Journal of Official Statistics, Vol. 13, No. 3, 1997, pp. 203–225

The Practical Specification of the Expected Error of Population Forecasts

Juha M. Alho and Bruce D. Spencer [1]

Forecasts of future population are uncertain, but the language of probability theory can be used to give the forecast users a realistic indication of their *ex ante* uncertainty. Knowing the magnitude of the expected error allows the users to prepare for contingencies they might otherwise neglect. The commonly offered high and low forecast variants fail to do the same because the users have no idea of how extreme they are. The purpose of this article is to review the central concepts needed in the analysis of uncertainty, and to present some approaches that are now available for the practical computation of the prediction intervals.

Methods of constructing prediction intervals include the application of formal time-series methods to past data, data analytic methods such as naive forecasting, analysis of *ex post* error of past forecasts, and subjective or judgmental analysis of error. Examples of each are given. In particular, we will comment on the compatibility of the results with our prior knowledge of the vital processes.

In the probabilistic handling of the *ex ante* error we must be able to combine nonlinearly the different sources of error without making unrealistic correlational assumptions, such as the (implicit) perfect correlation assumption of most national forecasts. These propagation of error calculations can be carried out either using analytic Taylor-series approximations or via simulation. We present empirical data on the forecast errors of mortality and fertility to motivate a simplified model that is suited to the programming of these vital rates and their errors with the present-day computational resources.

Key words: Demography; forecasting; population projections; propagation of error.

1. Introduction

The uncertainty of population forecasts can be assessed in two complementary ways. We can study the accuracy of past forecasts *ex post*, or after the future values have become known. Alternatively, we can try to assess *ex ante*, or beforehand, how accurate a forecast we are making is likely to be. The purpose of this article is to discuss practical ways of doing the latter. We will see that if a formal time-series model is used then the two approaches are closely related, when the model is correct.

Our discussion builds on the work of Lee (1974), Lee and Carter (1992), Stoto (1983), Cohen (1986), and Keilman (1990), who have made use of statistical time-series methods and modern data analytic techniques to assess uncertainty of population forecasts. Alho and Spencer (1991) and Lee and Tuljapurkar (1994) have developed methods for carrying

Acknowledgment: This article was prepared for presentation in the symposium "Analysis of Errors in Demographic Forecasts with Implications on Policy," in Koli, Finland, March 30–April 2, 1995. The authors gratefully acknowledge the support of this work through grant NIA-5-RO1-AG10156. Part of Alho's work was done while he was a Senior Research Fellow of the Academy of Finland.

out the propagation of error in forecasting. That is, if a forecast of the vital rates (fertility, mortality, migration) is given, how does the uncertainty of the rate forecasts translate into the uncertainty of the future population numbers? Since forecasting involves both judgmental and statistical elements, we will draw on our earlier work on modeling error (Alho and Spencer 1985) and expert judgment (Alho 1992a). We will also show how the concept of naive forecasts (Alho 1990) can be used to provide conservative, statistical *ex ante* assessments of uncertainty for purely judgmental forecasts. By a *conservative* assessment we mean a procedure that does not underestimate the level of uncertainty, but may overestimate it. Finally, we will use some results of Alho and Spencer (1997) to build simple representations for the forecast errors of the vital processes. A novel contribution is a stochastic model that combines a data-based model of uncertainty for the short run and a subjective specification for the long-run forecast.

Hoem (1973, p. 11) developed a detailed classification of forecast errors in population forecasting into six main categories: estimation and registration errors, pure randomness, random vital rates, unincorporated gradual changes in mean vital rates, gross shifts, and serious model misspecification. Keilman (1990, pp. 19–20) specifically adds errors in the jump-off population to the list. Both classifications are valuable in that they draw attention to important data analytic aspects of error. When the primary emphasis is on the *ex ante* error, we have found it useful to think of forecast error from the point of view of statistical modeling. Then, the following categories seem to be the most important ones:

> "(1) model misspecification: the assumed parametric model is only approximately correct; (2) errors in parameter estimates: even if the assumed parametric model would be the correct one, its parameter estimates will be subject to error when only finite data series are available; (3) errors in expert judgment: an outside observer may disagree with our judgments or prior beliefs about the parameters of the model, or the weight we give to our beliefs in forecasting; (4) random variation, which would be left unexplained even if the parameters of the process could be specified without any error: since any mathematical model is only an approximation one would expect there to be residual error." (Alho 1990, p. 523).

We will start in Section 2 by analysing the uncertainty of the rate of growth of total population, ignoring age structure. This analysis is feasible for most countries of the world at this time. Of course, it is deficient in that only the total population size, not age, is considered. Sections 2.2.4.–2.2.5. discuss the relationship between *ex post* and *ex ante* analyses. Next, in Section 3, we will show how simple calculations can be made for a sex- and age-structured population, if very simple models are used for the vital rates. The results are detailed enough to be implemented with modest effort in any country that uses cohort-component methods in forecasting. In Section 4 we conclude with practical recommendations for the producers of official statistics.

This article interweaves two themes. One examines particular kinds of models such as Brownian motion and random line models for forecast error in 2.2.1, ARIMA and regression models for the rate of growth in 2.2.2., scaled and Ornstein-Uhlenbeck models for error in 3.2, and a model for error for census survival rates in 3.3. We fit the models into a general framework for describing the propagation of error. The second theme illustrates the practical use of this framework: Section 2 analyses the total population of the U.S.,

Section 3 uses data from the Nordic countries, the U.S., and the Netherlands to discuss error models for age-structured populations.

2. A Simple Analysis of Uncertainty for the Total Population

Let $V(t)$ be the size of population at time $t = \ldots, -2, -1, 0, 1, 2, \ldots$ Assume that $V(t) > 0$ for all t, permitting us to define $v(t) = \log(V(t))$ and $\delta(t) = v(t) - v(t-1)$. Suppose that $t = 0$ is the jump-off time of the forecast, i.e., it is the last time for which we have data. As a consequence $v(t) = v(0) + \gamma(t)$, where $\gamma(t) = \delta(1) + \ldots + \delta(t)$.

Denote the forecast of $V(t)$ by $\hat{V}(t)$ for $t > 0$. Similar notation is used for other forecasted or estimated values. Define the forecast errors $\epsilon_v(t) = v(t) - \hat{v}(t)$, $\epsilon_\delta(t) = \delta(t) - \hat{\delta}(t)$, and $\epsilon_\gamma(t) = \gamma(t) - \hat{\gamma}(t)$. Then, we can write

$$\epsilon_v(t) = \epsilon_v(0) + \epsilon_\gamma(t)$$

We will now go over the major steps involved in any assessment of the uncertainty of population forecasts in this simple setting.

2.1. *Errors in the jump-off population*

The term $\epsilon_v(0)$ represents jump-off error. In a country like the U.S. or the U.K., the main source of population data are the decennial censuses. Suppose that $t = 0$ corresponds to the time of a census. We can forecast three different types of populations. *(1) The population as enumerated in the census.* In this case $\epsilon_v(0) = 0$, by definition. *(2) The population that should have been enumerated in the census according to the law.* Estimates of census undercount depend on statistical modeling and judgment, so all four sources of error we have defined are present: erroneous parametric models may be used in dual system estimation or demographic analysis to assess undercount; parameters of such models are subject to sampling error; judgments concerning some census procedures such as imputation, or some modeling assumptions such as lack of correlation bias, are uncertain; and the process of enumeration in the census and the possible post enumeration survey can be inherently random (Mulry and Spencer 1993). *(3) The de facto population that is actually present in the country.* This is different from the *de jure* population that is legally resident in the country, and which is targeted in the census. In this case, many supplementary statistical estimates are needed for the military and diplomatic personnel, seasonal workers etc. If $t = 0$ does not correspond to the time of a census, additional estimates are also needed to update census figures by subsequent births, deaths, and migration (Spencer 1980).

In some other countries,such as the Nordic countries, population data are obtained from a population register. The situation parallels that described above, but the magnitudes of error may be different. If we forecast the population as present in the register (cf., case (1) above), then $\epsilon_v(0) = 0$, by definition. It is known that the Nordic population registers are not fully free of error, but it is believed that the errors are negligible. Hence, even if the population to be forecasted is the population that *should* be in the register according to the official rules (case (2)), we may often take $\epsilon_v(0) = 0$ in practice. Since the registers operate continuously, the establishment of the jump-off population is easier than in the case of a census based system. However, if the objective is to forecast the *de facto* population

(case (3)), then similar additional estimates are needed as in the census-based system, so all the four sources of error are potentially present.

Our impression is that in many industrialized countries the statistical agencies ostensibly prepare forecasts for the true *de jure* population (2), but they are prepared as if the enumerated population (1) were known without error. For many government users of the forecasts this is convenient, because with this method there is only one set of figures that needs to be used in administration. However, we suspect that many non-government users believe that forecasts are prepared for the actual population (3).

