BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

RESEARCH REPORT SERIES, No. RR–92/07

THE CONTRIBUTION OF DIFFERENT SOURCES OF ERROR TO THE ACCURACY OF NATIONAL POPULATION FORECASTS IN THE U.S.

by

Juha M.Alho[1]
Department of Statistics
University of Joensuu
Finland

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233.

[1] The author wrote this as the final report in connection with Joint Statistical Agreement (JSA) 90–49 between the University of Illinois, where he was employed at the time, and the Bureau of the Census.

Report issued: June 15, 1992

THE CONTRIBUTION OF DIFFERENT SOURCES OF ERROR TO THE ACCURACY OF NATIONAL POPULATION FORECASTS IN THE U.S.

Juha M. Alho[1]

[1] The author would like to thank the U.S. Bureau of the Census for supporting this research through a Joint Statistical Agreement. Thanks are also due to Professor Bruce D. Spencer, Northwestern University, for comments on a draft manuscript. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

Contents:

1. Introduction

Consider a female population disaggregated by single years of age. Suppose we have available a forecast that is thought to be unbiased for the logarithm of the population vector. The problem we consider is, what is the contribution of the different vital processes to the forecast error of different ages? Alho and Spencer (1991) provided approximate propagation of error formulas needed in such an analysis. The derivation assumed that certain covariance terms would be so small that they could be neglected. A related issue is, therefore, how accurate are these approximate formulas in different circumstances?

In the general case forecast accuracy depends not only on the methods used to forecast vital rates, but also on how variable (or "volatile") they are (cf. Alho, 1991, Fig. 1, p. 524). In particular, since mortality, migration and fertility all influence the uncertainty of births and surviving births, it is critical to have an idea of the relative uncertainty of these processes (cf. Keilman, 1990, pp. 85-104). In addition, the propagation of error calculations are potentially influenced by the age-structure of the jump-off population and the trend in forecast births (cf., Smith, 1987). For example, a large error in the size of a child bearing age group has a large impact on the error of births when the cohort is in ages 20-30 in which fertility is high, but has less of an impact when the cohort is in ages 40-50 in which fertility is low.

A key element in any propagation of error calculations is the handling of "births to births". In other words, when the cohorts born after the jump-off year will begin to have children of their own, the uncertainty of fertility influences the uncertainty of births in two ways: (1) uncertainty of fertility at current forecast year has a direct impact, (2) the uncertainty of births 16 or more years earlier influences the uncertainty in the number of women in child bearing ages and will, thus, have an indirect influence.

Because of these factors, a two-pronged approach to the problem will be adopted. First, we will define the simplest possible setup that still retains the "births to births" dynamics. The key simplification is to assume that the errors due to different vital processes have a uniform impact (in a log-scale) for all ages. Although unexciting as an error structure of a forecast, this setup permits an analytical treatment of the propagation of error. Up to three generations of births will be considered. A program written in (Turbo) C that implements the calculations is presented in Appendix I. In particular, we note the impact of *stable population growth*, in which the population grows or declines exponentially, but does not change its age structure. Second, the actual U.S. female population will be considered. Slightly simplified versions of the vector ARIMA forecasts developed by Bell and Monsell (1991) are used to derive the moments of the prediction errors. A (Turbo) C program implementing these calculations is in Appendix II. In both cases both the contributions of the different sources of error and the accuracy of the propagation of error formulas of Alho and

Spencer (1991) will be considered.

2. Linear Growth

Following the notation of Alho and Spencer (1991) we 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.

Age j refers to an age-group of females who have had their jth birthday, 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 $(s+1)x(s+1)$ 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.

The *linear growth model* specifies that

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

This model describes a closed female population (i.e., there is no migration). By replacing the life table survival rates by the so-called *census survival rates*, migration can also be incorporated.

Suppose that the *jump-off population* V(0) is strictly positive, or V(j,0) > 0 for j = 0,...,s, and that the age-specific fertility rates and survival rates are strictly positive. Define,

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

Under the linear growth model,

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

The following formulas are among those derived in Alho and Spencer (1991, formulas (2.1)-(2.7)). Consider the survivors of the jump-off population. For $1 \leq t \leq j \leq s-1$ we have

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

A somewhat more complex formula is required for the last age j = s (Alho and Spencer, 1991). However, no new issues arise from the consideration of the last age, so we will ignore it in the sequel. Consider the births. 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,...,16,

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

Surviving the first sixteen birth cohorts we have

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

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

For the "second generation of births", or for the years t = 17,...,32, we have that

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

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

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". The same formula works for t > 32 as well, if we use the convention that the first sum over k is taken to be zero for $t \geq 32$, and the upper limit of the second sum is kept at 44 for t > 46.

Surviving the birth cohorts born during the years t = 17,...,32, we apply the above formula for $\max\{0, t-32\} \leq j \leq t-16$.

A *probabilistic version* of the linear growth model is obtained when we define $\tilde{v}(j,t)$ as the true (random) value with the unbiased forecast $E[\tilde{v}(j,t)] = v(j,t)$. The true vital rates $\tilde{f}(j,t)$ and $\tilde{r}(j,t)$ will be similarly related to their forecasts $f(j,t)$ and $r(j,t)$. All formulas given above will remain valid when we replace v, f, and r, by $\tilde{v}$, $\tilde{f}$, and $\tilde{r}$.

3. Assumptions Concerning the Components of Error

We assume that the errors come from four sources: jump-off population, mortality, migration, and fertility. Although these sources may, in some circumstances, be to some extent correlated, it appears both on theoretical and empirical grounds that the assumption of independence is quite reasonable (cf., Alho and Spencer, 1991; Lee and Tuljapurkar, 1991). *Independence will be assumed throughout.*

We will now present models and assumptions that will be used in the subsequent analyses. We first review the existing literature on what is known about the error structure of each source, and then determine the simplest realistic model that can be used in our analytical work. The simple models typically assume that errors for a component are perfectly correlated over age and they have a constant coefficient of variation. As such they will permit an approximate comparison of the relative contribution of the four sources to the forecast error. In the numerical analyses, a more refined set of assumptions will be used.

3.1. Jump-off Population

Errors in the jump-off, or starting, population of a forecast derive from the imperfections of censuses (or a population register, as the case may be); errors in the registration of births, deaths, and migration; and classification errors that put individuals into wrong age/sex/region etc. categories. An example of this type of uncertainty is the error of the dual system estimator and the estimators based on the so-called demographic analysis that are used to adjust U.S. censuses for undercount (cf., Mulry and Spencer, 1991). Although some element of forecasting may be involved in the specification of a jump-off population, for example when preliminary statistical tabulations are used, the error basically derives from the uncertainty of the basic demographic data. As such the errors in the jump-off population might be expected to be smaller than the forecast errors of the vital rates. However, for very short term forecasts of the survivors of the jump-off population the jump-off error can, in some ages, dominate.

In Alho and Spencer (1985, p. 310) a simple specification that was based on the estimated accuracy of the 1980 census was developed. It assumed that the $\hat{v}(j,0)$'s are perfectly correlated with age, $\hat{v}(j,0) = v(j,0) + e_{JO}$ with

$$e_{JO} \sim N(0, \sigma_{JO}^2).$$

(Note that the subscript of e_{JO} and σ_{JO}^2 refers to "jump-off population" and not to age j.) Based on the uncertainty of the amount of illegal immigration it was estimated that the value σ_{JO} = 0.00325 might adequately represent the uncertainty of the census. This means that a 95% confidence interval for population size would have been ±0.65% of the point forecast. Note that this specification assumes that the coefficient of variation of any population aggregate, from single years of age to the total population, has a constant coefficient of variation of 0.00325. Other considerations, such as the number of homeless, might imply a larger standard deviation.

3.2. Mortality

Alho and Spencer (1990a, 1990b) have analyzed the mortality forecasts of the U.S. Office of the Actuary. The so-called *random line model* that assumes that forecast errors are perfectly correlated over time, and a more realistic *Brownian motion model* in which the forecast errors behave like Brownian motion (or random walk in discrete time), were studied and estimated.

Bell and Monsell (1991) and Lee and Carter (1990) have used principal component techniques to analyze age-specific mortality in the United States. [This technique finds an orthonormal basis for the linear space spanned by the data matrix, consisting of the eigenvectors of the crossproduct matrix. Time series analysis is used to model the change in the coordinates, or the principal components, when observed time-series data are represented in the new basis]. Bell and

Monsell used a full set of principal components and built a vector ARIMA model for the five most important ones. From the point of view of variance calculations it is noteworthy that in their 1991 paper Bell and Monsell differenced the first three coordinate series. Later they have differenced the first five. In both cases first order autoregressive models were built for the series. The differenced series will eventually dominate the error variance, because their variance grows at least linearly, whereas the variance of the remaining stationary error components has an upper bound. Lee and Carter (1990) identified a random walk model with a drift for the first principal component, so their error variance of the same order of magnitude as that of Bell and Monsell.

All the analyses mentioned above suggest that a nonstationary model is necessary for describing the forecast error. At least for the purpose of our analytical setup, the simplest of such models, namely random walk appears also to be sufficient in the sense of giving the empirically observed order of magnitude for forecast error. More precisely, we will assume that

$$\tilde{r}(j,t) = r(j,t) + e_{\mu}(t),$$

where

$$e_{\mu}(t) = \sum_{k=0}^{t-1} \varepsilon_{\mu}(k),$$

with

$$\varepsilon_{\mu}(k) \sim N(0, \sigma_{\mu}^{2}) \text{ i.i.d. for } k = 0,1,...$$

Based on Alho and Spencer (1990a, Table 3, p. 616) a reasonable range of values for σ_μ can be derived as follows. Average value for $5\sigma_\mu$ is between 0.2 and 0.4 in ages 25-90 for cause specific mortality. Errors in causes were found to be close to independent with heart diseases, cancer, and vascular diseases typically claiming well over a half of mortality. This suggests that the logarithm of their sum, age-specific mortality, should have one half the standard deviation of the logarithms of the components, or it should range from 0.1 to 0.2. This yields values for σ_μ in the range of 0.02 to 0.04. How do these estimates translate into errors of the survival probabilities?

Since $\exp(\tilde{r}(j,t))$ is the one year probability of survival, then $-\tilde{r}(j,t)$ must be the mortality rate. Above, we have considered (approximately) the model $\tilde{r}(j,t) = r(j,t)(1 + e_\mu(t)) = r(j,t) + r(j,t)e_\mu(t)$. It follows that by defining $e_S(j,t) = r(j,t)e_\mu(t)$, we get the desired representation $\tilde{r}(j,t) = r(j,t) + e_S(j,t)$. (Note that the subscript "S" refers to "survival" here rather than last age s.) Even though $e_S(j,t)$ clearly depends on j, for our analytical framework we will replace all such factors by their average value, $e_S(j,t) \equiv e_S(t)$. In analogy, we define the annual increments $\varepsilon_S(k)$,

$$e_S(t) = \sum_{k=0}^{t-1} \varepsilon_S(k),$$

with

$$\varepsilon_S(k) \sim N(0, \sigma_S^2) \text{ i.i.d. for } k = 0,1,\ldots$$

It remains to specify a range for σ_S. Age-standardized death rates for females in the United States have been recently in the range of 0.006. Multiplying the σ_μ values above by this gives an approximate range of 0.0001 - 0.0002 for σ_S. (This shows that the uncertainty of the jump-off population can, indeed, be dominant in some short term forecasts.) These estimates do not take into account the fact that mortality varies with age, so that it is about 1/10 of the standardized mortality rate in ages 10-20 and over ten times the standardized rate in ages 70+. It follows that in applying the model for mortality to young ages we may want to consider values as low as 1/10 of the range given above, and for old ages we may want to use ten times as large values as the range given.

3.3. Migration

The uncertainty of migration forecasts appears to be less well studied than that of the other vital processes. Although migration is regulated through legislation in most, if not all, countries, both the presence of illegal immigration, undocumented out-migration, and unpredictable changes in regulations make migration hard to predict. Abrupt changes from year to year are possible.

For the analytical framework we hypothesize that migration has an unpredictable effect of the same relative magnitude from year to year. We can incorporate this into our model by assuming that any secular trends in migration are absorbed into the forecasts r(j,t) and by adding an error component to reflect the uncertainty of the forecast. We assume that

$$\tilde{r}(j,t) = r(j,t) + e_S(t) + e_M(t),$$

where

$$e_M(t) \sim N(0, \sigma_M^2) \text{ i.i.d. for } t = 0,1,2,...$$

are independent of the $e_S(t)$'s. This implies that the effect of migration is perfectly correlated over ages. The magnitude of the annual deviations was considered in Alho and Spencer (1985, p. 312). An analogous, but slightly simpler (and less realistic) calculation can be given as follows. U.S. Bureau of the Census (1982) assumed that the high low range for the annual net migration was 0.25-0.75 million. Assuming that half of the migrants are women, this translates into a range of 0.125-0.375 million. Assuming a female population of 125 million, this translates into a range of 0.001-0.003 out of the total population. Suppose the width of the interval, 0.002, can be interpreted as a 67% prediction interval, then we get that $\sigma_M = 0.001$. Obviously, somewhat smaller and larger values could be entertained.

