ISSN 1440-771X

MONASH University

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Stochastic population forecasts using functional data models for mortality, fertility and migration

Rob J Hyndman and Heather Booth

September 2007

Working Paper 14/06

Stochastic population forecasts using functional data models for mortality, fertility and migration

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University,
VIC 3800, Australia.
Email: Rob.Hyndman@buseco.monash.edu.au

Heather Booth
Australian Demographic & Social Research Institute,
The Australian National University,
Canberra ACT 0200, Australia.
Email: Heather.Booth@anu.edu.au.

21 September 2007

JEL classification: J11, C53, C14, C32

Stochastic population forecasts using functional data models for mortality, fertility and migration

Abstract: Age-sex-specific population forecasts are derived through stochastic population renewal using forecasts of mortality, fertility and net migration. Functional data models with time series coefficients are used to model age-specific mortality and fertility rates. As detailed migration data are lacking, net migration by age and sex is estimated as the difference between historic annual population data and successive populations one year ahead derived from a projection using fertility and mortality data. This estimate, which includes error, is also modeled using a functional data model. The three models involve different strengths of the general Box-Cox transformation chosen to minimise out-of-sample forecast error. Uncertainty is estimated from the model, with an adjustment to ensure the one-step-forecast variances are equal to those obtained with historical data. The three models are then used in the Monte Carlo simulation of future fertility, mortality and net migration, which are combined using the cohort-component method to obtain age-specific forecasts of the population by sex. The distribution of forecasts provides probabilistic prediction intervals. The method is demonstrated by making 20-year forecasts using Australian data for the period 1921–2003.

Key words: Fertility forecasting, functional data, mortality forecasting, net migration, nonparametric smoothing, population forecasting, principal components, simulation.

1 Introduction

Stochastic methods of population forecasting are rapidly gaining recognition. Stochastic population forecasts have been produced for the US, Australia and several European and other countries, as well as for the world and world regions. In the Netherlands, official statistical agencies now use stochastic methods, and other countries, such as the US, are also adopting their use in official forecasts.

Forecasts of the size and structure of the population are central to social and economic planning, from the provision of services in the short term to policy development in the long term. Not least of the demographic challenges facing developed countries is the rapid ageing of the population. Already developed-country populations are experiencing unprecedentedly large elderly proportions. The major driver of this ageing process is the fertility fluctuations of the past, notably the post-war baby boom coupled with the low fertility of recent times, but declining mortality is also significant. One response to population ageing and the attendant shortage of labour to provide for the elderly has been an increase in immigration to 'replace' or make up for past shortfalls in births (United Nations, 2001). Immigration has thus become a major driver of population change in many developed countries, and in some cases amounts to as much as 50% of the number of births; Australia and Spain are examples.

Population forecasting must take proper account of all three of these components of demographic change, that is mortality, fertility and migration. Mortality forecasting has received considerable attention in recent years. Methods for forecasting fertility and migration are less well developed: as with human behaviour in general, these demographic behaviours are difficult to forecast. A further problem for demographic forecasting is the estimation of uncertainty: estimates may vary considerably depending on the method of estimation (Keilman, 2001).

Any forecasting exercise presupposes that the data on which the forecast is based exist in suitable form. For mortality and fertility, this is generally the case for developed countries: vital registration provide lengthy series of data with the necessary detail. For migration, however, data are often lacking: where they exist at all, they tend to be in shorter series and often inadequately represent actual migration flows. A solution to this lack of data is to estimate net migration as the difference between the increment in population size and natural increase using the demographic growth-balance equation. For subnational population forecasting, where internal migration forecasts are required, this method is often the only approach available because data are not collected.

Several different approaches to demographic forecasting have been developed. The most widely used are those that involve some form of extrapolation, often using time series methods. Functional data methods fall under this category, but they have only recently been adopted in demographic forecasting (Hyndman and Ullah, 2007). Functional data methods have the advantage of providing a flexible framework that can be used for all three demographic processes.

This paper applies functional data models in forecasting mortality, fertility and net international migration for use in national population forecasting. These forecast components are combined using the cohort-component method to produce probabilistic population forecasts by age and sex. The method is illustrated using Australian data for 1921–2003 with a forecast horizon of 20 years. As complete and reliable data on international migration are lacking, annual net migration for 1972–2003 is estimated using the growth-balance equation.

1.1 Stochastic population forecasting

Fully probabilistic population forecasts have the major advantage of probabilistic consistency among all forecast variables including ratios and other derived indices (Lee and Tuljapurkar, 1994). They are generated through stochastic population renewal using the cohort-component method of population projection (Preston et al., 2001, pp.119–129). The forecast is achieved either analytically using the stochastic Leslie matrix (Lee and Tuljapurkar, 1994; Sykes, 1969; Alho and Spencer, 1985) or more simply by Monte Carlo simulation to produce a distribution of possible outcomes. In both approaches, it is necessary to specify the mean (or median), variance-covariance structure and distributional form for each demographic component. These parameters may be estimated using informed judgment and ex-post evaluation of errors (Keilman, 1997; Lutz et al., 1996) or by formal statistical models (e.g., Lee, 1992; Alho and Spencer, 1997).

In this paper, the Monte Carlo simulation approach is adopted, and the parameters of the component forecasts are derived using functional data models and time series methods. The use of extrapolative methods presupposes that the trends of the past will be continued into the future. This assumption has proved to be a better basis for forecasting than either structural modelling involving exogenous variables or methods based on expectation (Booth, 2006). The following brief review of forecasting the three components focuses on extrapolative methods; for a more comprehensive review, see Booth (2006).

1.2 Mortality

One of the simplest methods for extrapolating mortality, commonly used by actuaries, applies empirical age-specific geometric mortality reduction factors to current death rates to forecast future rates (Pollard, 1987). A more parsimonious approach is to forecast the parameters of a 'law' of mortality representing the age pattern, also ensuring regularity. Among the numerous existing models, the eight-parameter Heligman-Pollard (1980) model and the multi-exponential model have been used in forecasting with limited success (McNown and Rogers, 1989; McNown et al., 1995; Forfar and Smith, 1987). Interdependencies among parameters call for multivariate ARIMA models.

The most prominent method in mortality forecasting is the Lee-Carter method (Lee and Carter, 1992) which uses singular value decomposition to reduce annual age-specific log death rates to a time-dependent index of the level of mortality and a set of time-independent parameters that modify the overall level at particular ages. Standard time series methods are used to model and forecast the level index over time. The Lee-Carter method has the advantages of parsimony and simplicity in application. These stem from the fact that the overall downward trend in mortality accounts for almost all of the variation. Further, in most applications, a random walk with drift has been an appropriate model for the mortality index, indicating constant rates of decline in age-specific death rates (e.g., Tuljapurkar et al., 2000).

Several variants and extensions of the original Lee-Carter method have been developed. Lee and Miller's (2001) variant is now widely used. A variant by Booth et al. (2002) has been shown to be at least as accurate as Lee-Miller in the short term (Booth et al., 2005, 2006). The Lee-Carter method has been further developed to incorporate a heteroscedastic Poisson error structure (e.g., Wilmoth, 1993; Brouhns et al., 2002) and to be applicable to mortality reduction factors (Renshaw and Haberman, 2003a). Booth et al. (2002) and Renshaw and Haberman (2003b) examine the use of more than one term. Parallel approaches within the GLM framework have also been developed (Renshaw and Haberman, 2003c).

Two recent extensions of the Lee-Carter method involve incorporating nonparametric smoothing into the model. De Jong and Tickle (2006) combine spline smoothing and estimation via the Kalman filter to fit a generalized version of the Lee-Carter model. Hyndman and Ullah (2007), following the functional data paradigm, propose smoothing the mortality curves for each year using constrained regression splines prior to fitting a model using principal components decomposition. These methods are compared in Booth et al. (2006). We extend the Hyndman-Ullah approach in this paper, and apply it to mortality, fertility and migration.