To proceed with the propagation of error it is necessary to say something about the distribution of $\epsilon_v(0)$. In the simplest case we may want to specify the *mean squared error* (MSE). We have $E[\epsilon_v(0)^2] = \sigma_v(0)^2 + E[\epsilon_v(0)]^2$, where $\sigma_v(0)^2 = Var(\epsilon_v(0))$. If erroneous models are used in estimation, then we would typically have bias, or $E[\epsilon_v(0)] \neq 0$. Sampling error in parameters and residual random variation contribute to $\sigma_v(0)^2$. Errors in judgment would typically cause both bias and influence variance. From the total error analysis in Mulry and Spencer (1993) of the 1990 U.S. census, one can conclude that it would be reasonable to assume that $E[\epsilon_v(0)] = 0$ and $\sigma_v(0) = .0036$, if the *bias-corrected*, adjusted census count is used as the estimate of the U.S. population (case (2)). The primary sources of uncertainty are modeling error and sampling error. (Mulry and Spencer 1993, Tables 1 and 2, p. 1082).

2.2. *Errors in the forecasted growth rate*

The growth rate $\delta(t)$ is essentially equal to: *(crude birth rate) – (crude death rate) + (net migration "rate")* during the period $[t-1, t]$. In the industrialized countries forecasts for the rate of growth are obtained as a by-product of a cohort-component forecast. In countries with poor data, the crude vital rates, or the past rates of growth may be the basis of forecasting. In either case, the simplest assessment of uncertainty derives from the rate of growth. Next we present four different ways of arriving at a probabilistic characterization of $\gamma(t)$, and a simple way to quantify the effect of choice of data period for model fitting.

2.2.1. Judgmental specification

A method of forecasting that has been used in demography at least since the time of Whelpton (1947) is to set a target value at some future year $t = T$ for the rate of interest, in this case $\delta(T)$. As already noted by Whelpton, a possible source of information may be another country that is thought to lead in demographic development. Denote the target value by $\hat{\delta}(T)$. The usual approach would supplement the middle target by high and low targets $\delta_H(T)$ and $\delta_L(T)$. However, to develop a probabilistic version of the forecast, we would give a variance $\sigma_\delta(T)^2 \equiv Var(\epsilon_\delta(T))$, instead. Think of the true value as $\delta(T) = \hat{\delta}(T) + \epsilon_\delta(T)$. The simplest way to complete the specification of the *predictive distribution* of $\delta(T)$ would be to assume a distributional form. The normal (Gaussian) distribution is by far the easiest to handle technically. Provided that the forecast is an unbiased one, the mean is given by $\hat{\delta}(T)$. We can specify a variance as follows. Suppose that based on past level of error in forecasts, experience in other countries etc., we think that the probability is $(1-p)$ that the future value is inside the interval $[\delta_L(T), \delta_H(T)]$ centered around

$\hat{\delta}(T)$. Then we must have $\sigma_\delta(T) = (\delta_H(T) - \delta_L(T))/(\Phi^{-1}(1 - p/2) - \Phi^{-1}(p/2))$, where Φ is the distribution function of the standard normal distribution.

Suppose now that we have derived the model $\delta(T) \sim N(\hat{\delta}(T), \sigma_\delta(T)^2)$. A particularly simple way to derive a joint distribution for $\delta(t)$ with $1 \leq t \leq T$ is to assume a linear change for the mean, or $\hat{\delta}(t) = \hat{\beta}_0 + \hat{\beta}_1 t$, where $\hat{\beta}_0 = \hat{\delta}(0)$ and $\hat{\beta}_1 = (\hat{\delta}(T) - \hat{\delta}(0))/T$. This implies that

$$\hat{\gamma}(t) = \hat{\beta}_0 t + \hat{\beta}_1 t(t+1)/2$$

If a smooth start from the jump-off value and a leveling off at the target year is desired, then the formula of Andrews and Beekman (1987, p. 21) can be used instead of the linear curve. The forecast functions of ARIMA(1,1,0) processes provide alternate models that level off asymptotically (Box and Jenkins 1976, p. 155).

Having decided on the linear (or some other) model for the mean we need to specify the error structure. A particularly simple but fairly realistic model says that $\epsilon_\delta(t) = \xi_\delta(1) + \ldots + \xi_\delta(t)$, where the summands $\xi_\delta(t) \sim N(0, \sigma_\delta^2)$ are independent (Section 2.2.2. considers the correlated case). Under this *Brownian motion*, or *random walk* specification (cf. Alho and Spencer, 1990a), the error of the forecast increases with t. The forecast error is highly autocorrelated, but not perfectly correlated, so it allows for some cancellation of error over time. Under this specification we have $\sigma_\delta^2 = \sigma_\delta(T)^2/T$. It follows that $\sigma_\gamma(t)^2 = \sigma_\delta^2(2t+1)(t+1)t/6$. A less realistic model for the errors would take $\epsilon_\delta(t) = \epsilon_\delta(T)t/T$. This would be the same as assuming that $\hat{\delta}(t)$ is a straight line, but with a random slope. Hence, it can be termed a *random line model*. In this case, there could be no cancellation of forecast error over time, and we would have $\sigma_\gamma(t)^2 = \sigma_\delta(T)^2[(t(t+1)/(2T)]^2$. We see that the variance increases with the 4th power of t, whereas under the Brownian motion model it increased slower, with the 3rd power of t.

2.2.2. Formal models for forecasts

Suppose we have a time-series of values $\delta(t)$ *for* $t = -n, \ldots, -1, 0$ available. In the developed countries predictable changes in the age composition may provide explanations for changes in growth rates. In the developing countries we may only have a time-series of growth rates. In that case we have little other alternative than to model the $\delta(t)$ using univariate time-series techniques, such as ARIMA models or polynomial regression. These techniques provide us with both the point forecasts $\hat{\delta}(t)$ and estimates of the distribution of $\epsilon_\delta(t)$ for $1 \leq t \leq T$, provided that we can make sufficient assumptions about the distribution of the $\delta(t)$ (a functional form for the mean, stationarity of deviations from the mean, possible normality). Unfortunately, such time-series approaches can be problematic. We will illustrate this by presenting four alternate analyses of the U.S. growth rate data: (I) stationary model, (II) random walk model, (III) ARIMA model, (IV) regression model. We will then show how the error analysis can be done in the regression case.

(I) Pflaumer (1992) studied the series $\delta(t)$ for the U.S. along the lines that had been earlier suggested by Cohen (1986). The autocorrelation function declines slowly and the graph of the data suggests a declining trend (Pflaumer 1992, Exhibits 7 and 8, pp. 334–335). Nevertheless, the author fitted the stationary model: $\delta(t) = c + \phi\delta(t-1) + a_t$, where the a_t are i.i.d. error terms with $E[a_t] = 0$. The estimates were $c = .0029$ and

$\phi = .782$. The model implies that the growth rate *increases* from the 1988 value of about .009 to an asymptotic value, which is equal to the average of the data period 1900–1988, or .0133. This produces implausibly high forecasts after a few forecast years (see column PF of Table 1). For example, they are out of the 95 per cent prediction interval of Lee and Tuljapurkar (1994, p. 1185) by the year 2000. The Lee-Tuljapurkar forecast is given in column LT of Table 1. [The extremeness of the Pflaumer point forecast is further illustrated by the fact that at year 2050 it agrees closely with the most extreme scenario considered by Ahlburg and Vaupel (1990) which assumed a two per cent reduction of mortality annually, 1–2 million immigrants annually, and a 50-year cycle of variation of the total fertility rate from a ''boom'' of 3.2 to a ''bust'' of 1.84. While not impossible, these assumptions hardly qualify as the most likely point forecast.]

(II) Since both the graph and the autocorrelation function suggest that the series $\delta(t)$ is nonstationary, one might consider a stationary model for its first differences. The simplest such model assumes that $\delta(t)$ is a random walk. Random walk is a martingale which means that the best forecast for all future times is the current value. Since we had approximately $\delta(t) = .009$ at $t = 0$ (year 1988), a reasonable forecast for $\gamma(t)$ is $\hat{\gamma}(t) = .009t$. Since $V(t) = 246.1$ millions at $t = 0$, we get the forecast $\hat{V}(t) = 246.1 \times \exp(.009t)$. These values are in the column RW. The forecasts remain easily inside the Lee-Tuljapurkar 95 per cent prediction interval to 2070, at least. Under the assumption of a random walk, we have $\epsilon_\delta(t) = \xi_\delta(1) + \ldots + \xi_\delta(t)$, where $\xi_\delta(t) \sim N(0, \sigma_\delta^2)$ are i.i.d. From our data we get the estimate $\sigma_\delta = .00472$. Now we are back in the Brownian motion model of Section 2.2.1., if we take $\hat{\beta}_0 = \hat{\delta}(0)$ and $\hat{\beta}_1 = 0$. I.e., apart from the uncertainty connected with the jump-off population, we know the full predictive distribution of the forecast of the total population.