3.4 Fertility

The uncertainty of fertility has been extensively studied. Alho and Spencer (1985) and Alho (1984, 1991) have considered the adequacy of the official high-low intervals as prediction intervals and proposed alternatives. Lee (1974) and Carter and Lee (1986) have considered ARIMA type techniques for modeling fertility. Bozik and Bell (1987) and Thompson et al. (1989) have used principal components and vector ARIMA models to forecast fertility. McDonald (1979, 1981) considered econometric approaches. For our purposes it is important that all authors agree on one thing, namely that the birth and fertility series are *nonstationary*.

For the study of the accuracy of relatively short term forecasts, we believe that the random walk model for forecast error is again the simplest realistic model. In analogy of the model for mortality we will assume that

$$\tilde{f}(j,t) = f(j,t) + e_F(t),$$

where

$$e_F(t) = \sum_{k=0}^{t-1} \varepsilon_F(k),$$

with

$$\varepsilon_F(k) \sim N(0, \sigma_F^2) \text{ i.i.d. for } k = 0,1,\ldots$$

Based on Alho and Spencer (1985, Table 3, p. 313) a reasonable range of values for σ_F is from 0.05 to 0.10. We see a dramatic difference as compared to all the previous sources of error. Even the lower limit, which we view as an unrealistically low value in the sense that it would have produced too narrow confidence intervals in the past, implies more uncertainty than any of the previous sources of error.

4. Analytical Treatment

We will now summarize the assumptions used for the analytical treatment and evaluate the contributions of different sources of error in four stages. First, we evaluate the survivors of the jump-off population; second, the first generation of births during the forecast period, and their survivors; third, the second generation of births, i.e., those born in forecast years t = 17,...,32, and their survivors; and fourth, the third generation of births during the years t = 33,..., 48. Forecast horizons exceeding 48 years (for births) will not be considered.

Based on the preliminary discussion of Section 3, the forecast errors are specified as follows.

(i) Jump-off population: $\tilde{v}(j,0) = v(j,0) + e_{JO}$ with $e_{JO} \sim N(0, \sigma_{JO}^2)$.

(ii) Mortality and migration: $\tilde{r}(j,t) = r(j,t) + e_S(t) + e_M(t)$, where $e_S(t)$ and $e_M(t)$ are independent, with $e_M(t) \sim N(0, \sigma_M^2)$ i.i.d. for $t = 0,1,2,...$, and $e_S(t) = \varepsilon_S(0) +...+ \varepsilon_S(t-1)$ with $\varepsilon_S(k) \sim N(0, \sigma_S^2)$ i.i.d. for $k = 0,1,...$

(iii) Fertility: $\tilde{f}(j,t) = f(j,t) + e_F(t)$, where $e_F(t) = \varepsilon_F(0) + ...+ \varepsilon_F(t-1)$ with $\varepsilon_F(k) \sim N(0, \sigma_F^2)$ i.i.d. for $k = 0,1,...$

Assumptions (i) - (iii) leave one minor issue concerning future births unresolved. The births during the forecast year t − 1 that contribute to V(0,t) should be subject to migration and mortality during the year t − 1 in two ways. First, migration and mortality affect the number of women in child bearing ages during the fraction of the year before they give birth. Second, the children in age 0 are subject to migration and mortality before year t is reached. Both effects can easily be accounted for in the point forecast by adjusting the fertility rates. However, for the uncertainty analysis some additional convention is needed. Under our simplified assumptions that assume perfect correlation over time the simplest convention seems to be to add this uncertainty factor into fertility, and take

$$\tilde{f}(j,t) = f(j,t) + e_F(t) + e_S(t) + e_M(t).$$

This will be assumed below.

A major analytical advantage in the setup we have defined is that the jump-off population, mortality, and migration have the same impact on the uncertainty of all ages in a given forecast year. To formulate this result more precisely let us define

$$Q(t) = \sum_{k=0}^{t-1} (e_S(k) + e_M(k)).$$

In other words, Q(t) is the error component due to mortality and migration that we use to get from jump-off to forecast year t. Since,

$$\sum_{k=0}^{t-1} e_S(k) = \sum_{k=0}^{t-1} \sum_{j=0}^{k-1} \varepsilon_S(j)$$

$$= \sum_{k=0}^{t-1} (t - k)\varepsilon_S(k),$$

we have that

$$\mathrm{Var}(\sum_{k=0}^{t-1} e_S(k)) = \sigma_S^2 \sum_{k=1}^{t} k^2 = \sigma_S^2 h(t),$$

where we have written

$$h(t) = (2t + 1)(t + 1)t/6,$$

for short. It follows that

$$\mathrm{Var}(Q(t)) = t\sigma_M^2 + h(t)\sigma_S^2.$$

Lemma 1. The error component due to jump-off, migration, and mortality in any $\tilde{v}(j,t)$, $j = 0,\ldots,s$; $t = 1,2,\ldots,$ is e_{JO} + $Q(t)$, with the variance

$$\mathrm{Var}(e_{JO} + Q(t)) = \sigma_{JO}^2 + t\sigma_M^2 + h(t)\sigma_S^2.$$

Proof. The variance formula follows from the independence of e_{JO} and $Q(t)$. It remains to show that the same uncertainty factor works for all ages. Consider the survivors of the jump-off population. The claim follows directly from the formula of

$v(j,t)$, when $1 \leq t \leq j \leq s-1$ (we omit here the additional detail of the last age). Consider the first generation of births, i.e., births during the forecast years $1 \leq t \leq 16$, and their survival. Let $\max\{0, t-16\} \leq j < t$, and consider $\tilde{v}(j,t)$. These individuals are survivors of $\tilde{V}(0,t-j)$. The jump-off, mortality, migration component of the mothers who gave birth to these children is $e_{JO} + Q(t-j-1)$. We add to this the factor due to migration and mortality during the year of birth, $e_M(t-j-1) + e_S(t-j-1)$, and the subsequent factors after birth, $(e_M(t-j) + e_S(t-j)) +\ldots+ (e_M(t-1) + e_S(t-1))$. The sum is $e_{JO} + Q(t)$. The result follows by induction on birth generations.

4.1. Survivors from the Jump-off Population

First, consider the survivors of the cohorts present at jump-off. Specializing Lemma 1 to $j \geq t \geq 1$, yields $\mathrm{Var}(\tilde{v}(j,t)) = \sigma_{JO}^2 + t\sigma_M^2 + \sigma_S^2 h(t)$. Figure 1 indicates the relative importance of the jump-off error (solid line), error due to migration (dashed line), and error due to mortality (dotted line). The graph depicts the standard deviation due to each of these sources when the following parameter values are assumed: $\sigma_{JO} = 0.00325$, $\sigma_M = 0.001$, and $\sigma_S = 0.0001$. The top curve depicts the standard deviation of $\tilde{v}(j,t)$ that combines all these sources (dash-dotted line). Note that under our assumptions these curves do not depend on age j. We see that the uncertainty of survival (or mortality) dominates after about twenty years, so that after about thirty forecast years the contribution of the other factors could almost be ignored.

During the first 10 years, however, the uncertainty of the jump-off population is the most important, and even the uncertainty of migration is more important (with these parameter values) than that of mortality.

Note that Figure 1 can be also used to gauge the relative importance of the three error sources under alternative parameter values. The JO, M, and S-curves depicted are *proportional* to the σ-values used. Recall, in particular, that the uncertainty of survival can be as little as 1/10 of the value used in Figure 1 for young ages. Therefore, when considering the survival in, say, ages 1-10 at jump-off, the true curve would be initially about 1/10 of the curve depicted. It would start to increase at an accelerating rate in later forecast years because of the rapid increase in mortality. However, it would stay clearly below the curve depicted in the whole range depicted. Similarly, consider those in, say, age 50 at jump-off. Their true curve would start slightly below the curve depicted, but it would increase faster and quickly exceed the curve depicted. We see that in all cases the uncertainty of survival will eventually exceed the uncertainty of the other two sources, and the uncertainty of migration will eventually exceed that of the jump-off population. During the first twenty forecast years all sources appear to be important with any reasonable parameter combinations.

4.2. The First Generation of Births and Their Survival

Consider the first 16 forecast years, when there are still no "births to births". We call these the *first generation* of births. Define $B(j,t)$ as the forecast number of births to women in age $15 \leq j \leq 44$ that are in age 0 at t, or $V(0,t) = B(15,t) + \ldots + B(44,t)$. As in Alho and Spencer (1991) we use a Taylor-series development to calculate the covariances of the births. For $\tilde{v}(0,t)$ we get, in particular, that

$$\tilde{v}(0,t) \approx v(0,t) + V(0,t)^{-1} \sum_{j=15}^{44} B(j,t)[\log(\tilde{B}(j,t)) - \log(B(j,t))].$$

The covariance between births in the year t and in the year u, $1 \leq t \leq u \leq 16$, is

$$\mathrm{Cov}(\tilde{v}(0,t),\tilde{v}(0,u)) \approx$$

$$V(0,t)^{-1}V(0,u)^{-1} \sum_{j=15}^{44} \sum_{k=15}^{44} B(j,t)B(k,u)[\sigma_{J0}^2 + t\sigma_F^2 + \mathrm{Cov}(Q(t),Q(u))].$$

Using the independence of the $e_S(t)$'s and $e_M(t)$'s, and the relation $e_S(t) = \varepsilon_S(0) + \ldots + \varepsilon_S(t-1)$ we get that

$$\mathrm{Cov}(Q(t),Q(u)) = t\sigma_M^2 + \mathrm{Cov}(\sum_{m=0}^{t-1} (t-m)\varepsilon_S(m), \sum_{n=0}^{u-1} (u-n)\varepsilon_S(n))$$

$$= t\sigma_M^2 + \sum_{k=0}^{t-1} (t-k)(u-k)\sigma_S^2.$$

The sum on the right hand side can be written as $\sigma_S^2(u + t + 1)$ $(t + 1)t/6$. It follows that for $1 \leq t \leq u \leq 16$ we have that

$$\mathrm{Cov}(\tilde{v}(0,t),\tilde{v}(0,u)) \approx \sigma_{JO}^2 + t\sigma_F^2 + \sigma_S^2(u + t + 1)(t + 1)t/6\}.$$

In particular, the variances are

$$\mathrm{Var}(\tilde{v}(0,t)) \approx t\sigma_F^2 + \sigma_{JO}^2 + t\sigma_M^2 + \sigma_S^2 h(t).$$

This permits the comparison of the contribution that the jump-off population, mortality, and migration on the one hand, and fertility on the other, make to the uncertainty of the births. The top curve of Figure 1 shows how the standard deviation due to the first three components $(\sigma_{JO}^2 + t\sigma_M^2 + \sigma_S^2 h(t))$ varies as a function of the forecast year. In particular, we see that the curve exceeds the value 0.02 at forecast year $t = 46$, approximately. However, this is less than half of the value of $t^{1/2}\sigma_F$ at $t = 1$, when $\sigma_F = 0.05$, a low value. Similarly, at $t = 16$, $\mathrm{Var}(\tilde{v}(j,t-1)) = 0.005$ and $t^{1/2}\sigma_F = 0.2$, so $\mathrm{Var}(\tilde{v}(j,t-1))^{1/2}/t^{1/2}\sigma_F < 0.03$. This shows that the effect of the uncertainty in the jump-off population, mortality, and migration is to multiply the widths of prediction intervals for births by a number between 1 and $(1 + 0.03^2)^{1/2} \approx 1.0005$, as compared to intervals that would take into account the uncertainty of fertility alone. This conclusion holds, a fortiori, because the value of σ_S used in

Figure 1 that is appropriate for average mortality is about ten times as large as the value appropriate for the youngest ages. *Obviously, for the purpose of analyzing the uncertainty of births during the first sixteen forecast years it is not necessary to take into account any other factors besides fertility.*

Consider the surviving births from the years $1 \leq t \leq 16$. Or consider ages j such that $\max\{0, t-16\} \leq j \leq t$. By Lemma 1 we know that the variance due to jump-off, mortality, and migration is simply $\sigma_{JO}^2 + t\sigma_M^2 + \sigma_S^2 h(t)$. Since the age group j at forecast year t are survivors of V(0,t−j), the variance due to fertility must be $(t - j)\sigma_F^2$. Or we have that for $\max\{0, t-16\} \leq j \leq t$

$$\mathrm{Var}(\tilde{v}(j,t)) = (t - j)\sigma_F^2 + \sigma_{JO}^2 + \sigma_M^2 + \sigma_S^2 h(t).$$

It is of particular interest to see how the role of fertility that dominates the uncertainty of births, changes as the birth cohorts age. When $t - j = 1$ (births during the first forecast year), the role of fertility is at its minimum. We see from Figure 1 that the square root of the remaining terms does not even reach on half of $\sigma_F = 0.05$ during the first 48 forecast years. For larger values of t the uncertainty of the other factors has even less influence.