1.3 Fertility

Fertility rates can be separated into level (quantum) and pattern (tempo) effects. These can be forecast independently if quantum and tempo are not highly correlated, which is generally the case for developed countries (e.g., Thompson et al., 1989). Fertility has proved difficult to forecast due to structural change, and estimates of uncertainty are highly dependent on the particular model.

Early forecasts of fertility used time series methods to forecast the total number of births, reflecting the role of births in population growth (Saboia, 1977; McDonald, 1979, 1981). Total fertility and independent age-specific rates have also been forecast by time series methods (McDonald, 1984; Miller, 1986). Ortega and Poncela (2005) extend this approach to exploit common trends by jointly modelling the total fertility rates for a homogeneous cluster of countries using a dynamic factor model with common and country-specific factors.

Parameterisation has been employed in the form of the gamma, beta and Hadwiger distributions (e.g., Hoem et al., 1981; Thompson et al., 1989; Congdon, 1993; Chandola et al., 1999; Keilman and Pham, 2000), and the multi-exponential model (Knudsen et al., 1993; McNown et al., 1995). Parameter interpretability and over-parameterisation can present difficulties for forecasting and multivariate time series methods are desirable.

Lee (1993) forecast fertility using a parallel method to the Lee-Carter method, but found it necessary to pre-specify the long-term mean value of total fertility because of structural change and to impose limits of 0 and 4 to reduce the width of the prediction interval (see also Lee, 1999). A principal components approach was also employed by Bozik and Bell (1987) in forecasting fertility using the first four components and multivariate ARIMA methods. Hyndman and Ullah (2007) applied a similar method in adopting a functional data approach, and this method is further developed in this paper.

1.4 Migration

Methods for forecasting migration are the least developed among the three components and are often extremely simple (George and Perreault, 1992). One reason for this stems from a lack of suitable data. Ideally migration is separated into its two distinct processes, immigration and emigration, each of which is independently forecast by age and sex (Rogers, 1990). Decomposition by reason for migration and disaggregation of immigration by citizenship are also ideal (Hilderink et al., 2002). Forecasting immigration presents particular problems because of the uncertainty inherent in any process that is determined more by political and socio-economic considerations than any other.

A problem with the use of immigration data, even if accurate in what they measure, is that coverage of actual in-migration may be incomplete. Undocumented migration is a substantial flow for some countries. The US, for example, receives an estimated 500,000 illegal immigrants per year, while in Spain the immigration amnesty of 2005 resulted in 700,000 registrations, equal to 2 per cent of the total population. In Australia, uncounted immigration has occurred through ‘category jumping’ whereby temporary migrants (not counted as immigrants in migration data) legally change status when inside the country, thereby becoming permanent immigrants (Khoo and McDonald, 2002). Further, immigration data may be misleading if repeat migrants are counted each time they enter the country: overseas students, for example, may be counted annually as long-term in-migrants but not counted as out-migrants when they return home for the vacation. These and other inaccuracies demand that data be decomposed as far as possible into the different types of migrants and checked for reliability.

Emigration data present even greater problems, not least of which is that they are often unavailable or incomplete in coverage. When data are available, changes of emigration status after leaving the country reduce data accuracy. Reliance on the immigration data of main destination countries suffers from incomplete coverage and problems arising from different classifications.

Estimating net migration as the residual from the demographic growth-balance equation may be seen as a partial solution to the problems of data inadequacy, because population data are generally more complete. The coverage of illegal migrants in population censuses increases with the passage of time and with the acquisition of permanent legal status through amnesties. However, difficulties can arise: even if the two processes of immigration and emigration are stable, the difference between them can be unstable and therefore difficult to model and forecast. Nevertheless, where detailed data are not available or are of poor quality, estimated net migration may be the best solution.

Migration numbers are often forecast rather than rates. The use of numbers is consistent with the specification of annual quotas for immigrants, and total numbers are often forecast. Further, Miller and Lee (2004) note that numbers are preferable in that they take into account the co-variation in the rate of immigration (or positive net migration) and the population denominator. This also applies to emigration. However, the use of numbers can be problematic if, for example, age-specific emigration is forecast to exceed the age group's population. Time series methods are often used to forecast aggregate immigration and emigration (e.g., De Beer, 1997; Keilman and Pham, 2004; Wilson and Bell, 2004); De Beer (1997) also forecast net migration with consistent results.

The use of rates demands specification of the relevant population denominator. While emigration may be meaningfully related to the domestic population, this is less relevant for immigration and net migration, though it is often used. Miller (2003) forecast the total net migration rate for California, multiplied it by the total population in the preceding year to obtain total net migration and then applied the most recent empirical age-sex distribution as a constant distribution for the duration of the forecast. A similar approach is used by Miller and Lee (2004).

Where detailed data by age are lacking, a model may be used; this also reduces the number of parameters to be forecast. The multi-exponential model (Rogers and Castro, 1981; Rogers and Little, 1994) with up to 13 parameters is widely known, but has been used with only limited success for forecasting (Rogers et al., 2005; George, 1994). Keilman and Pham (2004) used a six-parameter version of this model to disaggregate forecast totals by age.

In this paper, we estimate net migration numbers as the residual from the age-sex-specific demographic growth-balance equation. Though this estimate includes errors in vital registration data and annual population estimates, the resulting series is superior to available migration data in its coverage of years, events and single years of age.

1.5 Structure of the paper

In the following section, we discuss the data requirements of our approach, and explain how derived data are obtained. Section 3 describes the functional data models we fit to each of the components (mortality, fertility and net migration). In Section 4, we discuss how to simulate future sample paths for each of these components, thereby obtaining simulated projections of age-specific population numbers by sex. We apply the method to Australian data in Section 5 to obtain twenty-year probabilistic forecasts of the population by age and sex. Some conclusions and discussion including extensions to this approach are contained in Section 6.

2 Data requirements

We use the following data: age-specific birth and death numbers for each calendar year, age-specific population numbers at 1 January in each year, and exposures to risk (or populations at 30 June) in each year. All of these data are available from the Human Mortality Database (2006) for most of the last century for most developed countries. We use the following notation for these data:

$$\begin{aligned}
B_t(x) &= \text{Births in calendar year } t \text{ to females of age } x;\\
D_t^F(x) &= \text{Deaths in calendar year } t \text{ of females of age } x;\\
D_t^M(x) &= \text{Deaths in calendar year } t \text{ of males of age } x;\\
P_t^F(x) &= \text{Female population of age } x \text{ at 1 January, year } t;\\
P_t^M(x) &= \text{Male population of age } x \text{ at 1 January, year } t;\\
E_t^F(x) &= \text{Female population of age } x \text{ exposed to risk (30 June), year } t;\\
E_t^M(x) &= \text{Male population of age } x \text{ exposed to risk (30 June), year } t;
\end{aligned}$$

where $x = 0, 1, 2, \dots, p-1, p^+$ and $t = 1, \dots, n$. Here, p^+ denotes the open-ended upper age group. To avoid repetition, we will drop the sex-superscript when the same notation applies to each sex.

Let $m_t(x) = D_t(x)/E_t(x)$ be the age-sex-specific central death rates in calendar year t and let $f_t(x) = B_t(x)/E_t^F(x)$ be the fertility rate for females of age x in calendar year t. Births are divided by sex using ρ, the male:female sex-ratio at birth:

$$B_t^F = \frac{1}{1+\rho}B_t \quad \text{and} \quad B_t^M = \frac{\rho}{1+\rho}B_t.$$

Net migration, denoted G_t, is estimated for each sex using the demographic growth-balance equation:

$$\begin{aligned}
P_{t+1}(x+1) &= P_t(x) && - D_t(x,x+1) && + G_t(x,x+1) && x = 0,1,2,\ldots,p-2;\\
P_{t+1}(p^+) &= P_t(p-1) + P_t(p^+) && - D_t(p-1^+,p^+) && + G_t(p-1^+,p^+);\\
P_{t+1}(0) &= B_t && - D_t(B,0) && + G_t(B,0),
\end{aligned}$$

where $D_t(x,x+1)$ refers to deaths in calendar year t of persons aged x at the beginning of year t, $D_t(p-1^+,p^+)$ refers to deaths in calendar year t of persons aged $p-1$ and older at the beginning of year t and $D_t(B,0)$ refers to deaths in calendar year t of births during year t; and similarly for net migration $G_t(x,x+1)$, $G_t(p-1^+,p^+)$ and $G_t(B,0)$. In other words, these are cohort rhomboids in the Lexis diagram (see Preston et al., 2001).