(III) Instead of assuming the increments to be uncorrelated we could build an ARIMA model for them. Based on the autocorrelation structure, an MA(1) model seems reasonable for the first differences of $\delta(t)$. This is the same as fitting the model ARIMA(0,2,1) to the series $v(t)$. These forecasts are in column IMA. We find that the added ''refinement'' that takes into account the autocorrelation structure does not change the forecast much. However, as can be seen from column IMC, adding a constant term to the model causes the forecast to be very close to the Lee-Tuljapurkar forecast. In either case the calculations needed for the assessment of uncertainty are a bit more complex. In Alho and Spencer (1997) we give equations for carrying them out. However, in the simple situation at hand, we get the same results from most statistical packages. *Minitab* for example, gives the 95 per cent prediction intervals for the ARIMA(0,2,1) forecast of $v(t)$. By exponentiating these we get the intervals for the population size $V(t) = \exp(v(t))$.

(IV) An alternative ''old fashioned'' way to handle the nonstationarity is to use ordinary least squares (OLS) to fit the model $\delta(t) = \beta_0 + \beta_1 t + \epsilon_\delta(t)$ to the data $t = -n, \ldots, -1, 0$. This closely parallels the judgmental specification of the predictive distribution under a linear model for the mean. Using the U.S. total population $V(t)$ for the years 1900–1990 we get the estimates $\hat{\beta}_0 = .00983$ and $\hat{\beta}_1 = -.0000762$. Using the formula of $\hat{\gamma}(t)$ given in Section 2.2.1., we have calculated the forecast given in column REG of Table 1. We find that this forecast is close to the Lee-Tuljapurkar forecast for at least 60 years.

Again a more ''refined'' version of the regression model would be to use the estimated residuals from the OLS fit to estimate a correlation structure for the error terms. This could

be used by itself to improve the predictor, or it could be used as a basis for a generalized least squares (GLS) estimator. Either procedure would be particularly helpful, if the fitted line were far from the observed value of $\delta(t)$ at jump-off. This is not the case in our data, so the point forecast would not change appreciably.

2.2.3. Illustration of error analysis: regression

Under the regression method, we can use the estimated covariance matrix of the parameters to calculate prediction intervals for the future population. The covariance matrix is calculated as a part of the regression fit, and it can typically be easily accessed in any statistical package. Define $\beta = (\beta_0, \beta_1)^T$ and let $\hat{\beta}$ be its estimator. Define $\mathbf{W}(t) = (t, t(t+1)/2)^T$, so we can write $\epsilon_\gamma(t) = \epsilon_\delta(1,t) + \mathbf{W}(t)^T(\beta - \hat{\beta})$, where $\epsilon_\delta(1,t) \equiv \epsilon_\delta(1) + \ldots + \epsilon_\delta(t)$. To reflect error source (2) in Section 1, define $Cov(\hat{\beta}) = \mathbf{C}$. If we assume that the errors $\epsilon_\delta(t)$ are i.i.d., we get the simple equation

$$Var(\epsilon_\gamma(t)) = \sigma_\delta^2 t + \mathbf{W}(t)^T \mathbf{C} \mathbf{W}(t)$$

The first term on the right is of the order of t, or $O(t)$, and the second is $O(t^4)$; the latter will dominate in long-range forecasts. Ignoring the jump-off error, we construct a 100(1-p) per cent prediction interval for $V(t)$ as $(249.95)\exp(\mathbf{W}(t)^T\hat{\beta} \pm z_{1-p/2} Var(\epsilon_\gamma(t))^{1/2})$, where $z_{1-p/2}$ is the $1 - p/2$ fractile of the standard normal distribution.

We caution that these intervals do not reflect possible error in the model, which can be substantial. Often error in modeling produces too narrow prediction intervals. Error in the modeling *can* produce intervals that are too wide, depending on how the models are fitted. For example, fitting, say, a quadratic function to the past data would give ever so slightly better fits but would have larger prediction intervals. Although model-selection procedures might clearly indicate that a smaller submodel is adequate, the model-selection procedures themselves can be misleading (and will be, with some probability). The point is that overfitting can occur, and will increase the variances of the prediction errors. Allowances for model uncertainty can be made (cf., Draper 1995), but the details are mostly beyond the scope of this article.

Overly narrow prediction intervals do not seem to be the case here, however. Using the U.S. data we get the estimates $\mathbf{W}(t)^T\hat{\beta} = (.00948)t - (.0000381)t(t+1)$, and $Var(\epsilon_\gamma(t)) = (.0000207)[t + (.0437)t^2 - (.000733)t^2(t+1) + (.00000412)t^2(t+1)^2]$. For about 15–20 years the forecast intervals produced this way are comparable to those given by Lee and Tuljapurkar (1994), but after that the fourth order terms of the variance formula start to dominate, producing quadratically growing standard deviations for the prediction error. These are soon unacceptably big. We will take up this problem again in Section 3.

In this particular model, it is easy to analyze the errors. However, in many propagation of error problems it is simpler to resort to simulation. It is immediately clear that we can simulate the variables $\epsilon_\delta(1,t)$ by simulating the summands $\epsilon_\delta(t)$, but it is less obvious how to take care of the estimation error $\beta - \hat{\beta}$. A simple solution is to approximate the error by $\epsilon_\beta \sim N(0, \mathbf{C})$, and take ϵ_β and the $\epsilon_\delta(t)$ to be independent. This yields the approximate model $\epsilon_\gamma(t) \approx \epsilon_\delta(1,t) + \mathbf{W}(t)^T\epsilon_\beta$ that can easily be simulated. Bayesian readers will recognize this interpretation of the standard error as an approximation based on a flat prior on β, and a normal approximation for the distribution of $\hat{\beta}$. Note that the variance of ϵ_β decreases with the length of the base period on which the model is fitted. Offsetting this variance

reduction is the concern that the model does not apply too far back in the past. Thus, in practice, official forecasts are based on relatively short series. One may well question why, if one doubts the validity of the model far back in the past, one should accept the validity of the model as far or farther in the future.

We conclude that the standard statistical methods can be used to arrive at a formally correct predictive distribution, provided that the model is correct. Hence, any statistical agency can perform these analyses, as long as the growth-rate data are available.

2.2.4. Evaluation of data-period bias

Data on the U.S. growth rates illustrate another important aspect of the assessment of uncertainty. The U.S. growth rates first declined until the 1930s, then they increased until the 1950s, then they have declined. Restricting the data period to 1930–1988 would suggest that the trend in the growth rate is up. Restricting the data period to 1950–1988 would suggest that the trend is down. Therefore, we should be sensitive to the fact that we do not let our preconceived ideas about the future dictate the choice of the data period. On the other hand, since some choices always have to be made, it is important to recognize that they may create a bias in the forecast (''error in expert judgment''; cf. Section 1). This is analogous to the effect of the jump-off year in the analysis of *ex post* errors that has been noted by Stoto (1983) and Keilman (1990).

The magnitude of such a bias can be assessed by considering alternative data periods. Judgment is always involved in such a choice, but we can typically decide that some data are clearly too old to have any information for the current forecasts. The cut-off point may depend on the length of the forecast horizon: the longer the horizon the more data one would want to have! Similarly, we typically can decide the earliest data period that we would certainly want to keep. The actual choice is somewhere in between, but it is hard to say definitely where. In both ARIMA and regression based analysis we may systematically go through all such alternative starting values s of the data period that are deemed reasonable. In the U.S. data we have analyzed, these values could be 1900,...,1950, for example. We can now fix a future year $t > 1990$. We calculate the forecast of $\gamma(t)$, denote it by $\hat{\gamma}(t, y)$, for all such starting years y. Under the assumption that at least one of the chosen starting years yields an unbiased forecast (given the model we use), we can get a bound for the data period bias. Let $\hat{\gamma}(t)$ be the preferred forecast and define $d_\gamma(t) \equiv \max\{|\hat{\gamma}(t, y) - \hat{\gamma}(t)|\}$. Then, we have the inequality, $E[d_\gamma(t)] \geq |\ bias\ of\ \hat{\gamma}(t)|$. To illustrate the method, we used the ARIMA(0,2,1) model with a constant on the U.S. growth rate data, and defined $\hat{\gamma}(t)$ as the value with starting year 1950. The following estimates were obtained: $d_\gamma(t) = .019$ at $t = 10$; .070 at $t = 20$; and .153 at $t = 30$. In other words, the data period bias might increase from 2 per cent to 15 per cent during the first ten to thirty forecast years. For this model the bias is bigger than the standard deviation of the prediction error (obtained from the ARIMA model with constant): .008 at $t = 10$, .025 at $t = 20$, and .047 at $t = 30$. If we defined $d_\gamma(t)$ as the average of the values obtained with the starting years $y = 1900, \ldots, 1950$, the bounds for the data period bias would drop by approximately one third. Interestingly, using the ARIMA(0,2,1) model *without* a constant term, and starting year 1950, yields much lower values: $d_\gamma(t) = .0055$ at $t = 10$; .011 at $t = 20$; and .016 at $t = 30$, but nearly twice as high standard deviations. Again, the use of average predictions would cut the biases by nearly a half.