4.3. The Second Generation of Births and Their Survival

During forecast years $t = 17,\ldots, 32$ the births generated during the first sixteen forecast years contribute new births. However, no "third generation" births are yet possible. By Lemma 1, the contribution of jump-off population, mortality, and migration to the uncertainty of $\tilde{v}(0,t)$ is $e_{JO} + Q(t)$. The direct contribution of fertility is $e_F(t)$. The indirect contribution (via the size of the child bearing ages) is

$$I(t) = V(0,t)^{-1} \sum_{k=15}^{t-2} B(k,t) e_F(t-1-k).$$

Consider two types of covariances. First, take $1 \leq t \leq 16 < u \leq 32$. We have that

$$\mathrm{Cov}(\tilde{v}(0,t),\ \tilde{v}(0,u)) \approx$$

$$\mathrm{Cov}(e_{JO} + Q(t) + e_F(t),\ e_{JO} + e_F(u) + Q(u) + I(u)) =$$

$$\sigma_{JO}^2 + t\sigma_M^2 + \sigma_S^2(u + t + 1)(t + 1)t/6 + t\sigma_F^2 + \mathrm{Cov}(e_F(t),\ I(u)).$$

The last covariance term can be written as

$$\mathrm{Cov}(e_F(t),\ I(u)) = V(0,t)^{-1} \sum_{k=15}^{u-2} B(k,t)\mathrm{Cov}(e_F(t),\ e_F(u-1-k)),$$

where $\mathrm{Cov}(e_F(t), e_F(u-1-k)) = \sigma_F^2 \min\{t,\ u-1-k\}$.

Second, take $17 \leq t \leq u \leq 32$. Then the covariance formula becomes

$$\mathrm{Cov}(\tilde{v}(0,t),\ \tilde{v}(0,u)) \approx$$

$$\mathrm{Cov}(e_{JO} + Q(t) + e_F(t) + I(t),\ e_{JO} + Q(u) + e_F(u) + I(u)) =$$

$$\sigma_{JO}^2 + t\sigma_M^2 + \sigma_S^2(u + t + 1)(t + 1)t/6 + t\sigma_F^2 +$$

$$\mathrm{Cov}(e_F(t),\ I(u)) + \mathrm{Cov}(e_F(u),\ I(t)) + \mathrm{Cov}(I(t),\ I(u)).$$

Here,

$$\mathrm{Cov}(e_F(t),\ I(u)) = V(0,u)^{-1} \sum_{k=15}^{u-2} B(k,u)\sigma_F^2(u-1-k),$$

$$\mathrm{Cov}(e_F(u),\ I(t)) = V(0,t)^{-1} \sum_{k=15}^{t-2} B(k,t)\sigma_F^2(t-1-k),$$

$$\mathrm{Cov}(I(t),\ I(u)) =$$

$$V(0,t)^{-1}V(0,u)^{-1} \sum_{k=15}^{t-2} \sum_{j=15}^{u-2} B(k,t)B(j,u)\sigma_F^2\min\{t-1-k,\ u-1-j\}.$$

In particular, when $17 \leq t \leq 32$ the variances are

$$\mathrm{Var}(\hat{v}(0,t)) \approx \sigma_{JO}^2 + t\sigma_M^2 + \sigma_S^2(2t + 1)(t + 1)t/6 + t\sigma_F^2 +$$

$$2\cdot V(0,t)^{-1} \sum_{k=15}^{t-2} B(k,t)\sigma_F^2(t-1-k) +$$

$$V(0,t)^{-2} \sum_{k=15}^{t-2} \sum_{j=15}^{t-2} B(k,t)B(j,t)\sigma_F^2\min\{t-1-k,\ t-1-j\}.$$

It is evident that the role of the uncertainty in the jump-off population, mortality, and migration is, if possible, even less important for the second generation of births than for the first generation.

An interesting new issue arises when t = 17,..., 32, however. The direct component of uncertainty that is due to current fertility is $t\sigma_F^2$. The two remaining terms involving σ_F^2 arise because of the indirect component that operates through the size of the child bearing ages. The first one is due to the covariance of the direct and indirect components (= $2\cdot\mathrm{Cov}(e_F(t),\ I(t))$, and the second is the variance of the indirect component (= $\mathrm{Var}(I(T))$). What are the numerical magnitudes of the direct and indirect components? For concreteness, let us assume that the age distribution of the mothers, $B(k,t)/V(0,t)$, k = 15,..., 44, is proportional to the *net-maternity function* (see Section 4.5) of the United States in the year 1985.

The absolute and relative contributions of the three terms are given in Table 1 (in the units σ_F^2). We see that the direct contribution to error variance declines from over 98% during the forecast years t = 17-20 to about 90% at forecast year t = 24 and to about 68% by forecast year t = 32. This means that a prediction interval for t = 32 that would completely ignore the indirect effect would be based on a standard deviation that is $0.68^{1/2} \approx 0.82$ times the correct value. In other words the widths of the intervals would be approximately 18% too narrow. Note that this result does not depend on the exact value of σ_F, as long as its order of magnitude relative to the other sources of error is the one assumed.

When one follows the chort born at jump-off year V(0,0) and the first birth cohort V(0,1), one notices a dramatic increase in the variance of the forecast error that is due to fertility. However, no similar jump occurs when one compares the cohort V(0,16) and V(0,17). Table 1 shows that the increase in variance is mainly determined by $Var(e_F(t))$, and the terms $Cov(e_F(t), I(t))$ and $Var(I(t))$ (that are additional as compared to forecast years t = 1,..., 16) are initially quite small.

Table 1 permits also a ready comparison of the role of fertility versus the other sources of uncertainty among the survivors of the second generation of births. Consider the births during the forecast year t = 25, for example. The standard deviation due to fertility alone (but including both the direct and indirect components) would be $(25\cdot 0.05^2/0.88)^{1/2} \approx 0.267$. The uncertainty of the other three components does not reach one tenth of this during the first 48 forecast years.

4.4. The Third Generation of Births

Consider finally forecast years t = 33,..., 48. The interest centers on the relative magnitudes of the direct and indirect components of fertility to the uncertainty of births. Transparent analytical expressions for the second moments are fairly complicated in this case. However, relatively simple recursive formulas can be derived that permit a numerical evaluation. A program written in (Turbo) C that implements these calculations for forecast years t = 1,..., 48, is given in Appendix I.

The direct contribution of fertility is $e_F(t)$, as always. The indirect contribution is

$$I'(t) = V(0,t)^{-1} \sum_{k=15}^{\min\{44,t-2\}} B(k,t)\eta(t-1-k),$$

where $\eta(t-1-k)$ is the factor of uncertainty in $\tilde{V}(0,t-1-k)$ that derives from fertility alone. When $t - 1 - k \leq 16$, then $\eta(t-1-k) = e_F(t-1-k)$. When, $t - 1 - k > 16$, then $\eta(t-1-k) = e_F(t-1-k) + I(t-1-k)$, where I is as defined in Section 4.3.

In Section 4.3. we have calculated the covariances of the factors $\eta(t-1-k)$ when $t - 1 - k \leq 32$, so $\mathrm{Var}(I'(t))$ can be calculated recursively. The covariance remains to be evaluated. For $t - 1 - k \leq 16$ we have that

$$\mathrm{Cov}(\eta(t-1-k),\ e_F(t)) = \sigma_F^2(t - 1 - k).$$

For $t - 1 - k > 16$ we get

$$\mathrm{Cov}(\eta(t-1-k),\ e_F(t)) = \mathrm{Cov}(e_F(t-1-k) + I(t-1-k),\ e_F(t))$$

$$= \sigma_F^2(t - 1 - k) + \mathrm{Cov}(I(t-1-k),\ e_F(t)),$$

where

$$\mathrm{Cov}(I(t-1-k),\ e_F(t)) = V(0,t-1-k)^{-1} \sum_{j=15}^{t-3-k} B(j,t-1-k)\sigma_F^2(t-1-k).$$

A simplifying assumption to be used below is that the age distribution of the mothers is the same as the distribution of age-specific fertility in the U.S. in 1985. This is a reasonable assumption, unless the point forecast indicates highly nonstationary population growth.

Table 2 presents the relative contributions of the current fertility $e_F(t)$, the indirect component $I'(t)$, and their covariance to the uncertainty in births during forecast years t = 33,..., 48. We see that both $\mathrm{Var}(I'(t))$ and $2\cdot\mathrm{Cov}(e_F(t),\ I'(t))$ increase their relative share of the error. These indirect sources grow faster than the uncertainty of the current fertility, which grows linearly with the forecast year. They exceed the contribution of the current fertility after forecast year t = 41. At the same time the total error due to fertility grows from 0.126 at t = 33 to 0.282 at t = 48. Obviously, in long term forecasting the role of indirect terms cannot be neglected.

4.5. The Impact of Stable Growth

Population growth can impact the propagation of error calculations. In our simplified setup this is visible in the formulas for the second or third generations of birth that depend on the age-distribution of mothers. Since the age-distribution depends, not only on fertility, but also on past birth series, it is clear that population growth can potentially have an impact. However, it is somewhat reassuring to note that in cases of practical importance the impact is likely to be fairly small. The easiest way to see this is to consider stable growth.

Suppose the jump-off population has a stable age-distribution (Keyfitz, 1977, pp. 172-173). Consider a point forecast that specifies no change in fertility and mortality rates over time, and assumes zero net migration by age. It follows that the point forecast leaves the age-distribution of the population unchanged. The total population size (and the size of each age group) grows or declines exponentially, depending on whether the births exceed the deaths, or vice versa. This makes the assessment of the effect of the growth rate relatively simple.

Let p(j) be the probability of surviving to age j, write f(j,t) = f(j), j = 15,..., 44, for short, and let ρ be the growth rate. Then, the stable age-distribution is proportional to $p(j)e^{-\rho j}$, j = 0,..., s, for all t. Therefore, the age-distribution of the mothers is proportional to $f(j)p(j)e^{-\rho j}$, j = 15,..., 44, where f(j)p(j) is the *net maternity function*. Or,

$$B(j,t)/V(0,t) = f(j)p(j)e^{-\rho j}/ \sum_{k=15}^{44} f(k)p(k)e^{-\rho k}.$$

Qualitatively, the effect of growth is straight forward. An increase in ρ increases the importance of low ages. A decrease in ρ increases the importance of high ages. Table 3 has numerical estimates of the magnitudes of these effects for growth rates ranging from $\rho = 0.02$ to -0.02 when the 1985 U.S. net maternity is used for the age-distribution of the mothers. We see that the proportion of variance due to the indirect sources, out of the total contribution to uncertainty by fertility, ranges from 8.1% ($\rho = 0.02$) to 6.0% ($\rho = -0.02$) at forecast year $t = 25$, and from 24.0% ($\rho = 0.02$) to 20.3% ($\rho = -0.02$) at $t = 32$.

4.6. Accuracy of Approximate Formulas

The above calculations provide one framework for assessing the accuracy of the approximate propagation of error calculations proposed by Alho and Spencer (1991). Those formulas account for all sources of error (within the accuracy of Taylor-series approximation) for birth forecasts of years $t = 1,\ldots, 16$. After that various covariance terms are omitted with the effect that once the covariance structure of past births has been calculated, then in the calculation of later births we do not go back to trace the history of a birth cohort beyond their mothers birth.

For example, when considering the variance of births during the forecast years $t = 17,\ldots, 32$, the variance calculation excludes the term $2\cdot\mathrm{Cov}(e_F(t), I(t))$ but includes the term $\mathrm{Var}(e_F(t))$, which represents current fertility, and the term $\mathrm{Var}(I(t))$, which represents the uncertainty of in the number of mothers. The relative magnitudes of these terms can be determined from Table 1. We see that the omitted term represents less than 11% of the variance until forecast year t = 25. After that it increases to 25.3% at forecast year t = 32. The omission of this covariance term means that the width of the prediction interval is $(1 - 0.253)^{1/2} = 0.864$ times the correct value at t = 32. Or the interval is nearly 14% too narrow. Table 3 implies that positive population growth tends to increase the error slightly. Population decline has the opposite effect.

Table 2 shows the development during forecast years t = 33,..., 48. The approximate formulas exclude the covariance terms P_C. We see that by the year t = 48, the variance of the forecast error will be underestimated by 40%. This implies that prediction intervals based on the approximate formulas will be $(1 - 0.397)^{1/2} = 0.78$ times the width of the correct intervals. Or the intervals will be 22% too narrow.

One way to make the approximate formulas more accurate would be to apply proportional adjustment factors based on calculations such as these. Alternatively, it is possible to program the exact formulas for forecast years t > 16. Alho and Spencer (1991) provide exact formulas for t = 17,..., 32.

