Journal of Official Statistics
Vol. 7. No. 3, 1991, pp. 295–310

A Population Forecast as a Database: Implementing the Stochastic Propagation of Error

Juha M. Alho[1] *and Bruce D. Spencer*[2]

Abstract: We propose implementing population forecasts as databases. This has two advantages. First, users can retrieve custom forecasts to suit their particular needs. Second, the forecasts can include probabilistic prediction intervals that allow the user to assess the uncertainty of the forecast. A major contribution of the paper is the derivation of propagation of error formulas that are needed to calculate the probabilistic intervals. A computer program implementing a simple version of a stochastic population forecast is described.

Key words: Databases; population projections; propagation of error; stochastic methods; uncertainty.

1. Introduction

Statistical agencies are caught between limited resources for publishing statistics and pressure from data users to make more statistics available. For example, a user may require more detailed cross tabulations than the ones available in published statistics, or he or she may need data for unusual aggregates. These demands are hard to meet because every user has potentially different needs. In addition, sophisticated users have become increasingly aware that all statistics are not equally accurate. They expect estimates of the uncertainty of the statistics. This issue of the *quality of statistics* arises always, whether the statistics be based on censuses, sample surveys, or administrative records. However, the need to express uncertainty is paramount for forecasts because they contain the largest errors. Meeting the need to express uncertainty is also difficult in published statistics, because for every n estimates there are n variances and, more generally, $n(n-1)/2$ covariances of potential interest.

One emerging solution to these dilemmas is to make available a computerized database that users can tap and create custom statistics. The database can reside in a central computer (or network) that provides on-line access to authorized users, or copies of the database can be provided on diskette (or machine readable form) for microcomputer use. This paper focuses on how users can retrieve estimates of population size and estimates of uncertainty for population aggregates of their choice in forecasts produced by the so-called cohort-component method. A central contribution is the deri-

[1] Institute for Environmental Studies and Department of Statistics, University of Illinois at Urbana-Champaign, 1101 W. Peabody Drive, Urbana, IL 61801, U.S.A.
[2] Department of Statistics, Northwestern University and Methodology Research Center, NORC, 2006 Sheridan Road, Evanston, IL 60208, U.S.A.

Acknowledgment: The authors would like to thank Ms. Wen-Ling Peng for programming help. The support of this research by a DHHS Grant 89ASPE220A and NIA Grant AG06696-01 is gratefully acknowledged.

vation of the propagation of error formulas that are used to carry out the uncertainty analysis. Our approach (although not our particular models) applies to other statistics as well.

The *cohort-component* method of population forecasting typically consists of projecting future numbers of annual births, deaths, and migration by one or five-year age-sex groups, adding them to form a new population vector, and repeating the calculations for each forecast year (Shryock and Siegel 1976, pp. 443–444). This procedure was introduced by P.K. Whelpton in a sequence of papers beginning in 1928 (Whelpton 1928; Thompson and Whelpton 1933, ch. X; Whelpton 1936; Whelpton, Eldridge, and Siegel 1947). Essentially the same procedure is currently used in many if not most official population forecasts, including the *functional forecasts* used for social security purposes, such as the forecasts of the number disabled (e.g., Wade 1987; Andrews and Beekman 1987).

Ever since Whelpton's time the problem of expressing uncertainty about the likely future population growth has been solved by preparing three or more forecast variants that are hoped to cover the "reasonable" alternative future paths of development. Typically, no probability statement is given for the event that the high-low interval covers the true population size. For example, the U.S. Bureau of the Census (1984, p. 1) considers three principal alternative assumptions each about fertility, mortality, and migration. The low (high) forecasts are obtained by assuming that in each year and for each population subgroup fertility, life expectancy, and immigration are low (high). Such a practice contrasts to the usual statistical or stochastic approach to error, which recognizes some degree of independence among alternative occurrences. Fertility or migration or life expectancy might be unexpectedly low for some subgroups for some years but the chances are considerably smaller that all three components of change would be simultaneously so low in all subgroups for all years. A consequence of this deterministic or nonstatistical approach is that the probability content of the intervals (that is, the probability that the future population sizes fall within high-low intervals) is inconsistent for forecasts for different points in time or for different subgroups. For example, Alho and Spencer (1985, p. 313) found that the probability content of the U.S. Bureau of the Census' intervals for age groups surviving from the jump-off year 1980 were somewhat less than 0.90 initially but grew with time until they were about 0.90 after about 5 years and greater than that thereafter. Intervals for births, on the other hand, were much narrower than 0.67 level prediction intervals for all 15 forecast years considered. Combining the birth forecast with the forecast of survivorship produces high-low intervals with a highly erratic probabilistic content.

The first attempt at giving a probabilistic interpretation, which we are aware of, was due to L. Törnqvist in Finland in 1949 (Hyppölä, Tunkelo, and Törnqvist 1949, pp. 68–74). This early work contained many of the central features of a stochastic forecast. Törnqvist's lead was not followed up in official forecasts until the 1980s when a probabilistic interpretation was considered by the U.S. Bureau of the Census (1984). Although it is more complicated to take a stochastic rather than a deterministic approach to the assessment of error, the stochastic approach is the only way to achieve prediction intervals with a known, constant probability content.

Since Törnqvist there has been an extensive discussion of how to use time-series methods to produce interval estimates of

particular functions of the population vector, such as total population size, or the number of births or deaths. For a review of this literature until 1985, see Land (1986). More recent contributions include those of Rogers (1986), Bozik and Bell (1987), Bell (1988), Lee and Carter (1990), McNown and Rogers (1990), Alho (1991) and Alho and Spencer (1990a, 1990b).