It is interesting to note that the use of the constant term in ARIMA modeling provides a much stronger control of the forecast than the model without the constant, but at the same time it exposes us to a greater risk of data-period bias.

Having estimates $d_\gamma(t)$ available permits us to incorporate part of the uncertainty of the data period into our assessment of uncertainty for all t. This can be done in two ways. We might simply replace our old estimate of $Var\,(\epsilon_\gamma(t))$ by the MSE $Var\,(\epsilon_\gamma(t)) + d_\gamma(t)^2$, and proceed in the analysis as before, but with an inflated variance. This is similar to the method used by Stoto (1983), who breaks the total error into the sum of error due to the jump-off period and (a considerably smaller) residual error. This approach lends itself easily to simulation. Use of intervals based on MSE rather than variance can have appropriate coverage probabilities if the bias is small relative to the standard error (cf. Cochran 1977, p. 15). Depending on the model, this may or may not be the case here. Alternatively, we may treat bias separately from the variance.

As a complement to these observations we note that *it is dangerous to forecast for a longer period than the data period.* In fact, prudence suggests clearly shorter forecasting periods, in order for those features of the data that occur fairly rarely to have time to manifest themselves. With the U.S. data of nearly 90 years, a forecast of the total population may be carried out up to, say, 30 years into the future, and we still have a reasonable basis for assessing its uncertainty using essentially data-driven methods. For longer forecast periods it is critical to take into account the fact that the models used for forecasting may be biased over the longer term.

2.2.5. Ex post estimates

Stoto (1983) suggested the use of the empirical distribution of the forecast error to characterize the uncertainty of forecasts. Specifically, one should study past forecasts, in order to detect the error in the rate of growth for each jump-off year and lead-time combination. The method assumes nothing about the method of forecasting, nor does it require a probability model for the series $\delta(t)$. Using data from the industrialized countries Stoto concluded that the average growth rate has an *ex post* standard deviation of .0028–.0052, depending what data were used. Stoto essentially used the random line model mentioned in Section 2.2.1. for translating the *ex post* estimates into *ex ante* prediction intervals.

Spencer (1989) and Cheeseman Day (1993) have studied the RMSE of the past forecasts of the U.S. Bureau of the Census. They found that the forecasts made in 1950–1971 had an RMSE that increased from .0015 during the forecast year $t = 1$, to .0046 for $t = 15$. However, for the forecasts made after 1972 the RMSE showed a non-monotonic pattern from the value .0015 at $t = 1$ to .0003 at $t = 7$, and to .0020 at $t = 15$. The accuracy of the short term forecasts has not improved. Overall, they conclude that if the RMSE of the forecasts made during the past 30 years can be used as an estimate of *ex ante* errors, then the high-low forecast intervals of the U.S. Bureau of the Census for the total population can be viewed as approximately 67 per cent prediction intervals.

Although the method is intuitively very appealing, it is limited for many reasons. First, to be useful in *ex ante* error analysis, one must assume that the errors in the forecasts of the future growth rates are of the same order of magnitude as those of the past forecasts. All forecasters would not be willing to accept this premise. Second, the method requires that many past forecasts are available. Otherwise, the estimates of past errors are unreliable.

This is exemplified both by the wide range of error estimates Stoto got using different data sets, and the experiences from the forecasts of the U.S. Bureau of the Census. Third, it is cumbersome to extend the method to populations disaggregated by age, because the joint errors at different ages must be estimated.

2.2.6. Volatility based estimates

Most demographic forecasts do not use formal statistical methods, and in many countries only a small number of past forecasts are available for an *ex post* error assessment. Therefore, it would seem that only a judgmental specification of *ex ante* uncertainty is feasible. However, in Alho (1990) we developed a method that can be used as long as we have a past time series of the process of interest available. Typically we can devise a simple method of forecasting that produces reasonable forecasts.

In the case of the growth rate, a reasonable forecast is obtained by assuming that the current growth rate will continue indefinitely. We call these forecasts *naive*, because they do not utilize any expert judgment about the conditions during the time of the forecasting. We can produce naive forecasts of the growth rate using each of the available past years as the jump-off year, and find out what its error would have been. This gives us a distribution of forecast errors for a range of lead times. From the forecaster's point of view *error assessments based on these distributions are conservative for any method of forecasting that he/she believes to be more accurate than the naive method.* For example, Keyfitz (1981) examined the 20-year growth rates for 90 countries with varying rates and found an RMSE of .90 when the forecast assumed the average rate for the preceding five years, an RMSE of .60 for a somewhat less naive forecast *versus* .48 for the actual forecasts. Therefore, we can get an upper bound for the error of our preferred forecasting method from the error of the naive method. The errors of the naive forecasts are closely related to the *volatility* of the time series in question. In highly volatile time series naive forecasts have large errors, whereas in series of low volatility the errors are typically small.

We applied the method to the U.S. growth rates for the years 1900–1990, as follows. First, we calculated the differences of the series for lags 1–60. These are the *ex post* errors of the naive forecast for the annual growth rate for lead times 1–60. We then calculated the RMSE of the naive forecast for each lead time. We found that the RMSEs increased from .0047 at $t = 1$ to .0083 at $t = 29$. Then, they decreased to the value .0051 at $t = 49$, and increased again. The nonmonotonicity is similar to the one mentioned in Section 2.2.4. We smoothed the empirical estimates using RSMOOTH of *Minitab* to remove some local irregularities. The values of the first 29 years were used in volatility estimation. We assumed that forecast errors in growth rates were built of independent increments, and estimated variances for the increments. This specifies a complete probability structure for the growth rates. In particular, we can derive $Var\,(\gamma(t))$ from the variances of the increments. We omit the details (similar to those that will be discussed in Section 3.2.), but note that the standard deviation of the prediction error for the naive forecast was .021 at $t = 10$; .068 at $t = 20$; and .121 at $t = 30$. The corresponding standard deviations of the Lee-Tuljapurkar forecast under a lognormal interpretation are .019, .044, and .071; cf., Lee and Tuljapurkar (1994; Table 2, p. 1185). We would expect the naive forecast of the growth rate to be less accurate than the Lee-Tuljapurkar cohort-component forecast, but it is interesting to note that initially the difference in performance is not overwhelming. The

regression approach discussed in Section 2.2.2. yields the estimates .017, .057, and .131. However, interestingly the corresponding ARIMA(0,2,1) based estimates .034, .083, and .145, are the largest of all. Adding a constant term decreases these slightly.

Note that the latter methods yield *ex ante* estimates of error that exceed the *ex post* errors of the naive method.

2.3. *Propagation of error for the total population*

We have noted approaches to the assessment of the uncertainty of the jump-off population. We have also described four methods of specifying a predictive distribution for the future rate of growth. We saw that under a formal statistical model the *ex ante* error estimates derive essentially from the past lack of fit of the assumed model. This implies that there is a close relationship between *ex ante* and *ex post* error estimates, provided that the model is correct, and the data-period bias is small. As shown in Section 2.2.5., when the model is not correct, then the model-based *ex ante* error can be bigger than the *ex post* estimates of past error. Note also that for the purpose of understanding the predictive distributions it is the most natural to interpret the probabilities subjectively. A concrete example of this approach was the simple way standard errors could be treated in simulation, under the regression model.

In the simple case of forecasting the total population via the growth rate, the propagation of error calculation simply means that we combine the assessments of the jump-off error and error in the growth rate, to get an overall estimate of the predictive distribution of $\epsilon_v(t) = \epsilon_v(0) + \epsilon_\gamma(t)$. We have,

$$Var\,(\epsilon_v(t)) = Var\,(\epsilon_v(0)) + Var\,(\epsilon_\gamma(t)) + 2Cov\,(\epsilon_v(0), \epsilon_\gamma(t))$$

Therefore, once we have estimates of the variances available, we still have to specify the covariance. If the population size has been consistently over- or underestimated at the same rate, then the error can be uncorrelated with the forecast error of the growth rate. However, if there are changes in the error of the population numbers, the two sources may be positively correlated: if the jump-off population happens to be underestimated to an unusual extent, then so is the rate of growth at jump-off.

The basic principle of the probabilistic propagation of error (as opposed to deterministic calculations) is that we can entertain arbitrary correlations between $\epsilon_v(0)$ and $\epsilon_\gamma(t)$, not just the perfect correlation.

Table 1. Forecasts of the U.S. total population by Lee and Tuljapurkar (1994) (LT), by Pflaumer's AR(1) model for the growth rate (1992) (PF), by a random walk model for the growth rate (RW), by an ARIMA(0,1,1) model for the growth rate (IMA), by an ARIMA(0,1,1) model with a constant term (IMC), and by a linear regression model (REG) for the growth rate

Year	LT	PF	RW	IMA	IMC	REG
2000	273.8	284.9	274.2	275.1	272.7	274.6
2010	294.8	324.5	300.0	302.7	294.2	299.5
2020	316.0	373.2	328.2	333.1	313.8	324.1
2030	336.3	425.1	359.1	366.5	330.9	348.1
2050	371.5	556.8	430.0	443.9	355.7	392.4

3. Propagation of Error in a Simple Age-Structured Setting

We will now extend the above discussion to cover an age-structured population. The same four ways of specifying predictive distributions, as above, are relevant. Similarly, the basic principle of probabilistic propagation of error applies. The only difference is that we now have a multidimensional problem.