5. Numerical Evaluation of Actual Forecasts

The simple analytical model given above probably captures the broad features of realistic error propagation. However, it obviously does not utilize the best possible time-series model for forecasting the vital rates. For example, once we have differenced the (logs of the) fertility rates, it is natural to investigate the autocorrelation structure of the differenced series and to attempt to identify an ARMA model for it, rather than to simply assume that the differenced series is an uncorrelated stationary sequence. More generally, we may model the various time-series jointly using vector ARMA techniques. One would expect that the utilization of such techniques would decrease the estimate of error variance for short term forecasts, but one would not expect large differences as compared to the simpler model in long term forecasting, when the advantage provided by the autocorrelatedness has worn off.

The concrete models considered here are those of Bell and Monsell (1991). We first describe the fertility and mortality models used and then present the numerical results.

5.1. Fertility Model

Define the *total fertility rate* of the year t as

$$W(t) = R(0,15,t) +\ldots+ R(0,44,t),$$

with $w(t) = \log(W(t))$, and let the *fraction of fertility* due to age j during year t be

$$Q(j,t) = R(0,j,t)/W(t), \quad j = 15,\ldots, 44,$$

with $q(j,t) = \log(Q(j,t))$. Then, $R(0,j,t) = Q(j,t)W(t)$, and

$$f(j,t) = q(j,t) + w(t).$$

Bell and Monsell analyzed the processes $q(j,t)$ using principal components methods (actually they used ages 14-46, but we will combine the contributions of ages 14 and 15 to age 15, and the contributions of ages 44-46 to age 44). It was found that four principal components and the total fertility rate provided a fairly good representation for the time-series.

In other words, writing $U = (u_1, \ldots, u_4)$, where u_j's are the eigenvectors, $\beta_t = (\beta_{1t}, \ldots, \beta_{4t})^T$ for the coordinates (or the pricipal components), and $q(t)$ is the vector of $q(j,t)$'s, they considered the representations

$$q(t) = U\beta_t + e_t,$$

where e_t's are independent error vectors with $Cov(e_t) = \Sigma_e$. Defining $\mathbf{U} =[\mathbf{1}, U]$, where $\mathbf{1}$ is a vector of all ones, $\beta_t = (w(t), \beta_t^T)^T$, and $f(t)$ a vector of $f(j,t)$'s, we can write

$$f(t) = \mathbf{U}\beta_t + e_t$$

for the vector of the log-fertility rates.

The components of the coordinate vector β_t were jointly modeled. It was found that differencing the vector once was necessary to reduce the vector series to stationarity. Then, a vector ARMA model was fitted to the differenced vector $\nabla\beta_t$. It turned out that a reasonable approximation to the (slightly more complex) best fitting model was obtained by an AR(1) model

$$\nabla\beta_t = \Phi\nabla\beta_{t-1} + \varepsilon_t,$$

where Φ is a *diagonal* coefficient matrix, Φ = diag(0.673, 0.383, 0.716, 0.598, 0.707), and ε_t's are independent with $\mathrm{Cov}(\varepsilon_t) = \Sigma_\varepsilon$. In other words, each of the component series was modelled by a univariate AR(1) model with the parameter given on the diagonal of Φ. This Markovian model is particularly simple from the point of view of propagation of error calculations.

Assuming that we have observations for $t \leq 0$, then the point forecast for the year $t > 0$ is simply $\nabla\beta_t = \Phi^t\nabla\beta_0$. The covariance of the forecast error is

$$\mathrm{Cov}(\nabla\beta_t - \nabla\beta_t) = \sum_{k=0}^{t-1} \Phi^k\Sigma_\varepsilon\Phi^k.$$

More generally, the error covariance for the coordinates β_t and β_{t+k}, $k \geq 0$, is of the form

$$\mathrm{Cov}((\beta_t - \beta_t)(\beta_{t+k} - \beta_{t+k})^T) = \sum_{j=1}^{t} (\sum_{h=0}^{t-j} \Phi^h)\Sigma_\varepsilon(\sum_{h=0}^{t+k-j} \Phi^h)$$

$$\equiv C_\beta(t,t+k).$$

It follows that in analogy with Bell and Monsell (1991, p. 15) the forecast error $\tilde{f}(t) - f(t) \equiv e_f(t)$ has the covariance

$$\mathrm{Cov}(e_f(t),\ e_f(t+k)) = \mathbf{U}C_\beta(t,t+k)\mathbf{U}^T + \delta(k)\Sigma_e,$$

where $k \geq 0$, and $\delta(k) = 1$ for $k = 0$, and $\delta(k) = 0$ for $k > 0$.

5.2. Mortality Model

Bell and Monsell (1991) used principal components to model mortality much the same way they analyzed fertility. In the case of mortality the ages 0, 1, 2, 3, 4, 5-9, 10-14,..., 80-84, 85+ were considered. Define the log-mortality rate of age j during year t as $y(j,t) = \log(-r(j,t))$. Let $y(t)$ be the vector of the $y(j,t)$'s for $j = 0, 1, \ldots, s$, where s is the highest age to be considered ($s \geq 85$). Let Z be the matrix that has the five most important eigenvectors of the sums of squares and products matrix of the grouped mortality data, expanded in such a way that the columns of Z have one component for each age $j = 0, 1, \ldots, s$. (In other words, rows were added to Z so that, for example, the rows 6-10 of Z are equal to the component corresponding to the age-group 5-9.) Let β_t be the vector of coordinates in the representation of $y(t)$ in terms of U for the year t. Then, the model considered can be written as $y(t) = Z\beta_t + e_t$, where the e_t's are independent error vectors with a covariance matrix estimated from the data. Note that since $y(t)$ has component for each age, some of the components, such as the components 6-10, are perfectly correlated. Again, a reasonable approximation to the best fitting model was obtained by taking an AR(1) model for the first differences of each of the coordinates. Or, we have

$\nabla\beta_t = \Phi\nabla\beta_{t-1} + \varepsilon_t$, with $\Phi = \text{diag}(0.649, 0.331, -0.106, -0.140, -0.183)$. It follows that the forecast error for y(t) has the same form as the forecast error for f(t) in Section 5.1. (with $\mathbf{U}$ replaced by Z).

Despite the many similarities, the analysis of mortality is in one respect more complicated than that of fertility. The use of the log-transform had the effect that fertility rates were transformed precisely to the form in which they are needed, say, in the formulas of page 5 above (the f(k,t-1)'s). However, the use of the log-transform for mortality means that the sums of r(j,t)'s on page 5 become nonlinear functions of the forecast values. It follows that the propagation of error calculations for v(j,t)'s with even j > t (these are the survivors from the jump-off population) requires the use of a Taylor-series expansion.

From the point of view of the propagation of error calculations it would be much preferable to analyze the mortality rates as such, despite the fact that this may lead to negative estimates in the very youngest or the very oldest ages. These ages are not important in the analysis of the accuracy of the forecasts. One reason for preferring log-mortality rates is that it may permit simpler forecasting models, an issue we have not considered in this paper at all.

5.3. Results

5.3.1. Survival from Jump-Off

Appendix II contains a program written in (Turbo) C that was used to calculate the relative forecast error for forecast years t = 1,..., 48. The uncertainty of survival was the only component included in the calculation. We will comment on the effect of the uncertainty of the jump-off population and migration at the end.

Figure 2 has a plot of the standard deviation of the relative error for the age-groups that were in ages 0, 10,..., 40 at jump-off, when they are survived 48 years. For comparison purposes it also includes the corresponding (solid) curve $\sigma_S h(t)^{1/2}$ for the simple analytical model that was displayed in Figure 1 already. As in Figure 1, a standard deviation of, say, 0.05 means that with probability 95% the forecast error is not more than ±10% of the point forecast of the number of survivors.

Figure 2 indicates that the estimates of uncertainty increase with the increasing age at jump-off. The reason is that the uncertainty of survival is approximately proportional to the level of mortality. Therefore, in younger ages (say, 1-50) much less uncertainty accrues over the forecast years than in older ages.

We see also that the estimates are, in many ages, initially smaller than those based on the simple analytical model with σ_S = 0.0001. However, the older age-group depicted, i.e., those who survive from ages 20, 30, and 40 at jump-off to ages 68, 78, and 88 at forecast year t = 48, have forecast errors that become much larger than one would anticipate based on Figure 1. The difference can be large. The cohort of 40 year olds at jump-off have a relative error of 0.000048 at t =

1 but it increases to 0.299 at t = 48, whereas the respective values expected based on the analytical model are 0.0001 and 0.0195. (For the cohort of 50 year olds at jump-off the relative error reaches the value 0.56 at t = 48.)

In view of the discussion of Figure 1, differences of this magnitude are not surprising, when one recalls that mortality in, say, ages 70-74 is 100 times higher than mortality in ages 10-14. Hence, the level of uncertainty can vary much more from age to age than one would expect based on the analytical model. One should also recall that the solid curve depicted in Figure 2 is based on a relatively low level of error (σ_S = 0.0001) compared to observed data. Doubling the level of uncertainty might be quite reasonable. This would, nevertheless, leave the conclusions above qualitatively the same.

A comparison between the shapes of the SD curves for the cohorts and the one based on the analytical formula is also interesting. Asymptotically, $h(t)^{1/2}$ behaves like $t^{3/2}$ so the curve increases faster than linearly, a fact that can easily be seen in Figure 1. The curves for the most cohorts start out slower than the analytical formula, but increase faster, when the cohorts advance to older ages with sharply higher levels of uncertainty (cf. Figure 4 of Alho and Spencer, 1990, p. 615). However, for the very youngest cohort (age = 0 at jump-off) the uncertainty is initially even higher than that predicted by the analytical formula, but by the tenth forecast year it falls below all other curves. This is an anomalous case.

We conclude by two remarks relating to the other sources of uncertainty. First, consider the effect of the uncertainty in the jump-off population and migration that was not explicitly considered above. Assume, as we did in our analytical models, that in relative terms those effects are the same in all ages. Recall that the relative error of migration reached approximately the level SD = 0.003 at t = 30 and the jump-off had SD = 0.004. These error levels would dominate many cohorts through the 20-30 first forecast years. It is useful to note that this conclusion holds, *a fortiori*, if one considers how the assumption of age-independence should be relaxed. Presumably, the uncertainty of the jump-off population is the greatest in those ages in which the undercount is the largest. These would typically be ages with very low mortality, such as ages 20-30. Similarly, the uncertainty of migration would be expected to be the largest in ages in which the volume of migration is the greatest. These would be the early working years (again 20-30) and the early retirement ages (60-70). At least the first of these uncertainty "humps" is located in an age with low mortality. Both factors would increase the uncertainty due to migration and jump-off relative to the uncertainty of survival.

Secondly, we have seen that fertility overwhelmingly dominates the uncertainty of future birth cohorts under the analytical model. We expect it to do the same, when the level of uncertainty in survival in the youngest ages is estimated to be much less.

5.3.2. Births

Figure 3 has a plot of the standard deviation of the error in log-births that is caused by the uncertainty of period fertility for forecast years t = 1,..., 32. In other words, this the direct component of uncertainty that does not include the uncertainty in the numbers of women in child bearing ages. The solid curve has the level of error calculated based on the principal components regression approach of Bell and Monsell, the dashed line has the curve $\sigma_F t^{1/2}$ for $\sigma_F = 0.05$ and the dotted line has the same curve for $\sigma_F = 0.10$. These are the parameter values considered reasonable for the analytical model.

We see that the principal components model gives qualitatively very similar results as the analytical model. The major difference is that the error expected based on the principal components model increases initially much faster than the two analytical curves. Apparently, this is due to the fact that the analytical models assume independent increments for the error, whereas the AR(1) model estimated for the principal components assumes positively correlated increments.

As anticipated based on the analytical models, it is clear that for the purpose of analyzing the uncertainty of births, only the uncertainty in birth rates (both in its direct and indirect forms) needs to be considered. For example, at forecast year t = 1 the relative error in the pricipal component approach is 0.03, a value that is reached by the highest error curve of Figure 2 at t = 27. The contribution of survival and migration in younger ages is quite insignificant.

The second issue to be studied was the indirect contribution of fertility through "births to births" during forecast years t = 17,..., 32. As shown in Table 1, this contribution exceeded 30% at t = 32 in the analytical model. As Table 4 demonstrates, the indirect contribution appears to be somewhat smaller in the principal components approach. For example, at t = 32 the indirect sources contribute approximately 25% of the variance. One might conjecture that the difference is due to the more highly nonstationary model used in the principal components approach.

Figure 4 has a plot of the relative error under the principal components model (solid curve) and the analytical model (dashed curve). We see again that the level of error is higher in the former model. Note also how both curves start to increase faster after t = 17, when the indirect component of uncertainty due to "births to births" starts to gain importance.

6. Discussion

We have analyzed several aspects of how different vital processes influence the error of population forecasts. The most important observation has been that the uncertainty of fertility is so high compared to the uncertainty of other vital processes that for the purpose of analyzing the uncertainty of birth forecasts the other sources can be completely ignored. The conclusion was the same in both the analytical model and the forecasting model based on principal components.