Apart from the work of Törnqvist, analyses of the uncertainty of the whole population vector started out by considering branching process models, such as multitype Galton-Watson processes, which treat the vital rates as probabilities (e.g., Kendall 1949; Goodman 1968; Feichtinger 1970; Pollard 1973, ch. 9; Keiding and Hoem 1976). However, it is well known that the amount of uncertainty generated by this type of binomial, or Poisson, variability is much less than the level actually observed. Pollard (1968) and Sykes (1969) appear to have been the first to consider population processes with random vital rates. Later, Cohen (1977) considered a model in which the matrices of vital rates were allowed to follow a (possibly non-time homogeneous) Markov chain with a finite state space. However, none of these contributions considered the problem of statistical prediction of the vital rates as a part of the propagation of error. In Alho and Spencer (1985) the probabilistic analysis of the whole population vector including the statistical prediction of the vital rates was taken up. Although this work showed that it is feasible to derive analytical approximations to the distribution of future population vectors, many difficult problems remained. An alternative to the analytic approach is to use simulation, see, e.g., Pflaumer (1988).

A problem in the analytic approach involves the handling of births to future births (when the forecast period exceeds the lower limit of the child bearing ages). This will greatly complicate the propagation of error calculations. One solution to the problem has been given by Davis (1988), who sketched recursive formulas for the covariance matrix of the future population vector under alternative simplifying assumptions. The most realistic set-up assumes only that the prediction errors of the vital rates have *state-space representations* (Davis 1988, p. 24; cf., Akaike 1974). If forecasts of the future vital rates are available in terms of ARIMA models, then Davis's approximate Kalman-filter calculations may provide an attractive way of handling the analytic propagation of error.

Our approach for the propagation of error is based on the consideration of the logarithm of the population vector. This linearizes the process of population growth for all other ages except the births. For the latter, a linear Taylor-series expansion will be used. We assume that the covariance structure of the forecast errors of the vital rates is known, and provide approximate formulas for the propagation of error based on that. The approach we propose involves the specification of the covariance matrix of the vector of births for all forecast years. This matrix is calculated once and stored. Thereafter, all the propagation of error calculations follow a uniform pattern. The user specifies the age-group and forecast years of interest. A computer program utilizes a collection of data matrices to compute the point forecasts and their standard deviations on a logarithmic scale, and puts them into a text file. The file can be read into any statistical package (such as Minitab) or a suitable spreadsheet program, for the specification of prediction intervals of desired level of coverage, and for graphical output. This is what we mean by "implementing the forecast as a database." In short, we view the forecast as a collection of files containing the requisite data and distributional

specifications, coupled with computer programs that are capable of producing the user-specified prediction intervals.

In Section 2 we describe a version of the basic linear growth model ("Leslie model") used to describe the evolution of the population vector over time. The formulas presented provide a basis for the propagation of error calculations of Section 3. In Section 4 we define and discuss the special problems of functional forecasts. Section 5 has a description of prototype computer programs that implement the formulas developed. Section 6 gives empirical examples of how probabilistic intervals differ from the high-low intervals currently used. We conclude in Section 7 by pointing out directions for future research.

2. Linear Growth Models

Define $\mathbf{V}(t) = (V(0, t), \ldots, V(s, t))^T$, where

$V(j, t)$ = size of *female* population in age j at time t.

To be specific, we let t refer to a specific date during the year t, e.g., January 1, or July 1. Age j refers to an age-group of females who have had their jth birthday by the specific date, but who have not had their $(j + 1)$st birthday. When $j = s$, we take $V(s, t)$ to be the size of the female population in the last age s.

Define an $(s + 1) \times (s + 1)$ matrix $\mathbf{R}(t) = (R(i, j, t))$, $i, j = 0, \ldots, s$, with $R(0, 15, t), \ldots, R(0, 44, t)$ the *age-specific fertility rates* of year t, and $R(1, 0, t), \ldots, R(s, s - 1, t)$ the *age-specific survival probabilities* from age 0 to age 1, from age 1 to age 2, etc., during year t. In addition, define $R(s, s, t)$ as the average survival probability in the last age-group. All other elements in matrices $\mathbf{R}(t)$ are zero. For some purposes we may want s to be a fairly small number, such as $s = 85$. For other purposes we may want s to be much larger, such as $s = 110$. In the latter case we may assume that $R(s, s, t) = 0$, so there is no survival in the last age.

The *linear growth model* specifies that

$$\mathbf{V}(t + 1) = \mathbf{R}(t)\mathbf{V}(t). \tag{2.1}$$

This model is capable of describing a closed female population (i.e., there is no migration). If we replace the life table survival rates by the so-called *census survival rates*, then migration can also be incorporated. In order not to complicate the terminology below, we shall assume that we are dealing with a closed population.

We will assume that the *jump-off population* $\mathbf{V}(0)$ is strictly positive, or $V(j, 0) > 0$ for $j = 0, \ldots, s$, and that the age-specific fertility rates and survival rates are strictly positive. Define,

$$\begin{aligned} v(j, t) &= \log \{V(j, t)\}, \\ f(j, t) &= \log \{R(0, j, t)\}, \\ r(j, t) &= \log \{R(j + 1, j, t)\}, \\ r(s, t) &= \log \{R(s, s, t)\}. \end{aligned}$$

The linear growth model (2.1) implies that

$$\mathbf{V}(t) = \mathbf{R}(t - 1) \cdots \mathbf{R}(0)\mathbf{V}(0). \tag{2.2}$$

First, consider the survivors of the jump-off population. We get the following formulas. For $1 \leqslant t \leqslant j \leqslant s - 1$ we have

$$v(j, t) = v(j - t, 0) + \sum_{k=0}^{t-1} r(j - t + k, k) \tag{2.3}$$

and for $t \leqslant s$ we have

$$v(s, t) = \log \left\{ \sum_{k=0}^{t} \exp \left[v(s - k, 0) + \sum_{n=0}^{k-1} r(s - k + n, n) + \sum_{h=k}^{t-1} r(s, h) \right] \right\}. \tag{2.4}$$

Consider now the births in (2.2). Note that it takes fifteen years for those born to be in the age 15, and one more year to have children. Therefore, the first forecast year for which there are "births to forecast births" is $t = 17$. Before that, we have for $t = 1, \ldots, 16$,

$$v(0, t) = \log \left\{ \sum_{k=15}^{44} \exp \left[v(k - t + 1, 0) + \sum_{n=0}^{t-2} r(k - t + 1 + n, n) + f(k, t - 1) \right] \right\}. \quad (2.5)$$

Strictly speaking, the formula for births given here does not give the correct size of the "age 0" population because it does not involve a factor for women's survival until birth and children's survival from birth to age 0, during the year of birth. This is easily remedied if we multiply the fertility rates by the average survival probabilities.