We start out by summarizing findings from Alho and Spencer (1997), in which we study extensively both empirical and theoretical aspects of the forecast errors of the vital processes of mortality, fertility, and migration. We will then present a simple model under which the propagation of error can be carried out either via analytical approximations or via simulation. The setting is sufficiently general to contain the empirical models that Lee and Tuljapurkar (1994) estimated from the U.S. data as special cases. It is also well-suited to the formulation of stochastic versions of judgmental, deterministic forecasts.

We will frequently use naive forecasts to characterize the errors of forecasts in the industrialized countries. We assume that in the industrialized countries, a naive forecast of fertility is today's value. In the case of mortality, a naive forecast assumes that the recent rate of decline will continue. In the case of migration, a naive forecast may be the past average level.

3.1. *What are plausible error structures?*

In Alho and Spencer (1997) we considered U.S. mortality in ages 65–69, 70–74, 75–79, 80–84, and 85+, and white U.S. fertility in ages $14, 15, \ldots, 46$. We found that *all series were nonstationary*. Many of them would have to be differenced twice, if differencing were to be used to render them stationary. The rates are similar in other industrialized countries, such as the Nordic countries. We conclude that plausible autocorrelation structures for the forecast errors of age-specific mortality and fertility rates must be qualitatively similar to the autocorrelation structures of the forecast errors of processes whose first or second differences are stationary. Or, *the autocorrelations are positive, and high.*

To study the cross-correlations of the forecast errors of the logarithms of age-specific mortality rates, we produced (essentially naive) forecasts with an ARIMA(1,1,0) model with a constant term. The correlations varied, but we concluded that a positive constant correlation across ages provides a reasonable description of the empirical data. The results were similar to those we found for cause-specific mortality earlier (Alho and Spencer 1990b, pp. 223–225).

In the case of fertility, the cross-correlations between different ages seemed to fall off exponentially with increasing age difference. Hence, an AR(1) process across age can serve as an approximation. We also studied the cross-correlations of age-specific fertility rates across the Nordic countries. The result was that inasmuch as official forecast errors resemble naive forecasts, the forecast errors of age-specific fertility rates will be expected to be positively correlated, in the long run, across countries. In the short run, differential timing would, nevertheless, produce different errors in the different countries. This suggests that a positive constant correlation might be assumed.

Joop de Beer (1993) has used time-series methods to study net migration in the

Table 2. The estimated standard deviations $S(j,t)$ of the annual increments of the forecast error for the logarithm of the age-specific U.S. female mortality during forecast periods 1–5, 6–15, 25–35, 45–55, 85–95 years based on the perfect correlation of error increments assumption and independence of error increments assumption

	Perfect correlation Period				
Age	1–5	6–15	25–35	45–55	85–95
0	.016	.012	.003	.003	.002
1–4	.012	.011	.003	.003	.003
5–9	.013	.013	.004	.004	.003
10–14	.011	.010	.003	.004	.003
15–19	.007	.006	.003	.003	.003
20–25	.006	.005	.003	.003	.003
26–29	.008	.007	.003	.003	.003
30–34	.012	.011	.004	.004	.004
35–39	.014	.013	.005	.005	.004
40–44	.012	.012	.005	.005	.004
45–49	.010	.010	.005	.005	.005
50–54	.007	.006	.005	.005	.005
55–59	.006	.005	.005	.005	.005
60–64	.006	.006	.005	.005	.005
65–69	.007	.006	.005	.005	.004
70–74	.007	.007	.004	.004	.004
75–79	.008	.007	.004	.004	.003
80–84	.008	.008	.004	.003	.003
85–89	.008	.008	.003	.003	.003
90–94	.007	.007	.003	.003	.003
	Independence Period				
Age	1–5	6–15	25–35	45–55	85–95
0	.036	.057	.039	.040	.039
1–4	.027	.051	.039	.045	.048
5–9	.030	.057	.042	.044	.054
10–14	.024	.046	.036	.046	.048
15–19	.015	.028	.030	.037	.045
20–24	.013	.023	.029	.035	.044
25–29	.018	.032	.034	.039	.048
30–34	.027	.052	.043	.048	.058
35–39	.031	.059	.050	.058	.068
40–44	.027	.054	.050	.058	.069
45–49	.022	.044	.048	.057	.069
50–54	.016	.029	.042	.053	.068
55–59	.013	.025	.041	.052	.067
60–64	.013	.025	.039	.049	.063
65–69	.015	.028	.039	.048	.060
70–74	.017	.031	.038	.046	.056
75–79	.018	.033	.036	.042	.051
80–84	.019	.035	.034	.040	.048
85–89	.018	.034	.033	.038	.045
90–94	.016	.032	.031	.036	.043

Netherlands in 1960–1989. The time series show abrupt changes from one year to the next, but there does not seem to be systematic trends. There is some evidence of auto-correlatedness, and the author identifies an MA(1) model. After the second forecast year, this model uses the mean of the data period as the forecast, i.e., it is essentially equivalent to an uncorrelated series. Bäckman and Scheele (1995) have given estimates of the relative error of migration forecasts by age, and shown that in ages with high mobility the errors have been the greatest in the Stockholm region.

Finally, we note Keilman's (1990, Figure 5.1., p. 83) results on the cross-correlations of the Dutch fertility and mortality forecasts. Forecasts for both have been too high since the 1960s. The same is true in many other industrialized countries, such as the U.S., Canada, and the Nordic countries. However, during the 1940–1960 period, fertility rose rapidly, but mortality declined. Since the baby-boom came everywhere as a surprise, it seems clear that the errors of fertility and mortality forecasts have little to do with each other. This suggests that a zero cross-correlation between them is an appropriate approximation. To avoid any misunderstanding, it is important to note that we do not claim that mortality and fertility could not be behaviorally ''correlated.'' We merely point out that it is unusual that the *forecast errors* of mortality and fertility were correlated.

A further technical issue is the nature of error in long-term forecasting. We noted in Alho (1990) that prediction intervals of the total fertility rate rapidly exceed values that are considered plausible in the light of historical experience. This is particularly apparent if an ARIMA forecast of the logarithm of the total fertility rate is made. We reported an analogous finding in the case of the U.S. growth rate, in Section 2.2.2. If the prediction intervals are not compatible with our other information concerning the vital processes, then some modification of the error specification is called for. Alho (1990) and Lee (1992) have experimented with logistic transformations that constrain the rates into a plausible range. The mixed estimation approach discussed in detail in Alho (1992a) and the mean conditioning of Lee (1993) similarly contribute to the stability of the prediction intervals. However, below we will present a new method that is particularly simple to use and avoids the use of deterministic bounds, while still retaining a strong control over the forecast in the long run.

3.2. *Simple covariance models for prediction errors*

Based on the empirical analyses summarized in the previous section, we will define simple correlation structures that can be used as building blocks to approximate many demographic forecast errors. Even if we use formal statistical methods in forecasting, we could opt for the family to be introduced to provide a simple approximation for the often complex analytical formulas. A key feature of the formulation is that it is given in terms of random variables, rather than moments. Hence, it is directly applicable in simulation.

We will assume that the prediction intervals up to time $T \leq \infty$ may be determined by empirical data, such as a formal statistical model, volatility or *ex post* based estimates etc. For very long term forecasting ($t \geq T$) we may specify a subjective structure that continues smoothly from the earlier part, but remains bounded *ad infinitum*. The choice of T will depend on the series. If the forecast errors increase to levels that are considered

implausible by expert demographers, then we may want to switch to a subjective specification that incorporates such judgment.

Consider error processes $X(j, t)$, where $j = 1, \ldots, J$ may refer to age or region, for example, and $t > 0$ is the forecast year. Suppose that the processes are of the form $X(j, t) = \epsilon(j, 1) + \ldots + \epsilon(j, t)$, where the error increments are of the form

$$\epsilon(j, t) = S(j, t)(\eta_j + \delta(j, t))$$

Here, the $S(j, t)$ are known weights whose specification will be discussed shortly. Assume that for each j, the variables $\delta(j, t)$ are independent over time $t = 1, 2, \ldots$ In addition, we let the variables $\{\delta(j, t) | j = 1, \ldots, J; t = 1, 2, \ldots\}$ be independent of the variables $\{\eta_j | j = 1, \ldots J\}$. Furthermore, we assume that

$$\eta_j \sim N(0, \kappa_j), \delta(j, t) \sim N(0, 1 - \kappa_j)$$

where $0 < \kappa_j < 1$ are known. For most purposes we may assume that $Corr(\eta_i, \eta_j) = \rho_\eta^{|i-j|}$, or $Corr(\eta_i, \eta_j) = \rho_\eta$, for some $|\rho_\eta| \leq 1$. Similarly, $Corr(\delta(i, t), \delta(j, t)) = \rho_\delta^{|i-j|}$, or $Corr(\delta(i, t), \delta(j, t)) = \rho_\delta$ for some $|\rho_\delta| \leq 1$. Since the increments are scaled by the $S(j, t)$, we call this a *scaled model* for error.