Secondly, it turned out that for the purpose of analyzing the survivors of the jump-off population, the uncertainty of the jump-off population and that of migration are equally important as the uncertainty of mortality for forecast periods of about 15 years. These sources of error have been studied much less than fertility and mortality, presumably because of data problems. However, it is clear that in short term forecasting they cannot be ignored.

Thirdly, the simple analytical model appeared to give qualitatively the correct picture of the order of magnitude of various sources of error. Nevertheless, it could not (as it was not designed to) capture the variation in the forecast error among the survivors of the jump-off population by age. These variations depended on the fact that mortality varies by orders of magnitude in different ages and, hence, also the level of error in mortality forecasts.

Finally, we considered the accuracy of the approximate propagation of error formulas of Alho and Spencer (1991). The fact that the uncertainty of fertility dominated the other sources means that the approximations saying that we can consider the covariance of childrens' survival and their mothers' survival prior to the childrens' birth zero seems well justified. On the other hand, the assumption that current fertility is uncorrelated with the fertility that produced the current mothers becomes increasingly unacceptable when the forecast period exceeds, say, 25-30 years. Fortunately, that covariance is relatively easy to program (at least when one compares it to the programming of the covariances between the mortalities), so the idea of Alho and Spencer (1991) of implementing a stochastic forecast as a database, which uses as one input a previously calculated covariance matrix of the errors of the birth forecasts, appears quite feasible based on these investigations.

References

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. (1985) "Uncertain Population Forecasting", *Journal of the American Statistical Association*, 80, 306-314.

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

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

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

Bell, W.R. and Monsell, B. (1991) "Using Principal Components in Time Series Modeling and Forecasting of Age-Specific Mortality Rates", Paper presented at the Annual Meeting of the Population Association of America in Washington, D.C., in March 1991.

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

Carter, L. and Lee, R.D. (1986) "Joint Forecasts of U.S. Marital Fertility, Nuptiality, Births, and Marriages Using Time-Series Models", *Journal of the American Statistical Association*, 81, 902-911.

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

Keyfitz, N. (1977) *Introduction to the Mathematics of Population, with Revisions*, 2nd ed., Addison-Wesley: Reading, MA.

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

Lee, R.D. and Carter, L. (1991) "Modeling and Forecasting U.S. Mortality", *Journal of the American Statistical Association*, to appear.

Lee, R.D. and Tuljapurkar, S. (1991) "Stochastic Population Projections for the U.S.: Beyond High, Medium, and Low", Paper presented at the Annual Meeting of the Population Association of America in Washington, D.C., in March 1991.

McDonald, J. (1979) "A Time-Series Approach to Forecasting Australian Total Live Births", *Demography*, 16, 575-601.

McDonald, J. (1981) "Modeling Demographic Relationships: An Analysis of Forecast Functions for Australian Births", *Journal of the American Statistical Association*, 76, 782-792.

Mulry, M. and Spencer, B. (1991) "Total Error in PES Estimates of Population", *Journal of the American Statistical Association*, 86, 839-855.

Smith, S. (1987) "Test of forecast Accuracy and Bias for County Population Projections", *Journal of the American Statistical Association*, 82, 991-1003.

Thompson, P.A., Bell, W.R., Long, J.F., and Miller, R.B. (1989) "Multivariate Time-Series Projections of Parametrized Age-Specific Fertility Rates", *Journal of the American Statistical Association*, 84, 689-699.

Table 1. The values of the direct component of uncertainty in births t = $\mathrm{Var}(e_F(t))\sigma_F^{-2}$, the covariance of the direct component and the indirect component C = $2\cdot\mathrm{Cov}(e_F(t),\ I(t))\sigma_F^{-2}$, and the variance of the indirect component V = $\mathrm{Var}(I(t))\sigma_F^{-2}$ (all in the units σ_F^2) for t = 17,..., 32.

t	%	C	%	V	%
17	99.8	0.03	0.2	0.00	0.0
18	99.5	0.09	0.5	0.00	0.0
19	98.9	0.20	1.1	0.00	0.0
20	98.0	0.40	2.0	0.01	0.1
21	96.7	0.69	3.2	0.04	0.2
22	95.0	1.09	4.7	0.08	0.3
23	92.9	1.61	6.5	0.14	0.6
24	90.6	2.24	8.5	0.24	0.9
25	88.0	3.00	10.6	0.39	1.4
26	85.3	3.89	12.8	0.59	1.9
27	82.4	4.90	15.0	0.84	2.6
28	79.5	6.05	17.2	1.17	3.3
29	76.6	7.31	19.3	1.57	4.1
30	73.7	8.69	21.3	2.04	5.0
31	70.8	10.17	23.2	2.59	5.9
32	68.1	11.76	25.3	3.22	6.8

Table 2. The proportions of the covariance of the direct and indirect components of uncertainty $2\cdot \mathrm{Cov}(e_F(t), I'(t))$ (= P_C) and of the variance of the indirect component $\mathrm{Var}(I'(t))$ (= P_V) in percent out of the total uncertainty caused by fertility for t = 33,..., 48.

t	P_C	P_V
33	26.7	7.8
34	28.1	8.7
35	29.5	9.5
36	30.8	10.4
37	31.9	11.2
38	32.9	11.9
39	33.9	12.6
40	34.7	13.2
41	35.5	13.8
42	36.3	14.4
43	37.0	14.9
44	37.6	15.5
45	38.2	16.0
46	38.8	16.5
47	39.3	17.0
48	39.8	17.5

Table 3. The proportions of the covariance of the direct and indirect components of uncertainty $2\cdot\mathrm{Cov}(e_F(t), I(t))$ (= P_C) and of the variance of the indirect component $\mathrm{Var}(I(t))$ (= P_V) in percent out of the total uncertainty caused by fertility for t = 17, 25, 32, when the rate of stable growth is ρ = 0.02, 0.01. 0.00, -0.01, -0.02.

	t = 17		t = 25		t = 32	
ρ	P_C	P_V	P_C	P_V	P_C	P_V
0.02	0.2	0.0	11.9	1.7	26.3	7.6
0.01	0.2	0.0	11.2	1.5	25.7	7.2
0.00	0.2	0.0	10.6	1.4	25.0	6.8
-0.01	0.1	0.0	9.9	1.2	24.3	6.5
-0.02	0.1	0.0	9.3	1.1	23.6	6.1

Table 4. The relative contributions (%) of the uncertainty of past births (= V) and twice the covariance between the past births and current fertility (= C) to the variance of the log-births during forecast years t = 17,..., 32.

t	V	C
17	0.0	0.0
18	0.0	0.1
19	0.0	0.3
20	0.0	0.7
21	0.0	1.3
22	0.1	2.2
23	0.2	3.3
24	0.2	4.7
25	0.5	6.3
26	0.8	8.1
27	1.1	10.1
28	1.6	12.2
29	2.1	14.3
30	2.9	16.5
31	3.6	18.6
32	4.5	20.7

APPENDIX I

```
/* This program calculates the covariances of births under
the simple propagation of error model described in
Alho: The contribution of different sources of error
to the accuracy of population forecasts
*/

#include <stdio.h>
#include <alloc.h>
#include <math.h>
#define C(x,y) *(c+((x-1)*48)+y-1)
#define A(x,y) *(a+((x-1)*48)+y-1)
#define B(x,y) *(b+((x-1)*48)+y-1)
#define H(x,y) *(h+((x-1)*48)+y-1)

 float co1(int i1, int i2, float s[], float s1);
 float co2(int i1, int i2, float s[], float s1);
 float co3(int i1, int i2, float s[], float s1);
 float co4(int i1, int i2, float s[]);
 float co5(int i1, int i2, float s[]);

float *a, *b, *c, *h;

main()
{
 float sjo, sm, ss, sf, r1, dd[30], *d, wu, wt, ros;
 int t, u, k, j, m;
 FILE *fp1, *fp2;
 d=dd-1;
 sjo=pow(0.00325,2);
 sm=pow(0.001,2);
 ss=pow(0.0005,2);
 sf=pow(0.05,2);

/*allocate space for the matrices:
  C has the full covariances,
  A has the total contribution of fertility,
  B has the covariance terms omitted in the approxiamte
formulas,
  H has the direct contribution of current fertility */

if ((c = (float *) malloc(sizeof(float)*48*48)) == NULL)
printf("c allocate error\n");
if ((a = (float *) malloc(sizeof(float)*48*48)) == NULL)
printf("a allocate error\n");
if ((b = (float *) malloc(sizeof(float)*48*48)) == NULL)
printf("a allocate error\n");
if ((h = (float *) malloc(sizeof(float)*48*48)) == NULL)
printf("h allocate error\n");

/* open the output file */
 fp2=fopen("tulos", "w");

/* read the distribution of net maternity */
 fp1=fopen("dist.dat","r");
 for(u=1; u <= 30; u++)
   {
    fscanf(fp1,"%f", &r1);
    d[u]=r1;
   }
 fclose(fp1);

/* calculate covariances */
  for(u=1; u <= 48; u++)
    {
     for(t=1; t <= u; t++)
       {
        C(t,u) = t*sf+sjo+t*sm+(u+t+1)*(t+1)*t*ss/6.0;
        A(t,u) = t*sf;
        B(t,u) = 0.0;
        H(t,u) = t*sf;
        if (u > 16 && u <= 32)
          {
           ros = co1(t, u, d, sf);
           A(t,u) += ros;
           C(t,u) += ros;
           B(t,u) += ros;
           if (t > 16)
             {
              ros = co1(u, t, d, sf);
              A(t,u) += ros;
              C(t,u) += ros;
              B(t,u) += ros;
              ros = co2(t, u, d, sf);
              A(t,u) += ros;
              C(t,u) += ros;
             }
          }
        if (u > 32)
          {
           if (t <= 32)
             {
              ros = co3(t, u, d, sf);
              C(t,u) += ros;
              A(t,u) += ros;
              B(t,u) += ros;
              if (t > 16)
                {
                 ros = co1(u, t, d, sf);
                 C(t,u) += ros;
                 A(t,u) += ros;
                 ros = co4(t, u, d);
                 C(t,u) += ros;
                 A(t,u) += ros;
                 B(t,u) += ros;
                }
             }
           if (t > 32)
             {
              ros = co3(t, u, d, sf);
              ros += co3(u, t, d, sf);
             B(t,u) += ros;
             ros += co5(t, u, d);
             A(t,u) += ros;
             C(t,u) += ros;
            }
         }
      } /* end of t */

     fprintf(fp2, "%6f ", C(u,u));
     fprintf(fp2, "%6f ", A(u,u));
     fprintf(fp2, "%6f ", H(u,u));
     fprintf(fp2, "%6f ", B(u,u));
     fprintf(fp2, "%6f ", H(u,u)/A(u,u));
     fprintf(fp2, "%6f ", B(u,u)/A(u,u));
     fprintf(fp2, "%6f\n", 1-(H(u,u)+B(u,u))/A(u,u));
    }   /* end of u */
}     /* end of main */

/* co1 calculates cov(ef(i1), I(i2)) for 16 < i2 <= 48
   and i1 <= 48, at least */
float co1(int i1, int i2, float s[], float s1)
{
 int k1, w1, m1;
 float sto;
 sto=0;
 w1 = (44 < i2-2) ? 44 : i2-2;
 for (k1=15; k1 <= w1; k1++)
   {
    m1 = (i1 < i2-1-k1) ? i1 : i2-1-k1;
    sto += s[k1-14]*m1*s1;
   }
 return sto;
}

/* co2 calculates cov(I(i1), I(i2)) for 16 < i1, i2 <= 32, at
least */
float co2(int i1, int i2, float s[], float s1)
{
 int k1, k2, w1, w2, m1;
 float sto;
 sto=0;
 w1 = (44 < i1-2) ? 44 : i1-2;
 w2 = (44 < i2-2) ? 44 : i2-2;
 for (k1=15; k1 <= w1; k1++)
   for (k2=15; k2 <= w2; k2++)
     {
      m1 = (i1-1-k1 < i2-1-k2) ? i1-1-k1 : i2-1-k2;
      sto += s[k1-14]*s[k2-14]*m1*s1;
     }
 return sto;
}

/* co3 calculates cov(ef(i1), I'(i2)) for i1 <= 48 and i2 >
32, at least */
float co3(int i1, int i2, float s[], float s1)
{
 int k1, w1, m1;
 float sto, sto2;
 sto=0;
 w1 = (44 < i2-2) ? 44 : i2-2;
 for (k1=15; k1 <= w1; k1++)
   {

    m1 = (i1 < i2-1-k1) ? i1 : i2-1-k1;
    sto2 = (i2-1-k1 <= 16) ? m1*s1 : m1*s1 + co1(i1, i2-1-k1, s, s1);
    sto += s[k1-14]*sto2;
   }
 return sto;
}

/* co4 calculates cov(I(i1), I'(i2)) for 16 < i1 <= 32 and i2 > 32,
at least */
float co4(int i1, int i2, float s[])
{
 int k1, k2, w1, w2, m1, m2;
 float sto;
 sto=0;
 w2 = (44 < i2-2) ? 44 : i2-2;
 for (k2=1; k2 <= w2; k2++)
   {
    w1 = (44 < i1-2) ? 44 : i1-2;
    for (k1=15; k1 <= w1; k1++)
      {
       if (i2-1-k2 < i1-1-k1)
         {
          m1 = i2-1-k2;
          m2 = i1-1-k1;
         }
       else
         {
          m1 = i1-1-k1;
          m2 = i1-1-k2;
         }
       sto += s[k1-14]*s[k2-14]*A(m1,m2);
      }
   }
 return sto;
}

/* co5 calculates cov(I'(i1), I'(i2)) for 32 < i1, i2 <= 48, at least
*/
float co5(int i1, int i2, float s[])
{
 int k1, k2, w1, w2, m1, m2;
 float sto;
 sto = 0;
 w1 = (44 < i1-2) ? 44 : i1-2;
 w2 = (44 < i2-2) ? 44 : i2-2;
 for (k1=15; k1 <= w1; k1++)
   for (k2=15; k2 <= w2; k2++)
     {
      if (i2-1-k2 < i1-1-k1)
```

```
        {
         m1 = i2-1-k2;
         m2 = i1-1-k2;
        }
      else
        {
         m1 = i1-1-k1;
         m2 = i2-1-k2;
        }
      sto += s[k1-14]*s[k2-14]*A(m1,m2);
     }
 return sto;
}
```