Surviving the first sixteen birth cohorts we have

$$v(j, t) = v(0, t - j) + \sum_{n=0}^{j-1} r(n, t - j + n) \quad (2.6)$$

for $\max \{0, t - 16\} \leqslant j < t$.

For the "second generation of births," or for the years $t = 17, \ldots, 32$, (2.2) implies that

$$v(0, t) = \log \left\{ \sum_{k=t-1}^{44} \exp \left[v(k - t + 1, 0) + \sum_{n=0}^{t-2} r(k - t + 1 + n, n) + f(k, t - 1) \right] + \sum_{k=15}^{t-2} \exp \left[v(0, t - 1 - k) + \sum_{n=0}^{k-1} r(n, t - 1 - k + n) + f(k, t - 1) \right] \right\}. \quad (2.7)$$

Here the first sum over k corresponds to births due to the survivors of the jump-off population. The latter sum over k corresponds to "births to births," so the equation is actually a version of the well-known *renewal equation* of the births (cf., Keyfitz 1977, p. 99). Note that the terms $v(0, t - 1 - k)$ have been defined earlier in terms of the jump-off population and past survival rates. The covariance between those survival rates and the ones appearing explicitly in the formula above is one source of the analytical difficulties mentioned in Section 1. However, we shall see below that it is possible to carry out approximate propagation of error calculations based essentially on the above "renewal equation" formulation, by approximating the (small) covariance in question by zero.

Surviving the birth cohorts born during the years $t = 17, \ldots, 32$, we apply (2.6) for $\max \{0, t - 32\} \leqslant j \leqslant t - 16$.

The formulas (2.6) and (2.7) apply to later birth cohorts, as well, so long as we note that for $t > 45$ the sum over k from $k = t - 1$ to 44 vanishes and if we take $f(k, t - 1) = -\infty$ for $k > 44$.

3. Propagation of Error

We will now introduce stochastic elements into our model. The notation given above will be reserved for the point forecast of future population. The population itself is taken to be a random vector $\tilde{\mathbf{V}}(t)$. Other random variables will also be distinguished by a $\sim$. Define $\mathbf{v}(t) = (v(0, t), \ldots, v(s, t))^T$, $\mathbf{f}(t) = (f(15, t), \ldots, f(44, t))^T$, and $\mathbf{r}(t) = (r(0, t), \ldots, r(s, t))^T$. Assume for the jump-off population that

$$E[\tilde{\mathbf{v}}(0)] = \mathbf{v}(0), \quad \text{Cov}(\tilde{\mathbf{v}}(0)) = \boldsymbol{\Sigma}(0)$$

where $\mathbf{v}(0)$ is a known vector, which has been estimated based on past data, and $\Sigma(0)$ is a known covariance matrix, which is based on what is known about the estimation error in $\mathbf{v}(0)$ (cf., Alho and Spencer 1985, p. 310). Assume further that there are predictions $\mathbf{f}(t)$ of $\tilde{\mathbf{f}}(t)$ such that

$$E[\tilde{\mathbf{f}}(t)] = \mathbf{f}(t), \text{Cov}(\tilde{\mathbf{f}}(t), \tilde{\mathbf{f}}(u)) = \Gamma(t, u),$$

where $\mathbf{f}(t)$s are known vectors and $\Gamma(t, u)$s are known covariance matrices for $t, u = 0, 1, \ldots$ with elements $\gamma(i, j, t, u)$. An example of the structure of the prediction error of vital rates (albeit in the presence of modeling bias) is given in Alho and Spencer (1985, 3.17, p. 309). Finally, assume that there are predictions $\mathbf{r}(t)$ of $\tilde{\mathbf{r}}(t)$ such that

$$E[\tilde{\mathbf{r}}(t)] = \mathbf{r}(t), \text{Cov}(\tilde{\mathbf{r}}(t), \tilde{\mathbf{r}}(u)) = \Theta(t, u),$$

where $\mathbf{r}(t)$s are known vectors and $\Theta(t, u)$s are known matrices. Their (i, j) elements are denoted as $\theta(i, j, t, u)$. Factors influencing the specification of the matrices $\Theta(t, u)$ are discussed in Alho and Spencer (1990b, sec. 4).

To complete the probabilistic specification of the model we assume that the variables $\tilde{\mathbf{v}}(0)$, $\tilde{\mathbf{f}}(t)$, and $\tilde{\mathbf{r}}(t)$ are jointly normal with the *jump-off population independent of the vital rates, and survival rates independent of the fertility rates*. Note that the assumption of normality is important for statistical inference, but it is irrelevant for the derivation of the formulas below, which involves the first two moments only. The assumption of independence simplifies our formulas considerably. It is a plausible assumption in practice, because the information relevant for the forecasting of mortality is largely independent of the information that is relevant for the forecasting of fertility, or of the errors in the estimation of the jump-off population. Moreover, even though an undercount of the denominator population will lead to an overestimate of a vital rate (and, therefore, to a negative correlation between the jump-off value and the vital rate), the variation in the number of events (births, deaths) typically is much larger than the uncertainty concerning the denominator, so the correlation is quantitatively insignificant.

We note that there may be circumstances in which the forecast errors of the vital processes may become crosscorrelated. AIDS, for example, may have an unexpected effect on both fertility and mortality. If such effects become numerically important, more complex formulas for the propagation of error may be called for.