At most ages, $D_t(x,x+1)$ could be estimated as the average of observed deaths at ages x and $x+1$. However, at age 0, averaging leads to bias. To avoid such bias, we use the standard life table approach used in population projection: cohort deaths are estimated as the product of the population at time t, or births in year t, and the complement of the relevant survivorship ratio from the life table calculated using $m_t(x)$. Thus, for each sex,

$$\begin{aligned}
D_t(x,x+1) &= P_t(x)[1 - L(x+1)/L(x)] && x = 0,1,2,\ldots,p-2;\\
D_t(p-1^+,p^+) &= P_t(p-1^+)[1 - T(p)/T(p-1)];\\
D_t(B,0) &= B_t[1 - L(0)/l(0)]
\end{aligned}$$

where $L(x)$ denotes person-years lived at age x, $T(x)$ denotes person-years lived at age x^+ and $l(0)$ is the radix of the life table (Preston et al., 2001).

Then:

$$\begin{aligned}
G_t(x,x+1) &= P_{t+1}(x+1) - P_t(x) && + D_t(x,x+1) && x = 0,1,2,\ldots,p-2;\\
G_t(p-1^+,p^+) &= P_{t+1}(p^+) && - P_t(p^+) - P_t(p-1) + D_t(p-1^+,p^+);
\end{aligned}$$

$$G_t(B,0) = P_{t+1}(0) \qquad - B_t \qquad\qquad\qquad + D_t(B,0).$$

The estimation of net migration in the last year of the fitting period requires population data for 1 January, year $n+1$. Note that as $G_t(B,0)$ refers to the migration of those born in year t, it is generally smaller in size than $G_t(0,1)$. The $G_t(x,x+1)$ are forecast; this corresponds directly with the cohort approach of population dynamics.

3 Functional data modelling approach

To obtain forecasts of each component, we first develop functional time series models for the five sex-specific components: $m_t^F(x)$, $m_t^M(x)$, $f_t(x)$, $G_t^F(x,x+1)$, and $G_t^M(x,x+1)$. The five models will then be used in the simulation of the future population. We follow the approach of Hyndman and Ullah (2007) to model each of these components.

3.1 Functional data models

Let $y_t^*(x)$ denote the quantity being modelled—either mortality rates, fertility rates, or net migration numbers for age x in year t. First we use a Box and Cox (1964) transformation of $y_t^*(x)$ to allow for variation that increases with the value of $y_t^*(x)$. Thus

$$y_t(x) = \begin{cases} \frac{1}{\lambda}\left([y_t^*(x)]^\lambda - 1\right) & \text{if } 0 < \lambda < 1; \\ \log_e(y_t^*(x)) & \text{if } \lambda = 0. \end{cases}$$

The value of λ determines the strength of the transformation. Then we assume the following model for the transformed quantity $y_t(x)$:

$$\begin{aligned} y_t(x) &= s_t(x) + \sigma_t(x)\varepsilon_{t,x} \\ s_t(x) &= \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x) + e_t(x) \end{aligned}$$

where $s_t(x)$ is an underlying smooth function of x, $\varepsilon_{t,x} \sim \text{IID}(0,1)$ and $\sigma_t(x)$ allows the variance to change with age and time. That is, $s_t(x)$ is a smooth function of age which we observe with error. The second equation describes the dynamics of $s_t(x)$ evolving through time. In this equation, $\mu(x)$ is the mean of $s_t(x)$ across years, $\{\phi_k(x)\}$ is a set of orthogonal basis functions calculated using a principal components decomposition, and $e_t(x)$ is the model error which is assumed to be serially uncorrelated. The dynamics of the process are controlled by the time series coefficients $\{\beta_{t,k}\}$ which are assumed to behave independently of each other (this follows from using principal components decomposition).

This model was first proposed by Hyndman and Ullah (2007) for mortality and fertility rates although they used log transformations instead of the more general Box-Cox transformation. It has also been used by Erbas et al. (2007) for forecasting breast cancer mortality rates.

As Hyndman and Ullah (2007) point out, the model is a generalization of the well-known Lee-Carter (1992) model for forecasting mortality rates. In the Lee-Carter approach, $y_t^*(x)$ denotes mortality rates and $\lambda = 0$ so that $y_t(x)$ represents log mortality for year t and age x. The Lee-Carter method does not assume smoothness, so $\sigma_t(x) = 0$ and $y_t(x) = s_t(x)$. Then $\mu(x)$ is estimated as the average of $y_t(x)$ across years, $K = 1$, and $\phi_1(x)$ and $\beta_{t,1}$ are computed from the first principal component of the matrix of $[y_t(x) - \hat{\mu}(x)]$. Forecasts are obtained by fitting a time series model to $\beta_{t,1}$; in practice this is almost always a random walk with drift.

In this paper, we extend the method of Hyndman and Ullah (2007) by using a more general transformation, and by applying the model to net migration numbers, mortality rates and fertility rates. We also modify the method Hyndman and Ullah used in calculating the forecast variance to allow for better calibration with the observed data.

Hyndman and Ullah proposed robust estimation of the model terms in order to handle the effect of epidemics and wars on mortality data. We will avoid this additional complexity in modelling Australian data by restricting mortality data to 1950 onwards.

The modelling steps (described in detail in Hyndman and Ullah, 2007) are:

1. Estimate smooth functions $s_t(x)$ using nonparametric regression applied to $y_t(x)$ for each year t. (In our application to Australian data, we use weighted penalized regression splines.)

2. Estimate $\mu(x)$ as the mean of $s_t(x)$ across years.

3. Estimate $\beta_{t,k}$ and $\phi_k(x)$, $k = 1, \dots, K$, using principal components decomposition of $[y_t(x) - \hat{\mu}(x)]$.

4. Estimate time series models for $\beta_{t,k}$, $k = 1, \dots, K$. We use exponential smoothing state

space models.

The value of K must be specified. Hyndman and Ullah (2007) proposed selecting K to minimize the mean integrated squared forecast error. Since then, we have found that the method is insensitive to the choice of K provided K is large enough. That is, there is little cost (apart from computing time) in choosing a large K, whereas a K too small may result in poor forecast accuracy. Consequently, in this analysis we choose $K = 6$ for all components; this seems to be larger than any of the components really require.

The observational variance, $\sigma_t^2(x)$, depends on the nature of the data. For deaths, we estimate observational variance from $y_t^*(x) = m_t(x)$, assuming deaths are Poisson distributed (Brillinger, 1986) with mean parameter $m_t(x)E_t(x)$. Thus, $y_t^*(x)$ has approximate variance $E_t^{-1}(x)m_t(x)$, and the variance of $y_t(x)$ is (via a Taylor approximation)

$$\sigma_t^2(x) \approx [m_t(x)]^{2\lambda-1}E_t^{-1}(x).$$

For births, we assume a Poisson distribution (Keilman et al., 2002) with mean $f_t(x)E_t^F(x)$, which gives

$$\sigma_t^2(x) \approx [f_t(x)]^{2\lambda-1}E_t^{-1}(x).$$

For migration data, we make no distributional assumptions, and we estimate $\sigma_t^2(x)$ by a non-parametric regression of $[y_t(x) - s_t(x)]^2$ against x.

The success of the model depends on how well the bivariate surface $\{s_t(x) - \mu(x)\}$ can be approximated by the sum of a few products of univariate functions of time (t) and age (x). So far, we have applied this model to mortality data from about twenty populations, and to fertility data and migration data from Australia. This experience suggests that the model is good at producing point forecasts, but not quite so good at estimating forecast variance. Consequently, we propose below an adjustment to the forecast variance implied by the model.

