochastic Methods in Population Forecasting'. *International Journal of Forecasting*, 6: 521–30.

(1998). *A Stochastic Forecast of the Population of Finland*. Review no. 1998/4. Helsinki: Statistics Finland.

(2005). 'Simplified Approaches to Stochastic Multi-state Population Forecasts'. Paper presented at conference on *Stochastic Demographic Forecasting*, Salamanca, August.

Alho, J. M. and Spencer, B. D. (1997). 'The Practical Specification of the Expected Error of Population Forecasts'. *Journal of Official Statistics*, 13: 203–25.

(2005). *Statistical Demography and Forecasting*. New York: Springer.

Auerbach, A. J. and Lee, R. D. (2001), eds. *Demographic Change and Fiscal Policy*. Cambridge: Cambridge University Press.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis,* revised edn. San Francisco: Holden-Day.

Cannan E. (1895). 'The Probability of Cessation of the Growth of Population in England and Wales during the Next Century'. *The Economic Journal*, 5: 505–15.

Chib, S. and Greenberg, E. (1994). 'Bayes Inference in Regression Models with ARMA(p,q) Errors'. *Journal of Econometrics*, 64: 183–206.

DeGans H. A. (1999). *Population Forecasting 1895–1945*. Dordrecht: Kluwer.

Freedman, D., Pisani, R. and Purves, R. (1978). *Statistics*. New York: Norton.

Eurostat (2005). 'EUROPOP2004, Population Projections: EU25 Population Rises Until 2025, Then Falls'. News Release 448/2005, 8 April.

Keilman, N. (1990). *Uncertainty in National Population Forecasting: Issues, Backgrounds, Analyses, Recommendations*. Amsterdam: Swets and Zeitlinger.

(1997). 'Ex-post Errors in Official Population Forecasts in Industrialized Countries'. *Journal of Official Statistics*, 13: 245–77.

Keilman, N. and Pham, D. Q. (2004). *Empirical Errors and Predicted Errors in Fertility, Mortality and Migration Forecasts in the European Economic Area*. Discussion Paper no. 386. Oslo: Statistics Norway.

Keilman, N., Pham, D. Q. and Hetland, A. (2002). *Norway's Uncertain Demographic Future*. Social and Economic Studies 105. Oslo: Statistics Norway.

Lee, R. D. (1974). 'Forecasting Births in Post-transition Populations: Stochastic Renewal with Serially Correlated Fertility'. *Journal of the American Statistical Association*, 69: 607–17.

Lee, R. D. and Carter, L. R. (1992) 'Modeling and Forecasting the Time Series of U.S. Mortality'. *Journal of the American Statistical Association*, 87: 659–71.

Lee, R. D. and Tuljapurkar, S. (1998). 'Uncertain Economic Futures and Social Security Finances'. *American Economic Review*, 88: 237–41.

Modeen, G. (1934). 'The Future Development of the Population of Finland'. *Kansantaloudellinen aikakauskirja*, 6: 351–78. (In Finnish.)

National Research Council (2000). *Beyond Six Billion: Forecasting the World's Population*. Panel on Population Projections. Washington, DC: National Academy Press.

Rao, J. N. K. (2003). *Small Area Estimation*. New York: Wiley.

Shryock, H. S., Siegel, J. S. and associates (1976). *The Methods and Materials of Demography*, condensed edn by E. G. Stockwell. New York: Academic Press.

Stoto, M. (1983). 'Accuracy of Population Projections'. *Journal of the American Statistical Association*, 78: 13–20.

United Nations (2004). *The World Population Prospects: The 2004 Revision of the Population Database*. New York: United Nations.