To calculate the point forecast of future population we use directly the formulas of Section 2, or the equivalent recursive multiplicative formulas. The former will be helpful in the derivation of the approximate formulas for the propagation of error below. We define

$$E[\tilde{\mathbf{v}}(t)] = \mathbf{v}(t), \text{Cov}(\tilde{\mathbf{v}}(t), \tilde{\mathbf{v}}(u)) = \Sigma(t, u).$$

The elements of $\Sigma(t, u)$ are denoted by $\sigma(i, j, t, u)$. We approximate the distribution of $\tilde{\mathbf{v}}(t)$ by a normal distribution for all t. Based on the assumed log-normality of the jump-off population and the survival rates, the approximation is exact for the survivors of the jump-off population. For others an analytical approximation is involved. Given this set-up, we approximate the matrices $\Sigma(t, u)$ for $t, u = 1, 2, \ldots$ We will write $\Sigma(t, t) = \Sigma(t) = (\sigma(i, j, t))$ for short, i.e., $\sigma(i, j, t) = \text{Cov}(v(i, t), v(j, t))$.

In a number of instances we resort to Taylor-series approximations. Consider the function

$$g(x_1, \ldots, x_n) = \log\left(\sum_{i=1}^{n} \exp(x_i)\right).$$

A linear Taylor-series representation for g

at $(y_1, \ldots, y_n)$ is

$$g_L(x_1, \ldots, x_n) = g(y_1, \ldots, y_n) + \exp(-g(y_1, \ldots, y_n)) \times \sum_{i=1}^{n} \exp(y_i)(x_i - y_i)$$

and we will write $g \approx g_L$. We will use the second moments of g_L to approximate those of g.

3.1. *Survivors of the jump-off population*

We will assume that s is large, so that the death rate in age s can be taken to be infinite. Then we have for the covariance between the prediction error of population size in ages i and j at time t, when $1 \leqslant t \leqslant i \leqslant j \leqslant s$, that

$$\sigma(i, j, t) = \sigma(i - t, j - t, 0) + \sum_{k=0}^{t-1} \sum_{h=0}^{t-1} \theta(i - t + k, j - t + h, k, h). \tag{3.1}$$

3.2. *Cohorts born after the jump-off*

We have given above formulas for the calculation of the births $V(0, t)$. For the present discussion it is useful to define $B(k, t)$ as the number of children due to women in age k. We have $V(0, t) = B(15, t) + \cdots + B(44, t)$, where $B(k, t) = \exp[v(k, t - 1) + f(k, t - 1)]$. Using a Taylor-series expansion we have for $1 \leqslant t \leqslant u \leqslant 16$ the covariances for the births

$$\sigma(0, 0, t, u) \approx V(0, t)^{-1} V(0, u)^{-1} \times \sum_{j=15}^{44} \sum_{k=15}^{44} \Big\{ B(j, t)B(k, u) \times \Big[\gamma(j, k, t - 1, u - 1) + \sigma(j - t + 1, k - u + 1, 0) + \sum_{m=0}^{t-2} \sum_{n=0}^{u-2} \theta(j - t + 1 + m, k - u + 1 + n, m, n) \Big] \Big\}. \tag{3.2}$$

The variances of births are obtained by taking $t = u$.

We note that this formula is an approximate one, because it uses a Taylor-series expansion. However, apart from that, it does account exactly for all sources of uncertainty: jump-off population, survival to years $t - 1$ and $u - 1$, and fertility during $t - 1$ and $u - 1$.

The propagation of error for surviving births and for subsequent births can be handled in exactly the same way. However, the formulas become progressively more complicated. Fortunately, considerable simplification is possible by noting that fertility is the dominant source of uncertainty in the forecast of births and surviving births. Therefore, we shall use three approximations. The resulting final "renewal equation" for the covariance of the births is given as expression (A1) in the Appendix. Here we describe verbally the principles behind the approximations. Their exact mathematical interpretation is given in the Appendix.

First, consider the propagation of error calculations for births during a year $t \geqslant 17$. We shall set to zero (*I*) *the covariance between the survival rates of the mothers giving birth at t and the survival rates of their own mothers* (*17 or more years earlier*), and (*II*) *the covariance between the errors in the jump-off population and errors in births of the mothers giving birth at t*. Similarly, even though errors in fertility forecasts are expected to be highly correlated over the forecast years, their correlations decrease over time. For example, suppose the correlation follows an AR(1) pattern over time, with ρ^n the correlation between two ages n years apart. Then, with $\rho = 0.8$ we would only have the correlation 0.03 for $n = 16$. The value $\rho = 0.9$ corresponds to 0.19. Therefore, when considering births to age 0 at time t we shall ignore (*III*) *the covariance*

between the errors in the fertility forecasts for year $t - 1$ and the past births.

An implication of approximation (III) is that current fertility is treated as if it were uncorrelated with the current child bearing population. Some caution is called for when this approximation is used in practice. The forecast errors of ARIMA models, for example, can be highly correlated (e.g., Box and Jenkins 1976, p. 160) suggesting the possibility of a positive correlation. On the other hand, if the Easterlin hypothesis (i.e., cohort size is inversely proportional to cohorts fertility) holds, but is not adequately taken into account in forecasting, then the correlation might be negative.

Note that in applying (I)–(III) recursively, we do let the errors in the jump-off population and past birth rates influence the covariance structure of the births. However, once the covariance structure has been calculated, we act as if the births were generated independently of the jump-off population or survivorship. This has the advantage that we never need to trace the history of a cohort beyond birth in the on-line database implementation of the propagation of error calculations.

For the surviving births, or for $\max\{0, t - 16\} \leqslant j \leqslant i < t$, we get

$$\sigma(i, j, t) \approx \sigma(0, 0, t - i, t - j) + \sum_{m=0}^{i-1} \sum_{n=0}^{j-1} \theta(m, n, t - i + m, t - j + n) + \varepsilon_1(i, j, t). \quad (3.3)$$

Approximation (I) asserts that the residual term $\varepsilon_1(i, j, t)$ is negligible. The covariance between the surviving births in age i at year t ($\max\{0, t - 16\} \leqslant i < t$) and the survivors of the jump-off population in age j at year u ($u \leqslant j$) is given by

$$\sigma(i, j, t, u) \approx \sum_{m=0}^{i-1} \sum_{n=0}^{u-1} \theta(m, j - u + n, t - i + m, n) + \varepsilon_2(i, j, t, u). \quad (3.4)$$

Approximations (I) and (II) assert that $\varepsilon_2(i, j, t, u)$ is negligible. We note that the two above formulas are valid for *all* cohorts born after the jump-off year.