Appendix II

```
/* Compile with memory model 'large' */

/* This program calculates a point forecast of the survivors
of the jump-off population and a point forecast of all
mortality rates, both by single years of age.
After that it calculates the variance of the forecast error
for the survivors of the jump-off population.
     Then, a point forecast for all fertility rates is
calculated. These are used to obtain a forecast of female births.
Mortality forecasts are used to survive the births through
the forecast period. After that, a covariance matrix of
the births is calculated. */

#include <stdio.h>
#include <alloc.h>
#include <math.h>

#define P(x,y)  *(p+(x)*49+y)
#define M(x,y)  *(m+(x)*49+y)
#define U(x,y)  *(u+(x)*5+y-1)
#define C1(x,y)  *(c1+(x-1)*5+y-1)
#define C2(x,y)  *(c2+(x)*101+y)
#define PSI(x,y)  *(psi+(x)*5+y-1)
#define VAR(x,y)  *(var+(x)*48+y-1)

#define B(x,y)  *(b+(x-1)*49+y-1)
#define F(x,y)  *(f+(x)*49+y-1)
#define C3(x,y)  *(c3+(x-1)*49+y-1)
#define U1(x,y)  *(u1+(x)*5+y-1)
#define PSI1(x,y)  *(psi1+(x)*5+y-1)
#define C4(x,y)  *(c4+(x-1)*6+y-1)
#define C5(x,y)  *(c5+(x+1)*34+y+1)
#define V(x,y)  *(v+(x+1)*5+y+1)

float co1(int i1, int i2, int t2, int t3);
float co2(int i3, int t4, int t5);

float *p, *m, *u, *c1, *c2, *c3, *c4, *c5, *psi, *var, *v;
float *b, *f, *c3, *u1, *psi1;

float beta[] = {-30.0134, -0.74581, -0.13023, 0.11293, 0.08089};
float dif[] = {-0.029049, 0.006372, 0.015485, 0.037090, 0.021413};
float phi[] = {0.649, 0.331, -0.106, -0.140, -0.183};

float beta1[] ={0.536084, -26.6611, -2.3922, 0.4342, -0.6754};
float dif1[] = {0.00480884, -0.146975, -0.117353, -0.063485,
-0.052468};
float phi1[] = {0.673, 0.383, 0.716, 0.598, 0.707};

main()
{
 float r1, mort, fert, sto[22];
 int i, j, k, h, t, per, j1, t1, k1, h1;
 FILE *fp1, *fp2, *fp3, *fp4, *fp5, *fp6, *fp7, *fp8, *fp9;
 FILE *fp10, *fp11, *fp12, *fp13;

/* will the mortality analysis be done?
   set mort = 1, if yes, otherwise set mort = 0 */
   mort = 0;

/* will the fertility analysis be done?
   set fert = 1, if yes, otherwise set fert = 0 */
   fert = 1;

 /* allocate space for the matrices:
    P  has the forecast of the survivors from the jump-off pop.
    in columns 1,2,...and the jump-off pop. in column 0;
    M  has the forecasts of the ASMR's for each age (cols. 1,2,...)
    and the last observed vector in column 0;
    U  has the first five eigenvectors for log-mortality;
    C1  has the covariance matrix of the first differences of the
    first five coordinates of log-mortality;
    C2  has the covariance matrix of the error due to lack of fit
    caused by not using all eigenvectors;
    PSI  has the diagonal elements of the powers of the phi-
    matrices for mortality;
    VAR  has the variances of the prediction errors for the log of
    the number of the survivors from the jump off population;
    F has the point forecasts of log(ASFR)'s for ages 15,...,44;
    B  has the forecasts of births by age of mother;
    C3  has the covariance matrix of the log-births;
    U1  has the eigenvector matrix for log(TFR) and log of
    fertility distributions;
    PSI1  has the sums of the diagonal elements of the powers of
    phi-matrices for fertility;
    C4  has the covariance matrix of the first differences of the
    first five coordinates of log-fertility;
    C5  has the covariance matrix of the error due to lack of fit
    caused by not using all eigenvectors;
    V  has some results stored in it;
 */

 if ((p = (float *) malloc(sizeof(float)*101*49)) == NULL)
 printf("p allocate error\n");
 if ((m = (float *) malloc(sizeof(float)*101*49)) == NULL)
 printf("m allocate error\n");
 if ((u = (float *) malloc(sizeof(float)*101*6)) == NULL)
 printf("u allocate error\n");
 if ((c1 = (float *) malloc(sizeof(float)*6*6)) == NULL)
 printf("c1 allocate error\n");
 if ((c2 = (float *) malloc(sizeof(float)*101*101)) == NULL)
 printf("c2 allocate error\n");
 if ((psi = (float *) malloc(sizeof(float)*6*49)) == NULL)
 printf("psi allocate error\n");
 if ((var = (float *) malloc(sizeof(float)*101*49)) == NULL)
 printf("var allocate error\n");

if((f = (float *) malloc(sizeof(float)*34*49)) == NULL)
printf("f allocate error\n");
if((b = (float *) malloc(sizeof(float)*31*49)) == NULL)
```

```
printf("b allocate error\n");
if((c3 = (float *) malloc(sizeof(float)*49*49)) == NULL)
printf("c3 allocate error\n");
if((u1 = (float *) malloc(sizeof(float)*6*34)) == NULL)
printf("u1 allocate error\n");
if((psi1 = (float *) malloc(sizeof(float)*6*49)) == NULL)
printf("psi1 allocate error\n");
if((c4 = (float *) malloc(sizeof(float)*6*6)) == NULL)
printf("c4 allocate error\n");
if((c5 = (float *) malloc(sizeof(float)*34*34)) == NULL)
printf("c5 allocate error\n");
if((v = (float *) malloc(sizeof(float)*6*34)) == NULL)
printf("v allocate error\n");

 /* open the output files */
 fp1=fopen("spop", "w");
 fp2=fopen("asmrs", "w");
 fp8=fopen("sd", "w");
 fp9=fopen("cohort", "w");

 /* read in the jump-off population;
    produces P(x,0), x = 0,...,100 */
 fp3=fopen("wfpop88", "r");
 for(i=0; i <= 100; i++)
    fscanf(fp3,"%f", &P(i,0));
 fclose(fp3);

 /* read in the last observed values of the asmrs
    and expand them to all ages;
    produces M(x,0), x = 0,...,100 */
 fp4=fopen("mrate", "r");
 for(i=0; i <= 4; i++)
   fscanf(fp4,"%g", &M(i,0));
 for(i=5; i <= 20; ++i)
   {
    fscanf(fp4,"%g", &r1);
    for(k=0; k <= 4; ++k)
      M((i-4)*5+k,0) = r1;
   }
 fscanf(fp4,"%g", &r1);
 for(k=0; k <= 15; k++)
   M(85+k,0) = r1;
 fclose(fp4);

/* read in the eigenvectors and expand them to all ages;
   produces U(x,y), x = 0,...,100, y = 1,...,5 */
 fp5=fopen("morteig1.dat", "r");
 for(i=0; i <= 4; i++)
    for(j=1; j <= 5; j++)
      fscanf(fp5,"%g", &U(i,j));
 for(i=5; i <= 20; i++)
   for(j=1; j <= 5; j++)
     {
      fscanf(fp5,"%g", &r1);
      for(k=0; k <= 4; k++)
     U((i-4)*5+k,j) = r1;
     }
 for(j=1; j <= 5; j++)
   {
    fscanf(fp5,"%g", &r1);
    for(k=0; k <= 15; k++)
      U(85+k,j) = r1;
   }
 fclose(fp5);

/* read in the covariance matrix of largest koordinates;
   produces C1(x,y), x = 1,...,5, y = 1,...,5 */
 fp6=fopen("mortmlt.cov", "r");
 for(i=1; i <= 5; i++)
   for(j=1; j <= 5; j++)
    fscanf(fp6,"%g", &C1(i,j));
 fclose(fp6);

/* read in and expand the covariance of the error caused by lack of
fit;
   produces C2(x,y), x = 0,...,100, y = 0,...,100 */
 fp7=fopen("mortsig2.dat", "r");
 for(i=0; i <= 21; i++)
   {
    for(j=0; j <= 21; j++)
      fscanf(fp7, "%g", &sto[j]);
    if (i < 5)
      {
       for(j=0; j <= 4; j++)
      C2(i,j) = sto[j];
       for(j=5; j <= 20; j++)
      for(k=0; k <= 4; k++)
        C2(i,(j-4)*5+k) = sto[j];
       for(k=0; k <= 15; k++)
      C2(i,85+k) = sto[21];
      }
    else if (i < 21)
      for(h=0; h <= 4; h++)
     {
      for(j=0; j <= 4; j++)
        C2((i-4)*5+h,j) = sto[j];
      for(j=5; j <= 20; j++)
        for(k=0; k <= 4; k++)
          C2((i-4)*5+h,(j-4)*5+k) = sto[j];
      for(k=0; k <= 15; k++)
        C2((i-4)*5+h,85+k) = sto[21];
     }
    else
      for(h=0; h <= 15; h++)
     {
      for(j=0; j <= 4; j++)
        C2(85+h,j) = sto[j];
      for(j=5; j <= 20; j++)
        for(k=0; k <= 4; k++)
          C2(85+h,(j-4)*5+k) = sto[j];
      for(k=0; k <= 15; k++)
        C2(85+h,85+k) = sto[21];
     }
   }
 fclose(fp7);

/* calculate the PSI matrix for mortality;
   produces PSI(x,y), x = 0,...,48, y = 1,...,5 */
for(j=1; j <= 5; j++)
  PSI(0,j) = 1.0;
for(i=1; i <= 48; i++)
  for(j=1; j <= 5; j++)
    PSI(i,j) = PSI(i-1,j) + pow(phi[j-1],i);

/* calculate forecasts for age-specific mortality;
   produces M(x,y), x = 0,...,100, y = 1,...,48
   (note that M(x,0) was read in earlier) */
for(j=1; j <= 48; j++)
  {
   for(k=0; k < 5; k++)
     {
      dif[k] = phi[k]*dif[k];
      beta[k] += dif[k];
     }
   for(i=0; i <= 100; i++)
     {
      r1 = 0;
      for(k=0; k < 5; k++)
        r1 += beta[k]*U(i,k+1);
      M(i,j) = exp(r1);
     }
  }

 /* calculate the forecast of the survivors;
    produces P(x,y), x = 0,...,100, y = 1,...,48