Note first, that $\kappa_j = Corr(\epsilon(j, t), \epsilon(j, t + h))$ for all $h \neq 0$. Therefore, κ_j can be interpreted as a constant correlation between the error increments. Under a Brownian motion model mentioned in Section 2.2.1. the error increments would be uncorrelated with $\kappa_j = 0$. This may be appropriate in fertility forecasting, for example. Second, note that $Var(\epsilon(j, t)) = S(j, t)^2$. Suppose that we have an increasing sequence of error variances $\sigma(j, 1)^2 < \sigma(j, 2)^2 < \ldots < \sigma(j, T)^2$ available with $Var(X(j, t)^2 = \sigma(j, t)^2$. Such variances may be known by subjective specification, formal methods, or by *ex post* or volatility based estimates. We can now estimate the corresponding variances of the error increments by taking first $S(j, 1)^2 = \sigma(j, 1)^2$. One can show that at $t > 1$, we must have

$$S(j, t) = -\kappa_j s(j; 1, t - 1) + [\kappa_j^2 s(j; 1, t - 1)^2 + \sigma(j, t)^2 - \sigma(j, t - 1)^2]^{1/2}$$

where $s(j; 1, t - 1) = S(j, 1) + \ldots + S(j, t - 1)$. Note that in the case $\kappa_j = 0$, this simplifies to $S(j, t)^2 = \sigma(j, t)^2 - \sigma(j, t - 1)^2$. (This was the method we used in Section 2.2.6. to specify the probability structure.) A moment's reflection shows that when $\kappa_j = 1$, then $s(j; 1, t - 1) = \sigma(j, t - 1)$, and $S(j, t) = \sigma(j, t) - \sigma(j, t - 1)$.

The structure given above provides flexible approximations to many different error structures that are relevant in demographic forecasting. For example, consider the logarithm of age-specific mortality in age j at time t, denote it by $m(j, t)$. The Lee and Carter (1992) forecast of $m(j, t)$ leads to a model $m(j, t) = \hat{m}(j, t) + X(j, t)$, where the prediction is of the form $\hat{m}(j, t) = a_j + b_j \hat{\alpha} t$, where the a_j and b_j form the mean and the first principal component vectors (and are treated by Lee and Carter as fixed in the propagation of error) and $\hat{\alpha}$ is an estimated parameter with a standard error. The increments of the error process are of the form $\epsilon(j, t) = b_j(\epsilon_\alpha + \delta_t)$, where $\epsilon_\alpha \sim N(0, \sigma_\alpha^2)$ and $\delta_t \sim N(0, \sigma_\delta^2)$. We see that by defining $c = (\sigma_\alpha^2 + \sigma_\delta^2)^{1/2}$ and taking $S(j, t) = cb_j, \eta_j = \epsilon_\alpha / c, \delta(j, t) = \delta_t / c$, and $\rho_\eta = \rho_\delta = 1$, we have a special case of the proposed model with $\kappa_j = \sigma_\alpha^2 / c^2$ for all j.

We have argued that the usual time-series methods often produce prediction intervals that will eventually be too wide. This may happen if the methods do not incorporate

sufficient information about the boundedness of the vital processes. We propose to take such additional information into account by allowing for modifications in the error structure so that levels of error that contradict the additional information are excluded. The first proposal is a simple one. Suppose we judge that the error structure we have specified yields what should be a maximum variance by year T. We may then assume that from T on the error structure will follow an AR(1) process centered around the point forecast that has the standard deviation $Var(X(j, T))^{1/2}$ and the first autocorrelation $Corr(X(j, T-1), X(j, T))$. We will consider $X(T)$ as the first value of the AR(1) process, so there is a smooth transition from one process to the next.

To provide a theoretical basis for the eventual AR(1) assumption, it is useful to note that the AR(1) process is the discrete time version of the Ornstein-Uhlenbeck process of diffusion theory. There, the process is obtained from a Brownian motion as subjected to an elastic force towards a mean function (Feller 1971, pp. 99, 335–336). This notion seems to capture well the idea that the errors should be centered around the point forecast and have a bounded variance, in the long run.

However, the continuous time Ornstein-Uhlenbeck formulation also suggests a modification of the above construction that is slightly less general, but simpler to describe. Let us suppose that we have a sequence of variances $\sigma(j, 1)^2 < \ldots < \sigma(j, T)^2 < \sigma(j, \infty)^2$, where $\sigma(j, \infty)^2$ is approached asymptotically, and suppose we are given the autocorrelation $e^{-\lambda} = Corr(X(j, T-1), X(j, T))$ for some $\lambda > 0$. We can then define an Ornstein-Uhlenbeck process that is zero at time $\tau = 0$, and has the Gaussian distribution $N(0, (1 - e^{-2\lambda\tau})\sigma(j, \infty)^2)$ at time $\tau > 0$. By taking $\tau_t = -\log(1 - \sigma(j, t)^2/\sigma(j, \infty)^2)/(2\lambda)$ for $t = 1, \ldots, T$, we see that the Ornstein-Uhlenbeck process has the variances $\sigma(j, t)^2$ at times τ_t. Hence, by altering the time scale of the Ornstein-Uhlenbeck process we get sample paths for the error process. This formulation permits the same cross-correlation structures as above, but it is constrained to a specific autocovariance structure dictated by the Ornstein-Uhlenbeck assumption.

The reports documenting the production of the *ex ante* forecast intervals should clearly indicate what kinds of additional information were used to constrain the level of error, and what the effects of these constraints are, e.g., at what point in time do the constraints take effect.

3.3. *A model for the error of the total fertility rate*

As the first application of the model given above, let us consider the modeling of the error in a forecast of age-specific fertility. Let $j = 1, \ldots, J$ refer to single years of age (such as 15–44) and define $\hat{f}(j, t)$ as the forecast of age-specific fertility in year t. Assume that the true rate is of the form $f(j, t) = \hat{f}(j, t)\exp(X(j, t))$, where the age-specific error of the forecast is the same for each age j, or $X(j, t) \equiv X(t)$ with $E[X(t)] = 0$. This simplified specification assumes that the error can be approximated in terms of the error of the total fertility rate alone, and that the errors are perfectly correlated over age (i.e., $\rho_\eta = \rho_\delta = 1$). This matches the assumptions of Lee and Tuljapurkar (1994) in their fertility forecast. A more refined analysis would take into account errors in the forecasted distribution of fertility (cf., Cruijsen and Zakee 1991, p. 36), but one would not expect this to have a major effect on the analysis of the error in births.

Suppose the error process $X(t)$ is a Brownian motion with the annual increment $\epsilon(j, t) \equiv \epsilon(t) \sim N(0, .08^2)$, where the variance is based on volatility estimates presented in Alho (1990) for Finland and the U.S. In this case we would take $\kappa_j = 0$, and $S(j, t) \equiv S(t) = .08$. Suppose the forecaster deems it unacceptable that, say, the 95 per cent forecast interval exceeds twice the point forecast, or goes below one half of the point forecast. Then, we would switch to a judgmental specification at time $t = T$ such that $\exp(2(.08)T^{1/2}) = 2$, or at $T = [\log(2)/(.16)]^2$. Rounding to integers we have $T = 19$. Under the Brownian motion process the first autocorrelation between values at time t and $t - 1$ is $[(t - 1)/t]^{1/2}$, so at $t = 19$ we get .973. Similarly the standard deviation of the value at $t = 19$ is $(.08)(19)^{1/2} = .349$. Therefore, we would replace the initial Brownian motion model with an AR(1) model that has the standard deviation .349 and first autocorrelation .973, for $t \geq 19$. There is some interest to note that the innovation variance of the AR(1) process is $(1 - .973^2)(.349)^2 \approx .081^2$, within rounding error of the variance of the error increment of the Brownian motion, $.08^2$.

For comparison, we note that the Ornstein-Uhlenbeck version of the error process that has variance $(.349)^2$ and $\tau = \log(.973)$ should be observed at times $\tau_1 = .99$, $\tau_2 = 2.02$, $\tau_5 = 5.6$, $\tau_{10} = 13.6$, $\tau_{15} = 28.4$, and $\tau_{19} = 117.0$, for example, to match the empirical variances. For $t = 20, 21, \ldots$ we would take $\tau_t = t - 19 + \tau_{19}$.