3.3. *Analysis of aggregates*

Consider now an aggregate of ages from age a to age b ($a \leqslant b$) at time t. Denote its true size by $\tilde{V}(a, b, t)$, and let $\tilde{v}(a, b, t) = \log(\tilde{V}(a, b, t))$. The point forecast for the aggregate is simply $V(a, b, t) = V(a, t) + \cdots + V(b, t)$. Define $v(a, b, t) = \log(V(a, b, t))$. To calculate the variance of $\tilde{V}(a, b, t)$ we use a Taylor-series approximation, as before. The result is

$$\operatorname{Var}(\tilde{v}(a, b, t)) \approx V(a, b, t)^{-2} \times \sum_{i=a}^{b} \sum_{j=a}^{b} V(i, t)V(j, t)\sigma(i, j, t), \quad (3.5)$$

where the covariance terms $\sigma(i, j, t)$ have been given earlier. From the user's point of view, the availability of these variances and the resulting probabilistic interpretability are a major advantage of the stochastic approach to the propagation of error.

4. Functional Forecasts

Functional forecasts are forecasts of functions of the population vector, such as the total population size, the number of women in child bearing ages, the number of disabled individuals, or the size of the employed population. The two latter examples differ from the two former ones in that they require additional "participation" or *prevalence* rates, such as the fraction disabled per age-sex group, or the employment rate by age-sex

groups, for their calculation. In principle, forecasts of such subpopulations could be handled in terms of a linear growth model by disaggregating the population not just by age and sex, but also by labor force status or state of disability. This would lead to a situation in which all quantities of interest could be obtained by aggregating elements of the enlarged population vector. The propagation of error could then be handled using the analogues of the variance formula of Section 3.3. However, demographic vital rates are not always available for the finely disaggregated groups, so the only practical possibility is often to apply appropriate prevalence rates to the basic demographic groups. The prevalence approach is also simpler to implement, because it does not require the age-specific in-flow and out-flow rates between the states.

Assume that the population is divided into two mutually exclusive subpopulations. As an example we consider the "disabled" population and its complement, the "healthy" population (cf., Goss 1984, pp. 9–10, 47–48). Let $\Pi(i, t)$ be the forecast of prevalence of disability in age i at time t, with $\pi(i, t) = \log(\Pi(i, t))$. Let the true prevalence be $\tilde{\Pi}(i, t) = \exp(\tilde{\pi}(i, t))$ with $\mathrm{Cov}(\tilde{\pi}(i, t), \tilde{\pi}(j, t)) = \psi(i, j, t)$. In general, the forecast error of the prevalence rate may be correlated with the forecast error of the age-specific survival rates. However, for the present illustration we shall assume that the correlation is zero. This is a reasonable assumption for ages with small prevalence of disability.

Define $W(i, t) = \Pi(i, t)V(i, t)$ and $W(a, b, t) = W(a, t) + \cdots + W(b, t)$. Then, in analogy of Section 3.3., we get for the variance of the prediction error of the disabled population in ages from a to b, at time t, that

$$\mathrm{Var}(\tilde{W}(a, b, t)) \approx W(a, b, t)^{-2} \times \sum_{i=a}^{b} \sum_{j=a}^{b} W(i, t)W(j, t)(\sigma(i, j, t) + \psi(i, j, t)).$$

We see that the covariances add under the assumption of independence. Note that when $\psi(i, j, t) = 0$, the above formula gives the variance of an arbitrary linear combination of the elements of the population vector. A problem for future research is to consider the formulation allowing for the covariance between the prevalence rates and the other relevant variables.

We conclude by noting that there is another, separate class of problems that can be subsumed under the heading functional forecasts. We can consider functions of future vital rates (instead of the population vector), such as the total fertility rate, standardized mortality rates, or life-expectancy. The two former ones are linear functions of the rates, so they can be easily handled. For life-expectancy, which is a non-linear function of the vital rates, a linearizing transform is called for (cf., Alho and Spencer 1990b, p. 222).

5. Computer Implementation

The calculations described in Sections 2 and 3 have been implemented in several computer programs written in C language. Logically, it is convenient to combine the programs into three groups, which we call Programs I, II, and III. A description of program inputs and outputs follows.

Program I

Inputs:

- a file containing the female jump-off population $\mathbf{V}(0)$ by single years of age;
- a file containing age-specific mortality rates $R(j + 1, j, t)$ for each forecast year; if nonzero net migration is assumed, then the mortality rates must be adjusted to account for net migration;

– a file containing age-specific fertility rates $R(0, j, t)$ for child bearing ages for each forecast year.

Output:

– produces a file containing the female population $\mathbf{V}(t)$ by single years of age for each forecast year;

– produces a file containing the births $B(k, t)$ due to each child bearing age for each forecast year.

Program II

Inputs:

– a file containing the standard deviations $\sigma(i, i, 0)^{1/2}$ of the errors in the jump-off population; to get the covariance matrix $\mathbf{\Sigma}(0)$ an AR(1) correlation structure is assumed with the parameter given in the program itself;

– a file containing the standard deviations of the annual errors in the forecast of the rate of change in mortality; these are assumed to be time invariant; their correlation over time is assumed to follow an AR(1) process for each age with a parameter given in the program; their correlation over age is also assumed to follow an AR(1) process for each forecast year; these will determine the covariances $\theta(i, j, t, h)$;

– a file containing the standard deviations of the annual errors in the forecast of the rate of change in fertility; these are assumed to be time invariant; their correlation over time is assumed to follow an AR(1) process for each age with a parameter given in the program; their correlation over age is also assumed to follow an AR(1) process for each forecast year; these will determine the covariances $\gamma(i, j, t, h)$;

– the file containing the numbers of births by age of mother, produced by program I.