3.2 Functional forecasts

Suppose we have data up to time $t = n$, and we wish to estimate future values of $y_t(x)$ for $t = n+1, \dots, n+h$ and all x. Let $\hat{\beta}_{n,k,h}$ denote the h-step ahead forecast of $\beta_{n+h,k}$, let $\hat{y}_{n,h}(x)$ denote the h-step ahead forecast of $y_{n+h}(x)$ and let $\hat{s}_{n,h}(x)$ denote the h-step ahead forecast of

$s_{n+h}(x)$. Then

$$\hat{y}_{n,h}(x) = \hat{s}_{n,h}(x) = \hat{\mu}(x) + \sum_{k=1}^{K} \hat{\beta}_{n,k,h}\hat{\phi}_k(x). \tag{1}$$

A forecast of $y_t^*(x)$ is found through back-transformation.

Following Hyndman and Ullah (2007), we can give the following expression for forecast variance

$$V_h(x) = \text{Var}[s_{n+h}(x) \mid \mathscr{I}, \mathbf{\Phi}] = \hat{\sigma}_\mu^2(x) + \sum_{k=1}^{K} u_{n+h,k}\, \hat{\phi}_k^2(x) + v(x) \tag{2}$$

where $\mathscr{I} = \{y_t(x_i)\}$ denotes all observed data, $u_{n+h,k} = \text{Var}(\beta_{n+h,k} \mid \beta_{1,k}, \dots, \beta_{n,k})$ can be obtained from the time series model, $\hat{\sigma}_\mu^2(x)$ (the variance of the smooth estimate $\hat{\mu}(x)$) can be obtained from the smoothing method used, and $v(x)$ is estimated by averaging $\hat{e}_t^2(x)$ for each x. Thus, the smoothing error is given by the first term, the error due to predicting the dynamics is given by the second term, and the third term gives the error due to the unexplained dynamic variation. After the observational error is also included, we obtain

$$\text{Var}[y_{n+h}(x) \mid \mathscr{I}, \mathbf{\Phi}] = V_h(x) + \sigma_t^2(x).$$

Note that correlations between ages are naturally dealt with in this formulation due to the smooth functions of age (x). Also, correlations between years are handled by the time series models for the coefficients $\beta_{t,1}, \dots, \beta_{t,K}$.

For low values of h, we can check the validity of $V_h(x)$ by computing the in-sample empirical forecast variance

$$W_h(x) = \frac{1}{n-h-m+1} \sum_{t=m}^{n-h} [s_{t+h}(x) - \hat{s}_{t,h}]^2$$

where m is the smallest number of observations used to fit a model. In practice, considerable differences between $W_h(x)$ and $V_h(x)$ can occur. Consequently, we use the following adjusted variance expression:

$$\text{Var}[y_{n+h}(x) \mid \mathscr{I}, \mathbf{\Phi}] = V_h(x)W_1(x)/V_1(x) + \sigma_t^2(x). \tag{3}$$

This adjusts the variance so that the one-step forecast variance matches the in-sample empirical one-step forecast variance. It is assumed that the same multiplicative adjustment is applicable at higher forecast horizons.

Note that we could similarly check for bias in the point forecasts by averaging $s_{t+h}(x) - \hat{s}_{t,h}$, but empirical results suggest the model provides excellent forecasts.

4 Stochastic cohort simulation from functional data models

Population sample paths are simulated using the cohort-component method (see Preston et al., 2001) adapted to permit random observational error. The base (or jump-off) population is the observed population at 1 January in year $n+1$.

For each of $m_t^F(x)$, $m_t^M(x)$, $f_t(x)$, $G_t^F(x, x+1)$ and $G_t^M(x, x+1)$, we simulate a large number of future sample paths which are then used to compute future sample paths of the age-sex-specific population. For each sex-specific component, we use the time series models to generate random sample paths of $\beta_{t,k}$ for $t = n+1, \dots, n+h$ conditional on $\beta_{1,k}, \dots, \beta_{n,k}$. We also generate random values of $e_t(x)$ by bootstrapping the estimated values.

Having obtained the simulated values of $s_t(x)$ for $t = n+1, \dots, n+h$, we then apply the variance adjustment described in (3); the adjusted value of $s_{n+h}(x)$ is

$$\tilde{s}_{n+h}(x) = \hat{\mu}(x) + \sum_{k=1}^{K} \hat{\beta}_{n,k,h} \hat{\phi}_k(x) + \left[s_{n+h}(x) - \hat{\mu}(x) - \sum_{k=1}^{K} \hat{\beta}_{n,k,h} \hat{\phi}_k(x) \right] \sqrt{W_1(x)/V_1(x)}.$$

In order to take account of births and deaths to annual migrants, the population is first adjusted for net migration: by adding half of net migration at the beginning of the year and half at the end, migrants are assumed to spend on average half of the year exposed to events (see Preston et al., 2001). Net migration for age x at the beginning of year $n+h$, $G_{n+h}(x, x+1)$, is computed using the simulated smoothed mean $\tilde{s}_{n+h}(x)$ and resampling the errors $y_t(x) - s_t(x)$. As we are using cohort net migration, half of simulated annual net migration, $G_t(x, x+1)$, is added to the population aged x on 1 January of year t (i.e. $P_t(x)$), and half is added to the population aged $x+1$ at the end of year t (equivalent to adding to the population aged $x+1$ at 1 January of year $t+1$, i.e. $P_{t+1}(x+1)$). We denote the population adjusted for the first half of migration by R. Thus $R_t(x) = P_t(x) + G_t(x, x+1)/2$. For net migration of infants born in year t, half of $G_t(B, 0)$ is added to births in year t, i.e., $R_t(B) = B_t + G_t(B, 0)/2$, and half is added to the population aged 0 at the end of year t. At the oldest ages, $G_t(p-1^+, p^+)$ is assumed to be equally divided between migrants aged $p-1$ and p^+ at 1 January of year t.

Mortality is applied next. For each sex, we take a simulated sample path of $\{m_t(x)\}$, $t = n + 1, \dots, n + h$. In order to obtain a random number of deaths at age x in year t, the mid-year population $E_t(x)$ is needed; however, this depends on the number of deaths. This circularity is dealt with as follows. Using simulated $\{m_t(x)\}$, we obtain life table survivorship ratios and apply them to $R_t(x)$ to obtain the expected number of cohort deaths, denoted $\bar{D}_t(x, x + 1)$. We subtract these deaths from $R_t(x)$ to obtain an estimate of $R_{t+1}(x + 1)$, denoted $\bar{R}_{t+1}(x + 1)$. We average $R_t(x)$ and $\bar{R}_{t+1}(x)$ to obtain the estimated mid-year population $\bar{E}_t(x)$. Then we obtain the number of deaths in year t to persons aged x as a random draw from a Poisson distribution with mean $m_t(x)\bar{E}_t(x)$. Deaths in year t at each pair of adjacent ages are averaged (see below for age 0) to obtain cohort deaths, $D_t(x, x + 1)$, which is subtracted from $R_t(x)$ to obtain $R_{t+1}(x + 1)$. For the open-ended age group, $D_t(p - 1^+, p^+)$ is the sum of $D_t(p - 1)/2$ and $D_t(p^+)$. The addition of the second half of cohort net migration gives the population aged $x + 1$ on 1 January of year $t + 1$, i.e., $P_{t+1}(x + 1)$.

The number of births is then obtained using the simulated fertility rate $f_t(x)$ and the population adjusted for half of net migration. Using the female population of reproductive age, births at each age are assumed to be Poisson with mean $f_t(x)[R_t(x) + R_{t+1}(x)]/2$ for $x = 15, \dots, 49$. A random draw from this distribution determines $B_t(x)$, the number of births in year t to women of age x. Simulated B_t is then calculated as

$$B_t = \sum_{x=15}^{49} B_t(x).$$

Male and female births are allocated binomially with the probability of being male $\rho/(\rho + 1)$, where ρ is the male:female sex ratio at birth. We ignore any variation in ρ as the effect is small compared to the other sources of variation.