Define $B(t)$ as the number of births during year $t > 0$, and let $\hat{B}(t)$ be its forecast. Assume, for the moment, that we can forecast mortality and migration without error. Define $\xi(t)$ such that $B(t) = \hat{B}(t)\exp(\xi(t))$ for $t > 0$, and $\xi(t) = 0$ for $t \leq 0$. Therefore, $\xi(t)$ is the contribution of future fertility alone to the uncertainty of the birth forecast. Similarly, define the number of women in age x during year t as $P(x, t)$. We have

$$B(t) = \exp(X(t)) \sum_x \hat{P}(x, t)\hat{f}(x, t)\exp(\xi(t - x))$$

Define $\hat{b}(x, t) = \hat{P}(x, t)\hat{f}(x, t)/\Sigma_y \hat{P}(y, t)\hat{f}(y, t)$ as the forecast of the fraction of births that are due to women in age x during time t. Then, a Taylor series expansion for $\log(B(t))$ yields the following *approximate renewal equation for the error due to fertility*

$$\xi(t) = X(t) + \sum_x \hat{b}(x, t)\xi(t - x)$$

Here, the summation extends over such child-bearing ages x that have $x - t < 0$. The renewal equation can be viewed as a version of Equation (2.5) of Lee (1974, p. 609), except in our case the mean from which the deviations are measured is an arbitrary function given by the point forecast. Also, for our purposes we may allow the weights $b(x, t)$ to depend on t.

The above renewal equation is useful, because it shows how errors in fertility propagate according to weights $\hat{b}(x, t)$ that are given by the point forecast. A practical way to see the effect is to replace $X(t)$ by an indicator process that takes the value zero everywhere, except at $t = t_0$, where it takes the value one. Running this process through the filter shows how an error at $t = t_0$ produces new errors at later times.

The model also permits an analytical calculation of the autocovariance structure of the $\xi(t)$ based on that of the $X(t)$. However, since the simplest way to estimate the autocovariances is via simulation, it is just as easy to simulate the values $B(t)$ directly,

and to derive the autocovariances of $\xi(t)$ from them. In either case, we may effectively derive the matrix $Cov(\xi_t)$, where $\xi_t = (\xi(1), \ldots, \xi(t))^T$, for any $t > 0$.

Above, we assumed that fertility was the only source of error in a birth forecast. The result of the calculation was a full covariance structure for the vector of the logarithm of the future births for as long as the forecast data are needed. It is easy to see that if the errors of fertility forecasts can be viewed as being independent of the errors in mortality and migration forecasts, then we can *add* those effects separately to the current results for a first-order approximation.

Another matter is that, as we have noted earlier (Alho 1992b), the level of uncertainty in fertility is so high in most industrialized countries that the added uncertainty due to mortality is negligible for the birth cohorts born after the jump-off. The same is true for migration in many national populations. (However, in small subnational populations migration can be a major, even the largest, source of error.) A remarkable consequence is that whenever we may ignore the other sources of error, the error structure for all survivors of the birth cohorts born after the jump-off time, can be derived from $Cov(\xi_t)$ for t large enough. To see this, let $V(x,t) = \exp(\hat{v}(x,t) + \epsilon_v(x,t))$ be the number of males (or females) in age x at time $t > 0$. If fertility is the only source of error, then (in a single region model) the relative prediction error is $\epsilon_v(x,t) = \xi(t-x)$ for $x < t$.

3.4. *A model for the error of a mortality forecast*

In this section we will take judgmental high, medium, and low forecasts of age-specific mortality, and show how these can directly be turned into a stochastic forecast. This allows the forecaster to perform additional propagation of error calculations based on the forecast in a way that is compatible with the original forecast but that avoids the severe perfect autocorrelation assumptions that the deterministic calculations are equivalent to.

Consider the U.S. female mortality in ages 0, 1–4, 5–9, 10–14, ..., 90–94. The jump-off year is 1985 and the official high, medium, and low forecasts are given for the years 1990, 2000, ..., 2080, so the forecast periods are 5, 15, 25, ..., and 95 years (Wade 1987, pp. 11–12). We took the logarithms of the rates and used the average of the "high–medium" and "medium–low" differences as an estimate of the standard deviation of the forecast error in the log-scale. In other words, we interpreted the high–low interval as (approximately) a 67 per cent prediction interval for a normally distributed error with mean zero. [Should it be more appropriate to interpret the official intervals as, say, 95 per cent intervals, then the estimated standard deviations we present should be divided by two.]

We considered both independent and perfectly correlated error increments ($\kappa_j = 0$ or $\kappa_j = 1$). In both cases we assumed that the standard deviation was the same during the first five forecast years, during the next ten forecast years, and similarly for each of the following ten year segments. This gave us estimates of the standard deviations for each of the ten forecast periods, for each age. Table 2 gives results for five forecast periods for both models. The results for the intermediate periods can be accurately obtained by interpolation. The perfect correlation interpretation results in estimates that are much higher during the first fifteen forecast years than later. The independence model shows a more even

progression. Curiously, both models show that in ages 30–44 the relative forecast errors are thought to be higher than in the neighboring ages.

To compare these error estimates for mortality to those for fertility, suppose $X(j, t)$ is the error in the forecast of the logarithm of the mortality rate in age j at time t. Denote the logarithm by $m(j, t)$, its forecast by $\hat{m}(j, t)$, and let $p(j, t) = \exp(-\exp(\hat{m}(j, t) + X(j, t)))$ be the corresponding *survival* probability with the forecast $\hat{p}(j, t) = \exp(-\exp(\hat{m}(j, t)))$. Then, the delta method gives the approximation

$$Var(p(j,t)) \approx \hat{p}(j,t)^2 \exp(\hat{m}(j,t))^2 \sigma(j,t)^2$$

where $Var(X(j, t)) = \sigma(j,t)^2$. We see that the relative error of the survival rate has approximately the standard deviation $\exp(\hat{m}(j,t))\sigma(j,t)$, i.e., it is the mortality rate times $\sigma(j,t)$. Depending on the country the age-specific mortality rates are in ages 1–40 roughly 1/1000, in ages 40–65 roughly 1/100, and in ages 65–80 less than 1/10. We omit the details (see Exhibit 4 of Alho 1992b), but note that this is the reason why in industrialized countries uncertain mortality is not a major source of error for forecasts of the size of birth cohorts that are born after the jump-off, despite the fact that t annual errors are added to get the t-year survival rate.

To complete the judgmental specification of the *ex ante* error we need assumptions for the cross-correlations of the errors across age. As noted in Section 3.1., naive forecasts of age-specific mortality rates have had highly variable cross-correlations, but a constant correlation model can provide a reasonable approximation. It can be interpreted as saying that all the ages share a common cause of forecast error, but that there are also age-specific error factors that modify the error from the average error. While the constant correlation assumption could certainly be refined, one should note that it is more general than the perfect cross-correlation assumption of Lee and Carter (1992).

3.5. *A model for the error of age-specific migration*

A proper description of migration in a system of n regions would require data and analysis of $(n-1)^2$ outmigration rates per age and sex annually. However, migration data are poor in many countries, and even if the data are available, the demands on the analyst are heavy. For such reasons, migration is typically forecasted using summary procedures, such as specifying an annual additive vector for a country's international net-migration; or by specifying annual out-migration rates for sub-regions, and reallocating them according to historical shares, for internal migration.

Given the multitude of settings, we will only sketch one simple approach to the handling of the error in migration forecasts. We will reduce the analysis to the consideration of *census survival rates*. In other words, we assume that the migration forecast can be represented in terms of correction factors that modify survival rates, to account for migration. Most migration forecasts can be reformulated in this way, even if they had originally been made using other procedures.

Define the true correction factor in age j at time t as $Q(j,t) = \exp(\hat{q}(j,t) + \epsilon_q(j,t))$, where $\hat{q}(j,t)$ is the forecast of $\log(Q(j,t))$ and $\epsilon_q(j,t)$ is a zero-mean forecast error. At each survival step, we should add $\hat{q}(j,t) + \epsilon_q(j,t)$ to the logarithm of the survival rate to get the correct result, but we only add $\hat{q}(j,t)$. This produces the error we have to

characterize. A simple approach is the following. Suppose we write $\hat{q}(j,t) = \hat{r}(j,t) - \hat{u}(j,t)$, where $\hat{r}(j,t)$ is the correction that we would need, if there were in-migration only, and $\hat{u}(j,t)$ is defined by subtraction. Therefore, $\hat{u}(j,t)$ is approximately the factor we would need, if there were out-migration only, in age j during year t. (Essentially, $\hat{r}(j,t)$ and $\hat{u}(j,t)$ are the in-migration and out-migration "rates.") A natural model would seem to be that the true factors are of the form $r(j,t) = \hat{r}(j,t)\exp(\epsilon_r(j,t))$, and $u(j,t) = \hat{u}(j,t)\exp(\epsilon_u(j,t))$. Therefore, we have

$$\begin{aligned}\epsilon_q(j,t) &= \hat{r}(j,t)(\exp(\epsilon_r(j,t)) - 1) - \hat{u}(j,t)(\exp(\epsilon_u(j,t)) - 1) \\ &\approx \hat{r}(j,t)\epsilon_r(j,t) - \hat{u}(j,t)\epsilon_u(j,t)\end{aligned}$$

The family of Section 3.2. for $\epsilon(j,t)$ can now be used directly for the terms $\hat{r}(j,t)\epsilon_r(j,t)$ and $\hat{u}(j,t)\epsilon_u(j,t)$. This implies a model for $\epsilon_q(j,t)$.