Outputs:

– a file containing the covariances of the births $\sigma(0, 0, t, u)$ for $1 \leqslant t, u \leqslant 75$.

Program III

Inputs:

– a file containing the standard deviations $\sigma(i, i, 0)^{1/2}$ of the errors in the jump-off population, as above;

– a file containing the standard deviations of the annual errors in the forecast of the rate of change in mortality, as above;

– the file containing the covariances of the births produced by program II;

– the file containing the point forecast produced by program I.

Outputs:

– the program calculates the standard deviation of the prediction error for the logarithm of the population size of a given group of consecutive ages for each of given consecutive forecast years; the parameters are supplied from the keyboard; the logarithm of the population size and the standard deviation are printed into a text file (name supplied from the keyboard) from which they can be read into some statistical program (such as Minitab) for the purpose of producing probabilistic prediction intervals and graphical displays.

6. Examples of Probabilistic Intervals

How would probabilistic prediction intervals for the sizes of different age-groups be expected to differ from the intervals derived from the current high and low projection variants, in practice? Briefly, in the cases we have studied the high-low intervals for those ages that depend on future fertility have had a much lower probability content than the intervals for ages that do not. Some quantitative estimates of the differences follow.

Alho and Spencer (1985) considered a time-series regression model for age-specific fertility. The model permitted the incorporation of expert judgement, so it could be used to accurately replicate official fore-

casts. In addition, the technique allows for modeling forecast bias. A generous allowance for bias would make the model-based, or *ex ante*, prediction intervals wide, whereas forcing the modeling bias to zero results in the usual prediction intervals. We calibrated the level of bias so that the model would have produced accurate prediction intervals in the past. The adequacy of these intervals was later confirmed by Alho (1991, figure 2). Alho (1984, p. 104) compared the resulting 67% prediction intervals for births to the official Finnish, Norwegian, and U.S. high-low forecasts. Define H = official high projection, M = official middle projection, U = upper end point of a 67% prediction interval, and P = time-series point forecast. The following table can be deduced. It gives the ratios $(H - M)/(U - P)$ for various forecast years in percent (the value for Finland for forecast year 1 is not available).

Table 1. Ratios $(H - M)/(U - P)$ *for various forecast years (in percent)*

Forecast year	1	5	10	15
Finland	n.a.	43	84	66
Norway	21	54	81	71
United States	10	29	39	45

For example, the width of the high-low interval for the United States for the first forecast year was only 10% of the width of the appropriate 67% interval. We conclude that the official intervals would have to be inflated considerably to make them into 67% prediction intervals. It follows that a similar adjustment would be necessary for the survivors of the predicted births.

We now consider survival. Alho and Spencer (1990a, table 2, pp. 616) showed that the forecasts made by the U.S. Office of the Actuary, Social Security Administration, generally underestimate the variability of cause-specific mortality. For example, in ages 60–84 the true standard deviations of the prediction errors are 2–3 times as large as the ones implicitly assumed by the Office of the Actuary. However, when cause-specific mortality is aggregated into age-specific mortality, the intervals of the Office of the Actuary become wider and hence more realistic. This is caused, paradoxically, by the assumption (implicit in the calculation of deterministic projections) that the cause-specific mortality series are perfectly correlated. Overall, the official high-low intervals for age-specific mortality were slightly wider than 95% intervals for ages 40–64 but were narrower than the 95% intervals for other ages. It follows that the probability content of high–low intervals for a cohort of survivors varies from one forecast year to the next (Alho and Spencer 1990a, figure 4, p. 615; cf., Alho and Spencer 1990b, table 2, p. 313).

Combining high–low intervals that have a probability content of less than 67% for future births and the surviving births, and a probability content that sometimes exceeds 95% for certain survivors of the jump-off population, produces forecasts that are very hard to interpret. Using the stochastic methodology for producing prediction intervals would eliminate this problem.

More details of the *ex ante* estimation of forecast errors can be found in the references cited. Keilman (1990) has recently provided an extensive discussion of various aspects of *ex post* error estimation in national population forecasts, with emphasis on Dutch experiences. As indicated above (see also Alho 1991) such estimates can be used to derive *ex ante* error estimates for future forecasts.

7. Concluding Remarks

In this paper we have derived analytical equations for the propagation of error in stochastic population forecasts. The major

contribution is the approximate "renewal equation" for the covariance structure of the birth sequence. The results of these covariance calculations are stored into a file, which, together with other data files, provides a basis for the calculation of the variance of the prediction error for population aggregates and for functional forecasts.

This hierarchical system of calculation is easily implemented as a database. Database concepts have already been used to provide public access to the decennial census data tapes in the United States. Similar developments have occurred in many European countries. For example, an extensive system of regional and time-series data (ALTIKA, ASTIKA) is accessible on-line in Finland. We believe that databases will be the future form of implementing population forecasts, as well.

Our approach to the building of a database was not to try to come up with a completely self-contained system, because the likely users of stochastic forecasts would typically already have statistical programs that they are used to. Therefore, only those calculations that would be tedious to implement within a statistical package were written in C language. Our experience in writing program II was that the specific form of the covariance structures makes a tremendous difference in the speed of calculation. However, one of the useful features of our database implementation is that the very time-consuming calculations are only done once, and they can be carried out on a faster machine if the need arises. All our calculations were manageable on a 386-based micro computer.

Our work demonstrates the feasibility of the database implementation of stochastic forecasts. However, the example we considered was a simple one, and many questions remain. First, a comparison of the accuracy and practical feasibility of our analytic approach and a simulation approach would be illuminating. It may be that the results depend on the covariance structures assumed, as well as on the forecast horizon. Second, the study of possible covariance structures of the forecast errors is of interest. When forecasts are made using formal statistical methods, then the covariance structure can typically be derived as a by-product of the estimation procedure (e.g., Box and Jenkins 1976, pp. 159–160). However, since expert judgement is an integral part of at least the current official forecasts, other approaches, such as those in Alho (1991) are called for. Third, the implications of various ways of handling the propagation of error in functional forecasts is unclear. We suspect that the independence assumption used above is widely applicable, but the conditions under which it fails need investigation.