For each sex, mortality is applied to $R_t(B)$ to obtain $R_{t+1}(0)$. Again an estimate of the mid-year population is used. Using the life table derived previously for year t, the survivorship ratio from birth to age 0 is applied to migration-adjusted births to give an estimate of deaths in year t between birth and age 0, $\bar{D}_t(B, 0)$. These deaths are subtracted from $R_t(B)$ to give the estimate $\bar{R}_{t+1}(0)$. We then average $R_t(0)$ and $\bar{R}_{t+1}(0)$ to obtain the estimated mid-year population $\bar{E}_t(0)$. Deaths at age 0 in year t are obtained using a Poisson distribution with mean $\bar{D}_t(0) = m_t(0)\bar{E}_t(0)$. This randomly-drawn number of deaths, $D_t(0)$, is then divided into deaths to births in year t, $D_t(B, 0)$, and deaths at age 0 to births in year $t - 1$ using the separation factor, f_0, estimated as the ratio of expected deaths between birth and age 0 and expected deaths at age 0 in year t, i.e. $f_0 = \bar{D}_t(B,0)/\bar{D}_t(0)$. (The separation factor is also used to obtain cohort deaths at age 0 to 1 as the weighted average of $D_t(0)$ and $D_t(1)$, i.e. $D_t(0,1) = (1-f_0)D_t(0) + 0.5D_t(1)$). Subtracting $D_t(B,0)$ from $R_t(B)$ gives $R_{t+1}(0)$ to which the second half of simulated net migration, $G_t(B,0)/2$, is added to give $P_{t+1}(0)$, the population aged 0 on 1 January of year $t+1$.

Hence, using simulated births, deaths and migrants, we generate the population for the next year. This is repeated for years $t = n+1,\dots,n+h$, and the whole procedure is repeated for each population sample path. Thus we obtain a large number of future sample paths of age-sex-specific population and vital event numbers which can be used to estimate, with uncertainty, any demographic variable that is derived from population numbers and vital events, including life expectancies, total fertility rates, old-age dependency ratios, etc.

The entire procedure is summarized in the Appendix, showing the order in which the calculations need to be carried out.

5 Application to Australia

Most of the data were obtained from the Human Mortality Database (2006) and consist of central death rates, start-year and mid-year populations by sex and by age in single years for 0–99 and 100+ years for 1921–2003 (the start-year population for 2004 is also used in estimating net migration). The base population for the forecast refers to 1 January 2004. Age-specific fertility rates by single years of age for 15–49 for 1921–2003 were obtained from the Australian Demographic DataBank (located at the Australian Demographic and Social Research Institute, Australian National University).

To avoid difficulties with war years and the 1918 Spanish influenza epidemic and with structural change over the course of the twentieth century, we only use mortality data from 1950 onwards. When computing net migrant numbers, we only use data from 1972 onwards, as the population numbers are less reliable before that.

We follow the modelling framework outlined in Section 3. For mortality, we choose $\lambda = 0$ which is consistent with other studies (e.g., Lee and Carter, 1992; Booth et al., 2002; Hyndman and Ullah, 2007, etc.). As described in Hyndman and Ullah (2007), we constrain the fitted curves to be monotonically increasing for $x > 65$.

For fertility, there seems no consensus on the best transformation to use. Hyndman and Ullah (2007) use logarithms; Lee (1993) uses no transformation on age-specific fertility rates, but a logistic transformation (defined on the interval [0,4]) on the total fertility rate. We choose the value of $\lambda = 0.2$ as it gave a relatively small out-of-sample forecast errors (on the untransformed scale) and narrowest prediction intervals when applied to the Australian data for 1921–1993. The sex ratio at birth was set to be $\rho = 1.0545$ based on current Australian data.

For migration, the data are both positive and negative, so we do not use any transformation (i.e., $y_t^*(x) = y_t(x)$.)

For all models, $K = 6$ basis functions are used. This is larger than any of the components seem to require. As noted previously, the method is insensitive to the choice of K provided K is large enough; in other words, additional basis functions do not decrease forecast accuracy.

The time series models used are additive state space models for exponential smoothing, as described in Hyndman et al. (2002). Parameters are optimized by averaging the average of the squared in-sample h-step forecast errors for $h = 1, 2, \dots, 8$.

5.1 Results

Estimates of net migration for the period 1972–2003 are shown in Figure 1. It is seen that there is a high degree of variation at all but the oldest ages around a jagged age-specific mean. The variation in these estimated numbers is a combination of the variation in migration, errors in the age-specific population numbers, errors in estimated age-specific deaths, and errors in the estimated number of births. As population numbers are Estimated Resident Populations (ERPs) produced by the Australian Bureau of Statistics (ABS), rather than independent empirical counts, they are the product of models and assumptions and will be subject to (unknown) systematic bias as well as random error. Such bias may have contributed to the jagged mean age distribution, particularly at ages 35 to 60 where male and female age patterns are almost identical; it is also likely that digit preference occurs in the reported ages in the population, deaths and possibly original migration data used in official estimates of intercensal population size. These biases have undoubtedly produced the elevated estimates at age 80. These irregularities will be largely removed by smoothing. Positive net migration at very old ages is also