    (note that P(x,0) was read in earlier and P(x,y) = 0 for x < y)
*/
 for(j=1; j <= 48; j++)
   {
    for(i=0; i <= 100; i++)
      if (i < j)
        P(i,j) = 0;
      else if (i < 100)
        P(i,j) = P(i-1,j-1)*exp(-M(i-1,j-1));
      else
        P(100,j) = (P(99,j-1)+P(99,j-1))*exp(-M(99,j-1));
   }

 /* print the mortality rates */
 for(i=0; i <= 100; i++)
   {
    for(j=0; j <= 48; j++)
      fprintf(fp2,"%8.7f ", M(i,j));
    fprintf(fp2,"\n");
   }
 fclose(fp2);

/* error variances for the log of the survivors of the jump-off;
```

```
   produces VAR(x,y), x = 0,...,100, y = 1,...,48
   (note that VAR(x,y) = 0 for x < y) */

/* the following calculations are executed only when we have set
   mort = 1, above */

if(mort == 1)
{

for(i=0; i <= 100; i++)
  for(j=1; j <= 48; j++)
    VAR(i,j) = 0.0;
printf("variances intialized\n");
for(i=1; i <= 100; i++)
  {
   r1 = 0;
   for(j=1; j <= 5; j++)
     for(k=1; k <= 5; k++)
       r1 += U(i,j)*C1(j,k)*U(i,k);
   VAR(i,1) = M(i-1,1)*M(i-1,1)*(r1 + C2(i-1,i-1));
  }
printf("initial values for recursion calculated\n");
for(i=1; i <= 99; i++)
  {
   per = (48 < 101-i) ? 48 : 101-i;
   for(t=2; t <= per; t++)
     {
      j1 = i + t - 1;
      r1 = 0;
      for(k=0; k <= t-1; k++)
       for(j=1; j <= 5; j++)
       for(h=1; h <= 5; h++)
        r1 +=U(j1-1,j)*U(j1-1,h)*PSI(t-1-k,j)*PSI(t-1-k,h)*C1(j,h);
      VAR(j1,t) =
VAR(j1-1,t-1)+M(j1-1,t-1)*M(j1-1,t-1)*(r1+C2(j1-1,j1-1));
      for(k1=0; k1 <= t-1; k1++)
      {
       r1 = 0;
       for(k=0; k <= k1; k++)
     for(j=1; j <= 5; j++)
     for(h=1; h <= 5; h++)
      r1 +=U(j1-1,j)*U(j1-t+k,h)*PSI(t-1-k,j)*PSI(k1-k,h)*C1(j,h);
       VAR(j1,t) += 2*M(j1-1,t-1)*M(j1-t+k1,k1)*r1;
      }
     }
  printf("age %u at jump-off handled\n", i);
  }

/* print the standard deviations */
for(i=0; i <= 100; i++)
  {
   fprintf(fp8,"\nage = %u\n", i);
   for(j=1; j <= 48; j++)
     fprintf(fp8,"%10.9f ", pow(VAR(i,j),0.5));
   fprintf(fp8, "\n");
   }
```

```
fclose(fp8);

/* print the standard deviations for selected cohorts */
for(j=1; j <= 48; j++)
  {
   fprintf(fp9,"%3u ", j);
   for(i=0; i <= 5; i++)
     fprintf(fp9,"%10.9f ", pow(VAR(10*i+j,j),0.5));
     fprintf(fp9, "\n");
  }
fclose(fp9);

} /* end of the variance calculations relating to survival that are
executed only when the condition mort = 1 is true */

/* the fertility calculations start here;
they are executed only if the condition fert = 1 is true */

if(fert == 1)
{

/* open the input and output files */
fp10=fopen("ferteig.dat","r");
fp11=fopen("futfert","w");
fp12=fopen("fertmlt.cov","r");
fp13=fopen("fertsig2.dat","r");

/* read the first four eigenvectors into columns 2,...,5 of U1 and
add
a column of 1's before them */
for(i=0; i <= 32; i++)
  {
   U1(i,1) = 1.0;
   for(j=2; j <= 5; j++)
      fscanf(fp10, "%g", &U1(i,j));
   fscanf(fp10, "%g", &r1);
  }
fclose(fp10);

/* calculate the PSI-matrix for fertility;
   produces PSI1(x,y), x = 0,...,48, y = 1,...,5 */
for(j=1; j <= 5; j++)
  PSI1(0,j) = 1.0;
for(i=1; i <= 48; i++)
  for(j=1; j <= 5; j++)
    PSI1(i,j) = PSI1(i-1,j) + pow(phi1[j-1],i);

/* calculate a forecast of the log of age-specific fertility;
   produces F(x,y), x = 1,...,30, y = 1,...,48;
   note that F(0,y) that corresponds to age = 14 and F(31,y)
   and F(32,y) are also non-zero; these numbers have been added
   to F(1,y) and F(30,y), respectively */
for(j=1; j <= 48; j++)
  {
   for(k=0; k < 5; k++)
     {
        dif1[k] =  phi1[k]*dif1[k];
        beta1[k] += dif1[k];
       }
     for(i=0; i <= 32; i++)
       {
        r1 = 0;
        for(k=0; k < 5; k++)
       r1 += beta1[k]*U1(i,k+1);
        F(i,j) = r1;
       }
    }

 /* in the sequel we need a version of the eigenvectors that
    considers ages 15-44 only; hence we add contributions of
    ages 14 and 15 and ages 44-46 to ages 15 and 44 respectively */
 for(j=2; j <= 5; j++)
   {
    U1(1,j) += U1(0,j);
    U1(30,j) += U1(31,j) + U1(32,j);
   }

/* calculate birth forecasts year by year and survive the
births to the end of the forecast period; female births only!
store the births into the matrix B(x,y), x = 1,...,30, y = 1,...,48;
store the results on survivors into the matrix P(x,y) for x < y */
for(j=1; j <= 48; j++)
  {
   per = 48-j;
   r1 = 0;
   for(i=1; i <= 30; i++)
     {
      B(i,j) = 0.49*P(i,j-1)*exp(F(i,j));
      r1 += B(i,j);
     }
   P(0,j) = r1;
   for(k=1; k <= per; k++)
     P(k,j+k) = P(k-1,j+k-1)*exp(-M(k-1,j+k-1));
  }

 /* print the population */
 for(i=0; i <= 100; i++)
   {
    for(j=0; j <= 48; j++)
      fprintf(fp1, "%10.1f ", P(i,j));
    fprintf(fp1,"\n");
   }
 fclose(fp1);

/* read in the covariance matrix of the betas into C4(x,y),
   x = 1,...,5, y = 1,...,5  */
for(i=1; i <= 5; i++)
  for(j=1; j <= 5; j++)
    fscanf(fp12,"%g", &C4(i,j));

/* leave only the variance of log(tfr), set everything else to
   zero */
```

```
/*
for(i=1; i <= 5; i++)
  for(j=1; j <= 5; j++)
    C4(i,j) = (i*j == 1) ? 0.0025 : 0.0;
*/

/* read in the covariance matrix of the error due to lack of fit
   into C5(x,y), x = -1,0,1,...,31, y = -1,0,1,...,31; the first
   two rows and columns and the last row and column will not be
   used in calculations */
for(i=-1; i <= 31; i++)
  for(j=-1; j <= 31; j++)
    fscanf(fp13,"%g", &C5(i,j));


printf("start calculating covariance of log-births\n");
/* calculate the covariance matrix of the log-births;
   produces C3(x,y) for 1 <= x, y <= 16, and C3(x,x)
   for 16 < x <= 32, other elements od C3 are set to zero */
for(i=1; i <= 32; i++)
   for(j=1; j <= 32; j++)
     C3(i,j) = 0.0;
printf("C3 initialized\n");
for(i=1; i <= 32; i++)
  {
  per = (i <= 16) ? 16 : i;
  for(j=i; j <= per; j++)
    {
     printf("iterate: i = %u and j = %u\n", i,j);
     r1 = 0.0;
     for(k=1; k <= 30; k++)
       for(h=1; h <= 30; h++)
         r1 += B(k,i)*B(h,j)*col(k,h,i,j);
     C3(i,j) = r1/(P(0,i)*P(0,j));
     C3(j,i) = C3(i,j);
    }
  }

/* add variance of past births to the diagonal of C3 for
   x = 17,...,32 and store it into V */

for(i=1; i <= 32; i++)
  for(j=1; j <= 5; j++)
    V(i,j) = 0.0;

for(i=17; i <= 32; i++)
  {
   r1 = 0.0;
   V(i,3) = C3(i,i);
   for(k=1; k <= i-16; k++)
     for(h=1; h <= i-16; h++)
       r1 += B(k,i)*B(h,i)*C3(i-15-k,i-15-h);
   V(i,1) = r1/(P(0,i)*P(0,i));
   C3(i,i) += V(i,1);
  }
printf("\nvariances of past births added to C3\n start
covariances\n");
/* add twice the covariances between current fertility and
   past births to the diagonal of C3 for x = 17,...,32 and
   store the values into V */
for(i=17; i <= 32; i++)
  {
   printf("start year %u\n", i);
   r1 = 0.0;
   for(k=1; k <= 32; k++)
     for(h=1; h <= i-16; h++)
       r1 += B(k,i)*B(h,i)*co2(k,i,i-h-15);
   V(i,2) = 2*r1/(P(0,i)*P(0,i));
   C3(i,i) += V(i,2);
  }

/* print results */
for(i=1; i <= 32; i++)
  {
   printf("Year is %u, SD = %10.9f\n", i, pow(C3(i,i),0.5));
   fprintf(fp11, "%u   %10.8f  ", i, pow(C3(i,i),0.5));
   r1 = V(i,1);
   fprintf(fp11, "%10.8f  ", r1);
   r1 = V(i,2);
   fprintf(fp11, "%10.8f  ", r1);
   r1 = V(i,3);
   fprintf(fp11, "%10.8f\n", r1);

} /* end of fertility computations */
} /* end of main */

/* co1 calculates Cov(ef(i1,t2),ef(i2,t3)) for
   i1,i2 = 1,...,30, t2, t3 > 0 */
float co1(int i1, int i2, int t2, int t3)
{
 int k1, k2 ,k3, a1;
 float sto, *d1, *d2, d3[5], d4[5];
 sto = (t2 == t3) ? C5(i1,i2) : 0.0;
 a1 = (t2 <= t3) ? t2 : t3;
 d1=d3-1;
 d2=d4-1;
 for(k1=1; k1 <= a1; k1++)
   {
    for(k3=1; k3 <= 5; k3++)
      {
       d1[k3] = PSI1(t2-k1,k3)*U1(i1,k3);
       d2[k3] = PSI1(t3-k1,k3)*U1(i2,k3);
      }
    for(k2=1; k2 <= 5; k2++)
      for(k3=1; k3 <= 5; k3++)
     sto += d1[k2]*d2[k3]*C4(k2,k3);
   }
 return sto;
 }

/* co2 calculates the covariance between the current fertility
   and the uncertainty of the current mothers' own births,
   i3 is the age of the mother, t4 is the current year, and
   t5 is the mother's birth year */
float co2(int i3, int t4, int t5)
{
 int k5;
 float sto1;
 sto1 = 0.0;
 for(k5=1; k5 <= 30; k5++)
   sto1 += B(k5,t5)*co1(i3,k5,t4,t5);
 return sto1/P(0,t5);
}
```