Appendix

Propagation of error for cohorts born during $t \geq 17$

The residual terms ε_1 and ε_2 introduced in formulas (3.3) and (3.4) are defined by

$$\varepsilon_1(i, j, t) =$$

$$V(0, t-j)^{-1} \sum_{k=15}^{44} B(k, t-j) \sum_{m=0}^{i-1} \sum_{n=0}^{t-2-j} \theta(m, k-t+j+1+n, t-i+m, n)$$

$$+ V(0, t-i)^{-1} \sum_{k=15}^{44} B(k, t-i) \sum_{n=0}^{j-1} \sum_{m=0}^{t-2-i} \theta(k - t + i + 1 + m, n, m, t - j + n)$$

and

$$\varepsilon_2(i, j, t, u) = V(0, t-i)^{-1} \sum_{k=15}^{44} B(k, t-i) \left[\sigma(k - t + i + 1, j - u, 0) + \sum_{m=0}^{t-2-i} \sum_{n=0}^{u-1} \theta(k - t + i + 1 + m, j - u + n, m, n) \right].$$

The sums involving θ above involve covariances between the survival of mothers giving birth at $t - i$ or $t - j$, and the survival of their own mothers. For the purpose of computing the covariance structure of the birth sequence we set these to zero. This is the content of approximation (I). The remaining sum involving σ is set to zero according to assumption (II) because it arises from a covariance between the jump-off population (ages $j - u$) and births (year $t - i$).

Consider now the covariance between births in years $t = 17, \ldots, 32$. Their uncertainty derives primarily from four independent components: current birth rates, jump-off population, past births, and survival rates. To abbreviate notation it is convenient to define some operators. Let u and t be fixed and define

$$\mathbf{K}_L = V(0, u)^{-1} \sum_{k=15}^{u-2} B(k, u), \quad \mathbf{K}_U = V(0, u)^{-1} \sum_{k=u-1}^{44} B(k, u), \quad \mathbf{K} = \mathbf{K}_L + \mathbf{K}_U,$$

$$\mathbf{J}_L = V(0, t)^{-1} \sum_{j=15}^{t-2} B(j, t), \quad \mathbf{J}_U = V(0, t)^{-1} \sum_{j=t-1}^{44} B(j, t), \quad \mathbf{J} = \mathbf{J}_L + \mathbf{J}_U,$$

$$\mathbf{H}_j = V(0, t - 1 - j)^{-1} \sum_{h=15}^{44} B(h, t - 1 - j),$$

$$\mathbf{I}_k = V(0, u - 1 - k)^{-1} \sum_{i=15}^{44} B(i, u - 1 - k),$$

so that, for example, $\mathbf{K}_L x_k = V(0, u)^{-1} \Sigma_{k=15}^{u-2} B(k, u) x_k$. A first-order Taylor approximation shows that $\sigma(0, 0, t, u)$ is approximately equal to

$$\begin{aligned}
&\mathbf{K}\mathbf{J}\gamma(k, j, u - 1, t - 1) + \mathbf{K}_U\mathbf{J}_U\sigma(k - u + 1, j - t + 1, 0) \\
&\quad + \mathbf{K}_U\mathbf{J}_U \sum_{n=0}^{u-2} \sum_{m=0}^{t-2} \theta(k - u + 1 + n, j - t + 1 + m, n, m) \\
&\quad + \mathbf{K}_L\mathbf{J}_L\sigma(0, 0, u - 1 - k, t - 1 - j) \\
&\quad + \mathbf{K}_L\mathbf{J}_L \sum_{n=0}^{k-1} \sum_{m=0}^{j-1} \theta(n, m, u - 1 - k + n, t - 1 - j + m) \\
&\quad + \mathbf{K}_L\mathbf{J}_U \sum_{n=0}^{k-1} \sum_{m=0}^{t-2} \theta(n, j - t + 1 + m, u - 1 - k + n, m) \\
&\quad + \mathbf{K}_U\mathbf{J}_L \sum_{n=0}^{u-2} \sum_{m=0}^{j-1} \theta(k - u + 1 + n, m, n, t - 1 - j + m) \\
&\quad + \varepsilon_3(t, u) + \varepsilon_4(t, u) + \varepsilon_5(t, u) \qquad \text{(A1)}
\end{aligned}$$

with

$$\varepsilon_3 = \mathbf{K}\mathbf{J}_L\mathbf{H}_j\gamma(k, h, u-1, t-2-j) + \mathbf{K}_L\mathbf{J}\mathbf{I}_k\gamma(i, j, u-2-k, t-1),$$

$$\begin{aligned}\varepsilon_4 = {} & \mathbf{K}_L\mathbf{J}_L\sum_{n=0}^{k-1}\mathbf{H}_j\sum_{p=0}^{t-3-j}\theta(n, h+j-t+p+2, u-1-k+n, p) \\ & + \mathbf{K}_U\mathbf{J}_L\sum_{n=0}^{u-2}\mathbf{H}_j\sum_{p=0}^{t-3-j}\theta(k-u+1+n, h+j-t+p+2, n, p) \\ & + \mathbf{K}_L\mathbf{J}_L\sum_{m=0}^{j-1}\mathbf{I}_k\sum_{q=0}^{u-3-k}\theta(i+k-u+q+2, m, q, t-1-j+m) \\ & + \mathbf{K}_L\mathbf{J}_U\sum_{m=0}^{t-2}\mathbf{I}_k\sum_{q=0}^{u-3-k}\theta(i+k-u+q+2, j-t+1+m, q, m),\end{aligned}$$

and

$$\begin{aligned}\varepsilon_5 = {} & \mathbf{K}_U\mathbf{J}_L\mathbf{H}_j\sigma(k-u+1, h+j-t+2, 0) \\ & + \mathbf{K}_L\mathbf{J}_U\mathbf{I}_k\sigma(i+k-u+2, j-t+1, 0).\end{aligned}$$

Details of the derivation are available from the authors. Approximation (III) asserts that ε_3 is negligible, approximation (II) asserts that ε_5 is negligible, and approximation (I) asserts that ε_4 is negligible.