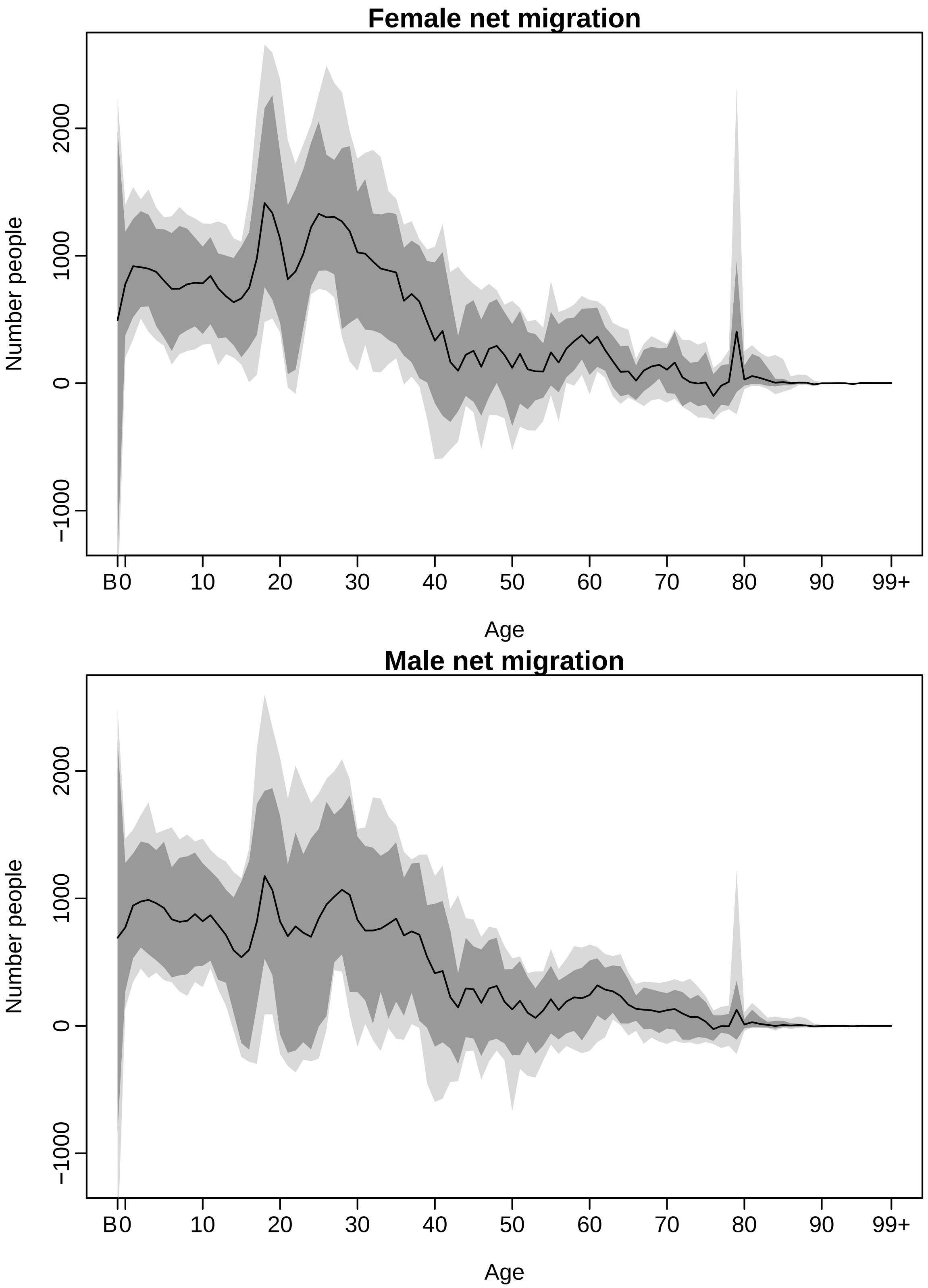

Figure 1: *Net migration estimates, 1972–2003. Light-shaded region shows 2.5% and 97.5% percentiles; dark-shaded region shows 10% and 90% percentiles; solid line shows the mean. On the age-axis, B represents the interval between birth and age 0 on 1 January in the following year.*

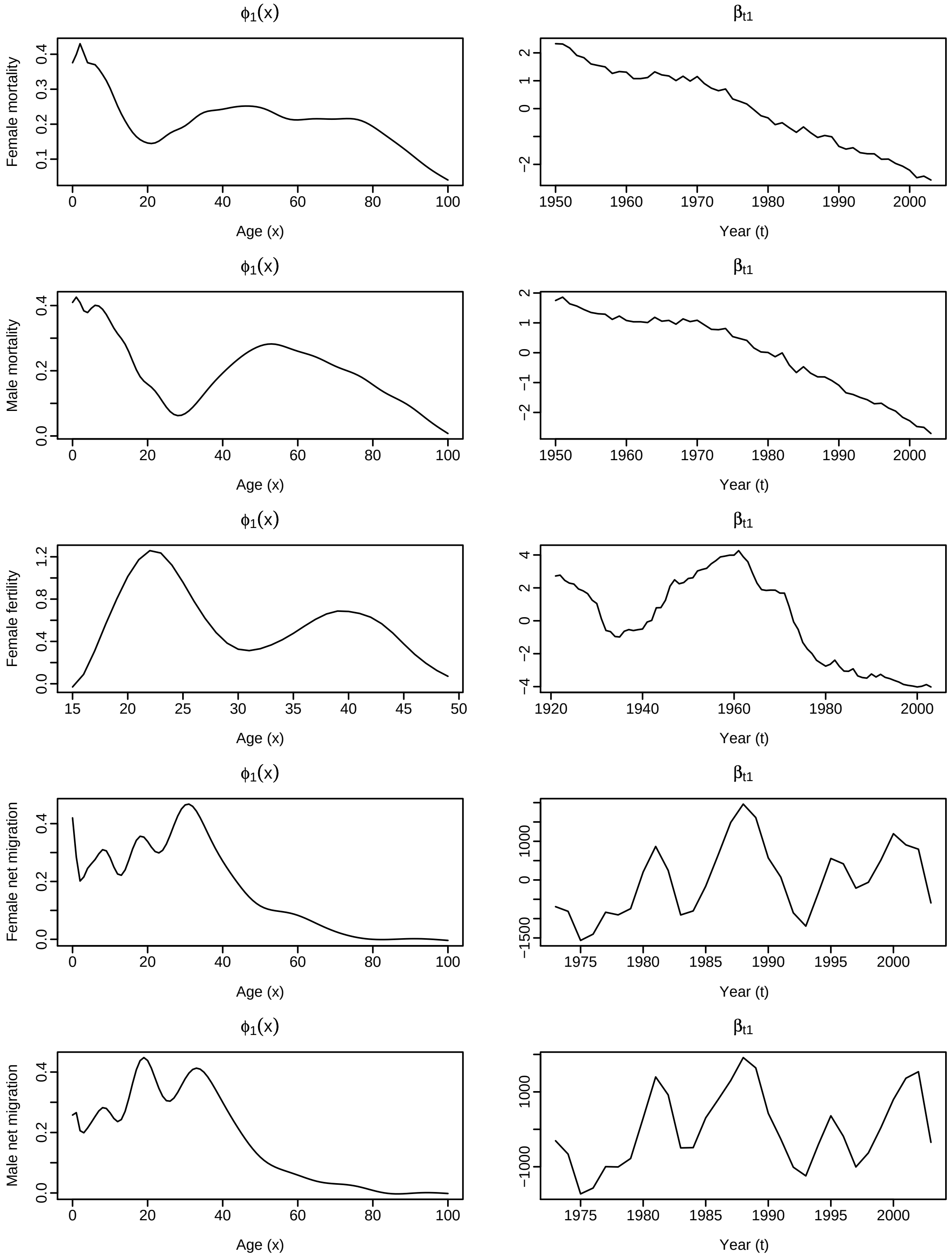

Figure 2: *First basis function and coefficient for each component of the Australian data.*

likely to be the result of bias; Wilson and Bell (2004) also found that the ERP inflates the population at ages 90+. The high volatility in estimated net migration between birth and age 0 on 1 January of the following year is likely to be an artifact of ABS estimation procedures, rather than a reflection of true variability in migration at this age.

While there is a high degree of variation, an age pattern can be readily discerned. The male and female patterns are similar. The mean suggests that the level of female net migration is higher than that for males at young adult ages; this is a result of male-dominated emigration rather than of female-dominated immigration. The number of children is also relatively high, a result of Australian migration policies concerning family migration. For both males and females, the first peak at young adult ages (at age 19) is due to overseas students, while the second (at age 29 for males and 26 for females) is due to labour migration and spouses. A further retirement-related peak occurs at around age 60. Compared with the labour-dominated migration model of Rogers and Castro (1981) (see also Rogers et al., 2005), these distributions are relatively flat across the age range.

Figure 2 shows the first basis function and first coefficient for all five models. The first term accounts for at least 94 percent of the variation in mortality, 66 percent for fertility and 51–60 percent for net migration. Taking female mortality as an example, the coefficient indicates a fairly steady decline in mortality over time, while the basis function indicates that the decline has been faster at very young ages and at ages 40–80. Figure 3 shows a further two terms for this model, which modify the dominant first term (the mean age pattern is also shown).

Figure 4 shows forecast log death rates by sex for the first and last years of the forecast period (2004 and 2023) with 80% prediction intervals. Also shown are observed rates for 2003. The forecast decline in male mortality is less rapid at young adult ages (about 30) than at older ages (40–80) reflecting recent trends.