Figure 1. Magnitude of Relative Error

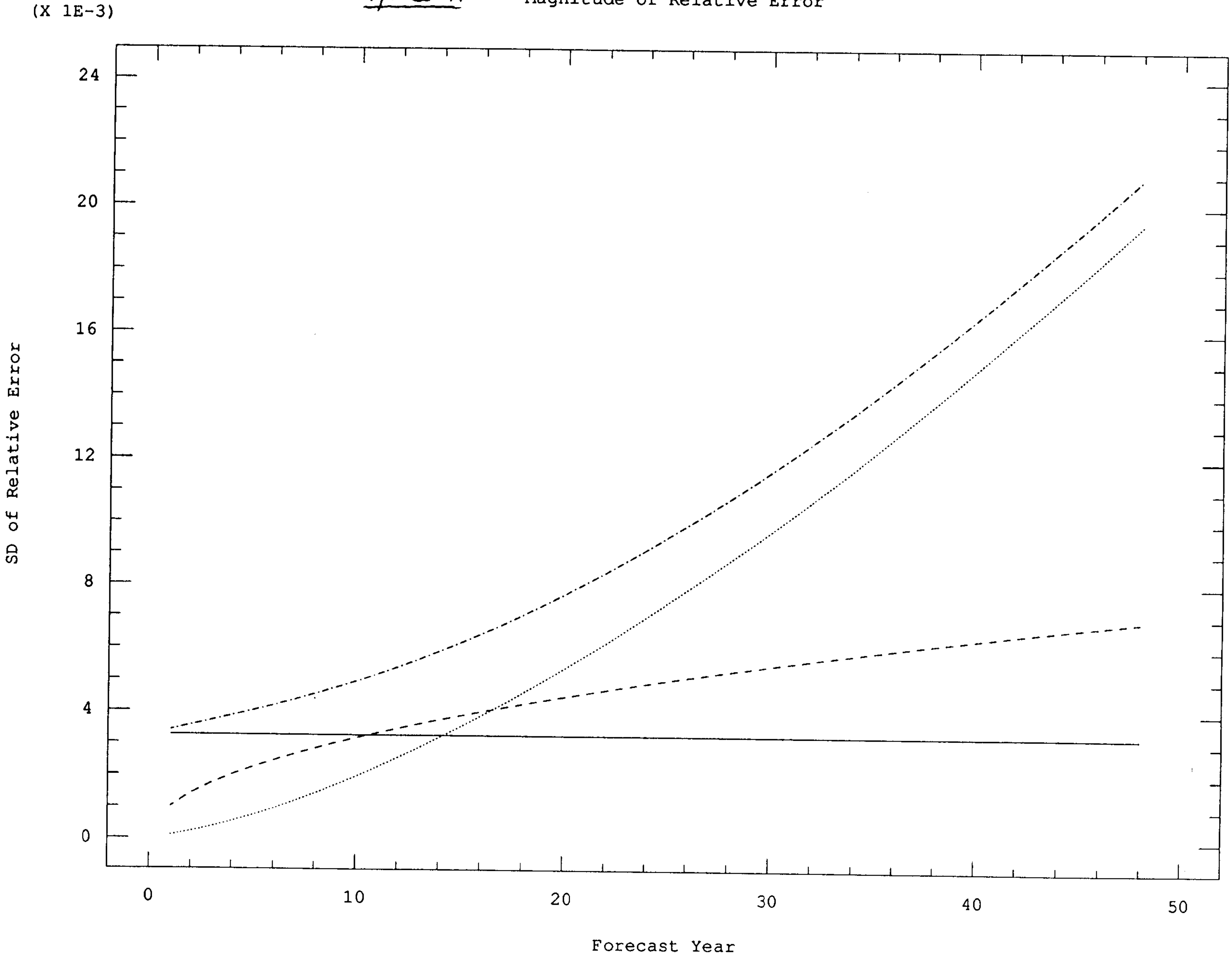


Figure 2. SD of Error for Different Cohorts

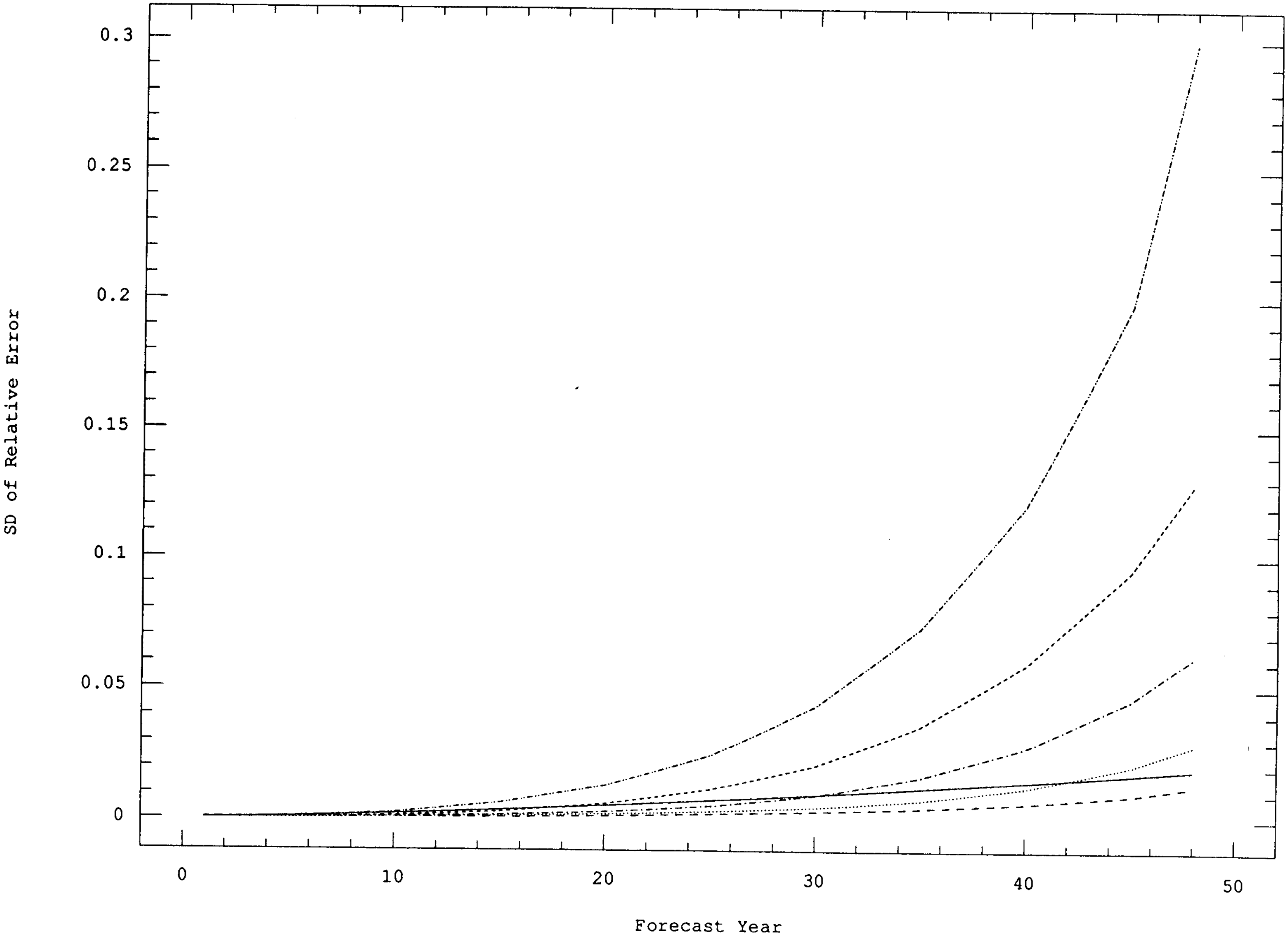


Figure 3. Direct Contribution of Fertility

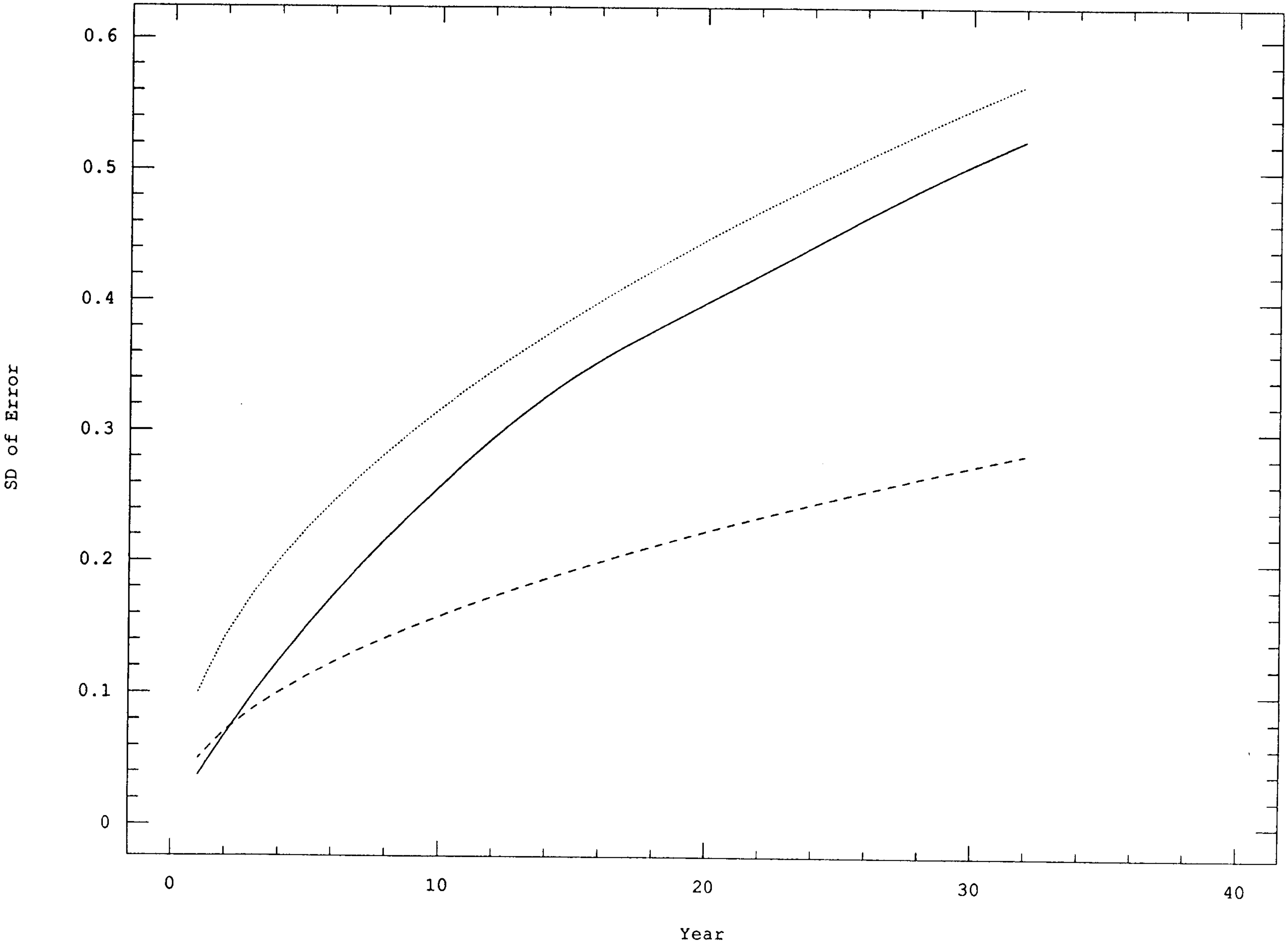
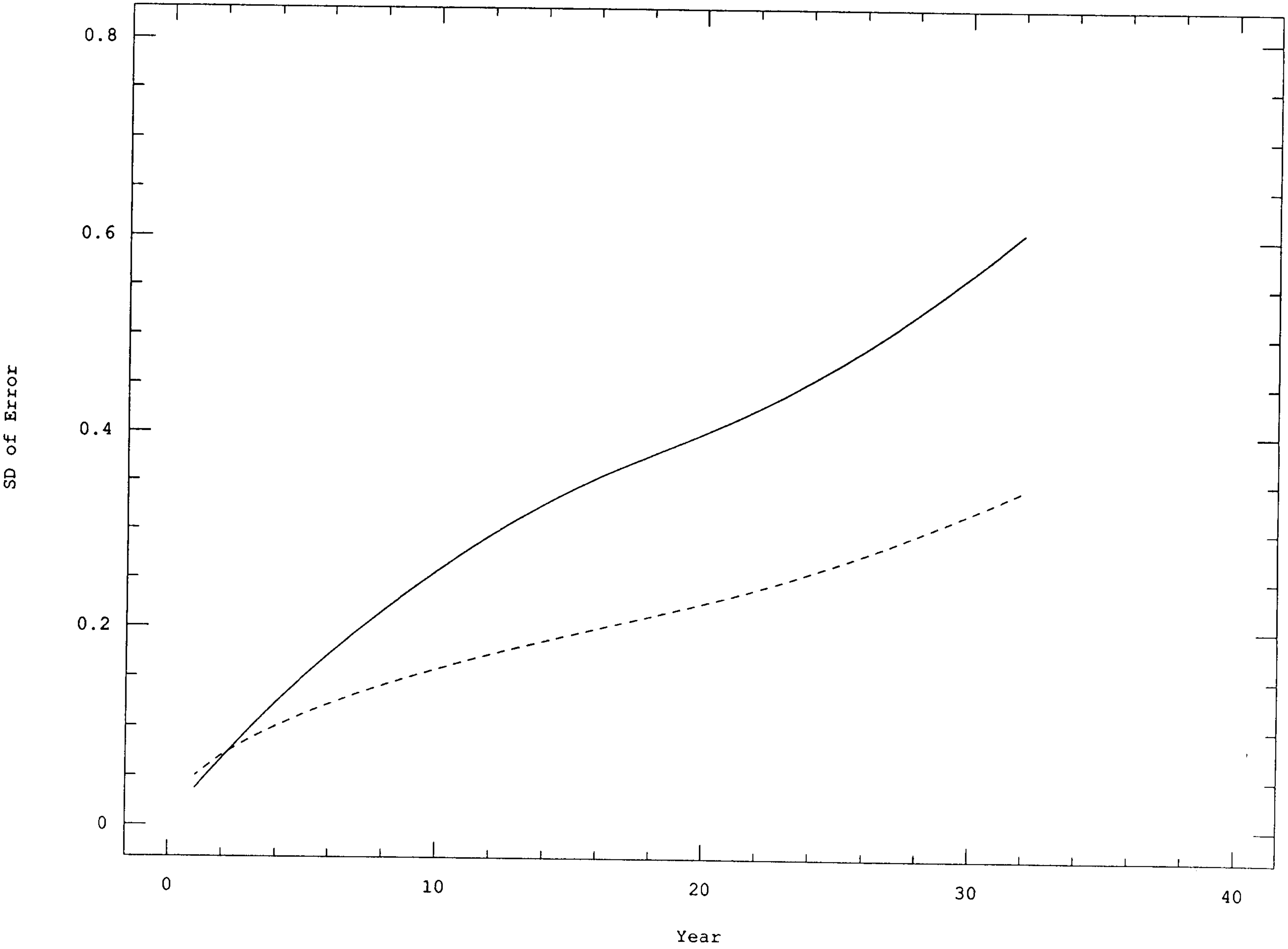


Figure 4. Error in the Principal Components and Analytical Models


SD of Error
Year