Approximations (I)–(III) imply that (A1) provides an approximation to $\sigma(0, 0, t, u)$ for *all* other pairs of birth years u, $t \geqslant 1$, if (i) we replace by zero any such sum in (A1), for which the lower limit of summation exceeds the upper limit, (ii) the upper limit of summation is not allowed to exceed 44, i.e., take $\min\{t-2, 44\}$ instead of $t-2$, and $\min\{u-2, 44\}$ instead of $u-2$, (iii) the lower limit of summation is not allowed to extend below 15, i.e., take $\max\{t-1, 15\}$ instead of $t-1$, and $\max\{u-1, 15\}$ instead of $u-1$. Since the above formula expresses the covariance $\sigma(0, 0, t, u)$ in terms of past covariances $\sigma(0, 0, t-1-k, u-1-h)$, we can think of the formula as a "renewal equation" for the covariances.

8. References

Akaike, H. (1974). Markovian Representation of Stochastic Processes and Its Application to the Analysis of Autoregressive Moving Average Processes. Annals of the Institute of Statistical Mathematics, 26, 363–387.

Alho, J. (1984). Probabilistic Forecasts. The Case of Population Projections. Scandinavian Housing and Planning Research, 1, 99–105.

Alho, J. (1991). Stochastic Methods in Population Forecasting. International Journal of Forecasting, 6, 521–530.

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

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

Alho, J. and Spencer, B.D. (1990b). Effects of Targets and Aggregation on the Propagation of Error in Mortality Forecasts.

Mathematical Population Studies, 2, 209–227.

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. Actuarial Education and Research Fund, Itasca, IL.

Bell, W. (1988). Applying Time Series Models in Forecasting Age-specific Fertility Rates. Statistical Research Division Report Series, Census/SRD/RR-88/19. Washington, D.C.: U.S. Bureau of the Census.

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

Bozik, J.E. and Bell, W.R. (1987). Forecasting Age-specific Fertility Using Principal Components. Statistical Research Division Report Series, Census/SRD/RR-87/19. Washington, D.C.: U.S. Bureau of the Census.

Cohen, J. (1977). Ergodicity of Age Structure in Populations with Markovian Vital Rates III: Finite State Moments and Growth Rate. Advances in Applied Probability, 9, 462–475.

Davis, W.W. (1988). Calculation of the Variance of Population Forecasts. Statistical Research Division Report Series, Census/SRD/RR-88/20. Washington, D.C.: U.S. Bureau of the Census.

Feichtinger, G. (1970). Stochastische Modelle Demographischer Prozesse. Berlin: Springer. (In German).

Goodman, L. (1968). Stochastic Models for the Population Growth of the Sexes. Biometrika, 55, 469–487.

Goss, S.C. (1984). Long-Range Estimates of the Financial Status of the Old-Age, Survivors, and Disability Insurance Program, 1983. Actuarial Study No. 91. Washington, D.C.: Office of the Actuary.

Hyppölä, J., Tunkelo, A., and Törnqvist, L. (1949). Calculations on the Population of Finland, Its Renewal, and Its Future Development. Tilastollisia Tiedonantoja 38. Helsinki: Central Statistical Office of Finland. (In Finnish).

Keiding, N. and Hoem, J. (1976). Stochastic Stable Population Theory with Continuous Time, I. Scandinavian Actuarial Journal, 150–175.

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

Kendall, D. (1949). Stochastic Processes and Population Growth. Journal of the Royal Statistical Society, Ser. B, 11, 230–264.

Keyfitz, N. (1977). Introduction to the Mathematics of Population, with Revisions. Reading, MA: Addison-Wesley.

Land, K. (1986). Methods for National Population Forecasts: A Review. Journal of the American Statistical Association, 81, 888–901.

Lee, R.D. and Carter L. (1990). Modeling and Forecasting US Mortality. A paper presented at the Annual Meeting of the Population Association of America in Toronto, Canada, May 1990.

McNown, R. and Rogers, A. (1990). Forecasting Cause-specific Mortality Using Time-Series Methods. Working Paper 90-4, Population Program, University of Colorado at Boulder.

Pflaumer, P. (1988). Confidence Intervals for Population Projections Based on Monte Carlo Methods. International Journal of Forecasting, 4, 135–142.

Pollard, J.H. (1968). A Note on Multi Type Galton-Watson Processes With Random Branching Probabilities. Biometrika, 55, 589–590.

Pollard, J.H. (1973). Mathematical Models for the Growth of Human Populations. London: Cambridge University Press.

Rogers, A. (1986). Parametrized Multistate Population Dynamics and Projections. Journal of the American Statistical Association, 81, 48–63.

Shryock, H.S., and Siegel and Associates, Condensed edition by E.G. Stockwell (1976). The Methods and Materials of Demography. New York: Academic Press.

Sykes, Z. (1969). Some Stochastic Versions of the Matrix Model for Population Dynamics. Journal of the American Statistical Association, 64, 111–130.

Thompson, W.S. and Whelpton, P.K. (1933). Population Trends in the United States. New York: McGraw-Hill.

U.S. Bureau of the Census (1984). Projections of the Population of the United States, by Age, Sex, and Race: 1983 to 2080. Current Population Reports, Ser. P 25, No. 952. Washington, D.C.: U.S. Department of Commerce.

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

Whelpton, P.K. (1928). Population of the United States, 1925 to 1975. American Journal of Sociology, 34, 253–270.

Whelpton, P.K. (1936). An Empirical Method of Calculating Future Population. Journal of the American Statistical Association, 31, 457–473.

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 D.C.: U.S. Government Printing Office.

Received August 1990
Revised May 1991