A new consideration is the correlation between $\epsilon_r(j,t)$ and $\epsilon_u(j,t)$. If in-migration and out-migration tend to increase or decrease simultaneously, then the correlation is positive. This can happen if, say, economic cycles sometimes accelerate and sometimes decelerate the rate at which people change their jobs, or housing. If unexpectedly high in-migration is associated with unexpectedly low out-migration, and conversely, then we have a negative correlation. This could be due to unanticipated change in the relative attractiveness of a region. Changes in the way *de jure* populations are defined may produce any kind of correlation, depending on composition of inmigrants and outmigrants.

The key issue in the above derivation is that even if we forecast net-migration directly, we need some idea about the magnitude and age-distribution of in- and out- migration. Even, if we were willing to assume that the two age-distributions are the same, we still need to have an estimate of what that common age-distribution is. The model migration schedules of Rogers and Castro (1981) may be valuable in this task. To get a handle on the magnitudes of the error terms we can always resort to naive forecasts. Recent past average level, if known, can often be taken as the naive forecast.

4. Discussion and Recommendations

We have presented an overview of how one can assess, using relatively simple procedures, the kinds of forecasts that are currently made by official statistical agencies. To be sure, these procedures could be refined to take account of special circumstances in different countries. They may also require modification to match a country's specific forecasting procedures and data capability. Nevertheless, we believe that our procedures are accurate enough to give a realistic assessment of the error to be expected in forecasting, and we hope that our discussion can serve as a basis for country-specific applications.

Our first recommendation is that the national statistical agencies do an *ex post* assessment of the error in their forecasts of the total population. These estimates should find their way into the future forecast reports as a crude warning that the users should be alert about the possibility of error.

Second, due to the central role of fertility in the error of population forecasting, the agencies should do an *ex post* evaluation of error, and a volatility-based assessment of error, for the total fertility rate, and compare the two. Something is wrong, if the *ex post* errors are systematically bigger than the volatility based estimates.

Third, the agencies should specify a Brownian motion model for the total fertility rate to match their country's circumstances and get an estimate of error for the birth vector. This may involve exercising judgment (recall Section 3.2), in order to avoid too large errors. The relative error in the size of a birth cohort will give a fairly accurate estimate of the relative error in the number of survivors from the cohort.

Fourth, the agencies should formulate a probabilistic version of their mortality forecast either based on time-series methods or judgmentally. This should be applied to the population already born at jump-off. Together with the previous point, this would produce a complete probabilistic specification of error for all ages.

Fifth, the agencies should assess the uncertainty in migration forecasts and in the jump-off population. In some countries these uncertainties would not be major sources of error on the national level, but in others they would. The estimates should be added to those obtained earlier.

Sixth, we believe that real advances in the usefulness of forecasts are obtainable, if they are implemented as databases. The procedures we have outlined here make it possible to produce relatively simple computer programs that will calculate error estimates for user-specified population aggregates. This will tremendously increase the value of the forecasts for the users (Alho and Spencer 1991).

Seventh, we would like to caution the forecasters not to simply follow the usual statistical practice of using five per cent as the risk level. It is likely that many users do not realize what the actual level of uncertainty in population forecasts is. Therefore, if the 95 per cent prediction intervals are much wider than the current high–low intervals, then this may create some unnecessary speculation as to what has caused the "decline" in the accuracy of the forecasts. Try 50% or 67% prediction intervals instead!

5. References

Ahlburg, D.A. and Vaupel, J.W. (1990). Alternative Projections of the U.S. Population. Demography, 27, 639–652.

Alho, J.M. (1990). Stochastic Methods in Population Forecasting. International Journal of Forecasting, 6, 521–530.

Alho, J.M. (1992a). Estimating the Strength of Expert Judgment. Journal of Forecasting, 11, 157–167.

Alho, J.M. (1992b). The Magnitude of Error Due to Different Vital Processes in Population Forecasts. International Journal of Forecasting, 8, 301–314.

Alho, J.M. and Spencer, B.D. (1985). Uncertain Population Forecasting. Journal of the American Statistical Association, 80, 306–314.

Alho, J.M. and Spencer, B.D. (1990a). Error Models for Official Mortality Forecasts. Journal of the American Statistical Association, 85, 609–616.

Alho, J.M. and Spencer, B.D. (1990b). Effects of Targets and Aggregation on the Propagation of Error in Mortality Forecasts. Mathematical Population Studies, 2, 209–227.

Alho, J.M. and Spencer, B.D. (1991). A Population Forecast as a Database: Implementing the Stochastic Propagation of Error. Journal of Official Statistics, 7, 295–310.

Alho, J.M. and Spencer, B.D. (1997). Statistical Demography and Forecasting. Unpublished manuscript.

Andrews, G.H. and Beekman, J.A. (1987). Actuarial Projections for the Old-Age, Survivors, and Disability Insurance Program of Social Security in the United States of America. Itasca, IL: Actuarial Education and Research Fund.

Bäckman, J.D. and Scheele, S. (1995). Uncertainty in Population Forecasts for Small Areas. Scandinavian Population Studies (to appear).

Box, G.E.P. and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control. Revised ed., San Francisco, CA: Holden-Day.

Cheeseman Day, J. (1993). Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1993 to 2050. Current Population Reports, Series P-25, No. 1018. U.S. Bureau of the Census, Washington, DC.

Cochran, W.G. (1977). Sampling Theory. New York, NY: John Wiley.

Cohen, J.E. (1986). Population Forecasts and Confidence Intervals for Sweden: A Comparison of Model-Based and Empirical Approaches. Demography, 23, 105–126.

Cruijsen, H. and Zakee, R. (1991). Population Forecasts for the Netherlands During the 1990s: How Far Were They Wrong? Maandstatitiek van de Bevolking 39(7), July. Statistics Netherlands.

De Beer, J. (1993). Forecast Intervals of Net Migration: The Case of The Netherlands. Journal of Forecasting, 12, 585–599.

Draper, D. (1995). Assessment and Propagation of Model Uncertainty (with Discussion). Journal of the Royal Statistical Society, Series B, 57, 45–97.

Feller, W. (1971). An Introduction to Probability Theory and Its Applications, Vol. II. New York, NY: John Wiley.

Hoem, J.M. (1973). Levels of Error in Population Forecasts. Artikler 61, Statistics Norway.

Keilman, N. (1990). Uncertainty in National Population Forecasting: Issues, Backgrounds, Analysis, Recommendations. NIDI GBGS Publications. Amsterdam: Swets and Zeitlinger.

Keyfitz, N. (1981). Can Knowledge Improve Forecasts? Population and Development Review, 8, 719–751.

Lee, R.D. (1974). Forecasting Births in Post-Transitional Populations: Stochastic Renewal With Serially Correlated Fertility. Journal of the American Statistical Association, 69, 607–617.

Lee, R.D. (1992). Stochastic Demographic Forecasting. International Journal of Forecasting, 8, 315–327.

Lee, R.D. (1993). Modeling and Forecasting the Time Series of U.S. Fertility: Age Distribution, Range, and Ultimate Level. International Journal of Forecasting, 9, 187–202.

Lee, R.D. and Carter, L. (1992). Modeling and Forecasting the Time Series of U.S. Mortality. Journal of the American Statistical Association, 87, 659–671.

Lee, R.D. and Tuljapurkar, S.D. (1994). Stochastic Population Forecasts for the United States: Beyond High, Medium, and Low. Journal of the American Statistical Association, 89, 1175–1189.

Mulry, M.H. and Spencer, B.D. (1993). Accuracy of the 1990 Census and Undercount Adjustments. Journal of the American Statistical Association, 88, 1080–1091.

Panel on Small Area Estimates of Population and Income (1980). Estimating Population and Income of Small Areas. National Research Council, Washington, DC: National Academy Press.

Pflaumer, P. (1992). Forecasting U.S. Population Totals with the Box-Jenkins Approach. International Journal of Forecasting, 8, 329–338.

Rogers, A. and Castro, L.J. (1981). Model Migration Schedules. RR-81–30, Laxenburg: I.I.A.S.A.

Spencer, B.D. (1980). Models for Error in Postcensal Population Estimates. In Panel on Small Area Estimates of Population and Income, Estimating Population and Income of Small Areas. National Research Council, Washington, DC: National Academy Press.

Spencer, G. (1989). Projections of the Population of the United States by Age, Sex, and Race: 1988 to 2080. Current Population Reports, Series P-25, No. 1018. U.S. Bureau of the Census, Washington, DC.

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

Wade, A. (1987). Social Security Area Population Projections: 1987. Actuarial Study No. 99. Washington, DC: Office of the Actuary.

Whelpton, P.K., Eldridge, H.T., and Siegel, J.S. (1947). Forecasts of the Population of the United States. U.S. Bureau of the Census, Washington, DC: U.S. Government Printing Office.

Received June 1995
Revised February 1996