Forecast fertility rates are shown in Figure 5. The relatively recent change of second derivative at ages 19 and 20, seen in the 2003 data, is a more marked feature in the forecast resulting in particularly low rates in the early to mid-twenties. While such a marked age pattern of fertility may at first appear extreme, it is not without precedent: an even more marked effect has been observed in Ireland (Chandola et al., 1999). This feature is emerging in several populations, including the United Kingdom (Chandola et al., 1999) and Germany (Betz and Lipps, 2004). It is likely to arise from heterogeneity with respect to marital status (as Chandola et al. suggest)

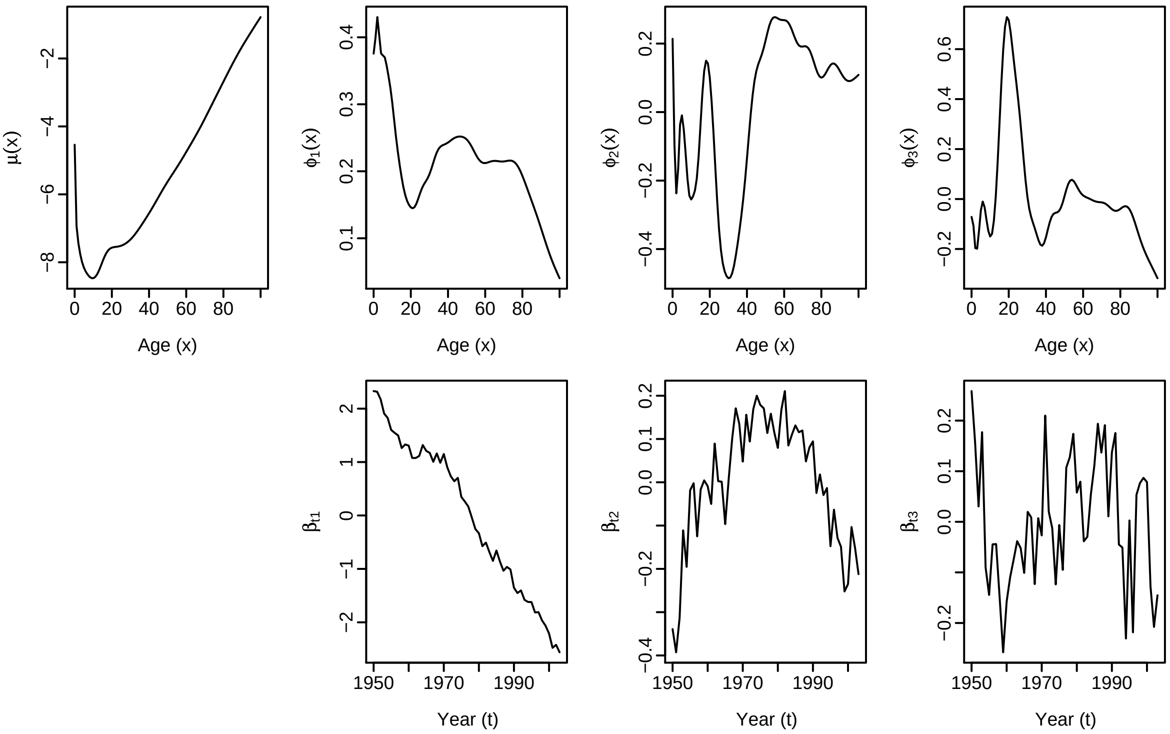

Figure 3: *Fitted basis functions and coefficients for Australian female log death rates.*

or socio-economic circumstance: one group experiencing early fertility which will decline only slowly (possibly towards a threshold beyond which rates will not decline) and another group experiencing a rapid decline in early fertility. However, the message contained in the very wide prediction intervals in 2023 should not be ignored.

Net migration forecasts are shown in Figure 6. The volatility in the data is reflected in the wide prediction interval which swamps the point forecast. The point forecasts are identical for every year of the forecast, as the point forecasts of all coefficients are constant over the forecast horizon. The prediction intervals increase with the forecast horizon. The forecast annual net migration is 42,700 females and 44,200 males. The sex ratio reflects that for recent estimates.

Forecast life expectancy by sex is shown in Figure 7. By 2023, forecast values are 86.2 years with an 80% prediction interval of 83.8 to 88.4 for females, and 82.7 years with an interval of 80.3 to 85.2 for males. Male life expectancy increases at a faster rate than female, a reflection of recent trends. While this is defensible for the 20-year forecast duration, it would lead to divergence in the longer term. Non-divergent models will be explored in future research. Forecast total fertility (Figure 8) shows a slight upward trend: by 2023 total fertility is 1.79, compared

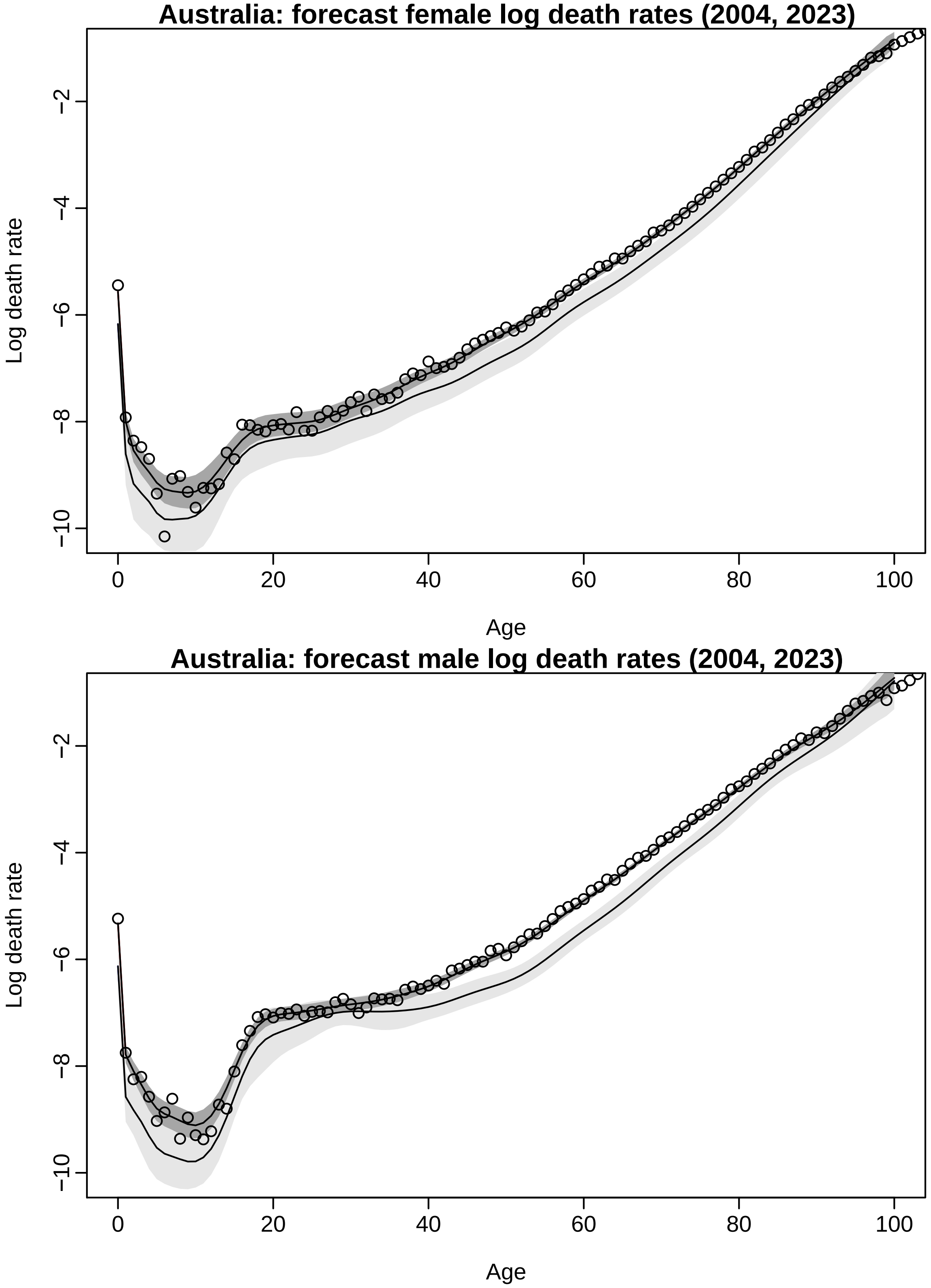

Figure 4: *Forecast log death rates for 2004 and 2023, along with 80% prediction intervals. Actual mortality rates for 2003 are also shown as circles.*

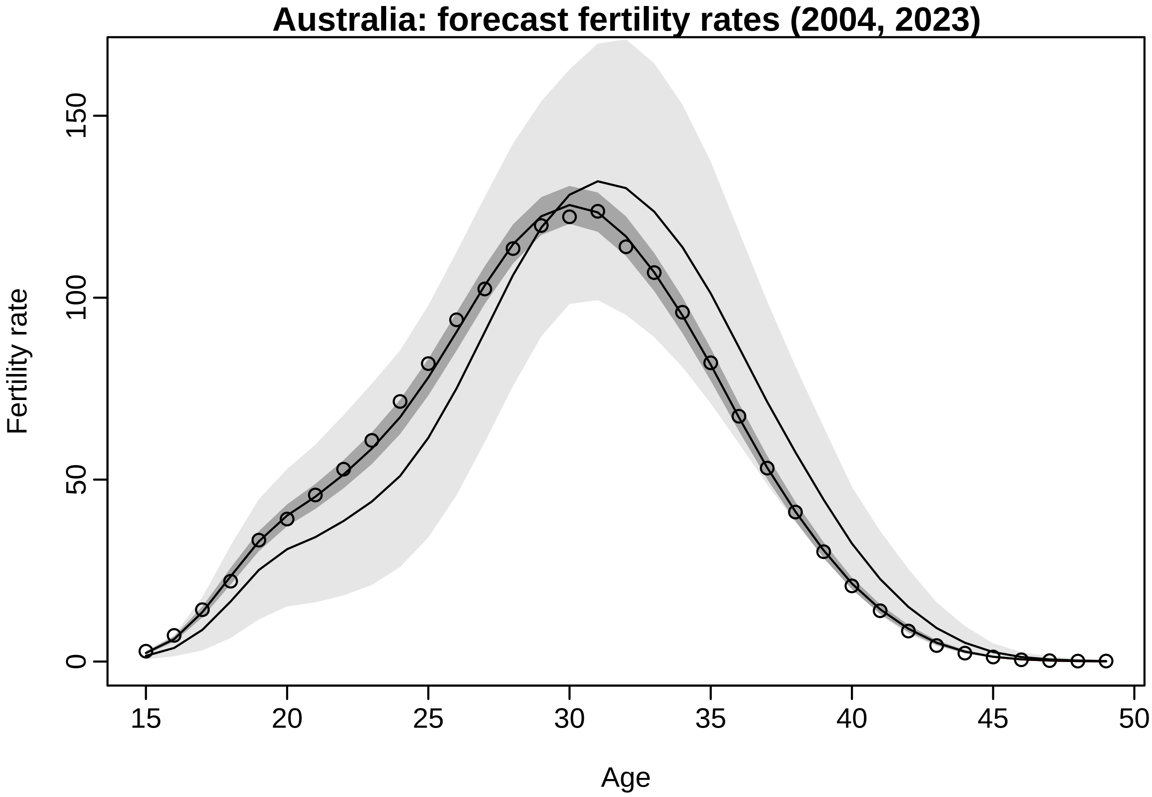

Figure 5: *Forecast fertility rates for 2004 and 2023, along with 80% prediction intervals. Actual fertility rates for 2003 are also shown as circles.*

with 1.75 in 2003. The prediction interval of 1.18 to 2.55 reflects the large uncertainty in forecasting fertility, and shows that this upward trend is not statistically significant.

The final population forecasts for males and females in 2023 are shown as a population pyramid with 80% prediction intervals in Figure 9, along with the 2003 base population. These are based on 10000 simulated sample paths. The broad prediction intervals at young ages reflect the greater uncertainty in forecast fertility, while uncertainty at 20–50 is largely due to migration.

Our forecasts are compared in Figure 10 with the population projections given in Australian Bureau of Statistics (2003). For both sexes, the forecast mean is close to the middle ABS projection (Series B) suggesting that ABS Series B provides a reasonably good point forecast. However, the high ABS projection (Series A) lies close to the upper 80% prediction limit so that there is a probability of about 20% of the population being greater than series A. The low ABS projection (series C) lies between the 80% and 95% lower prediction limits; thus, there is between 5% and 20% chance of the population being lower than series C.

Most users tend to interpret series A and series C as something like a prediction interval. Yet this

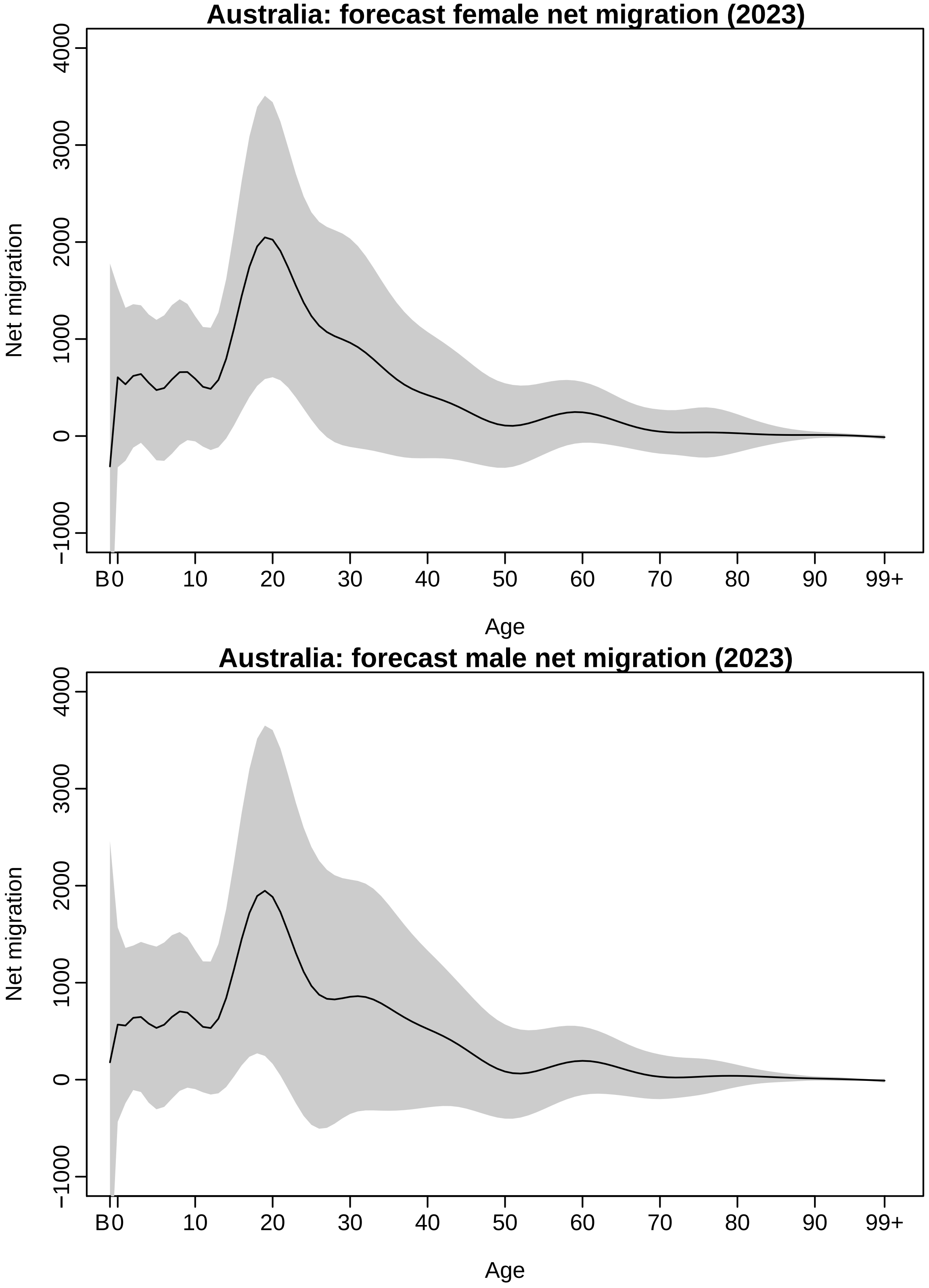

Figure 6: *Forecast net migration numbers for 2023, along with 80% prediction intervals. All other years are almost identical due to the low correlation between years.*

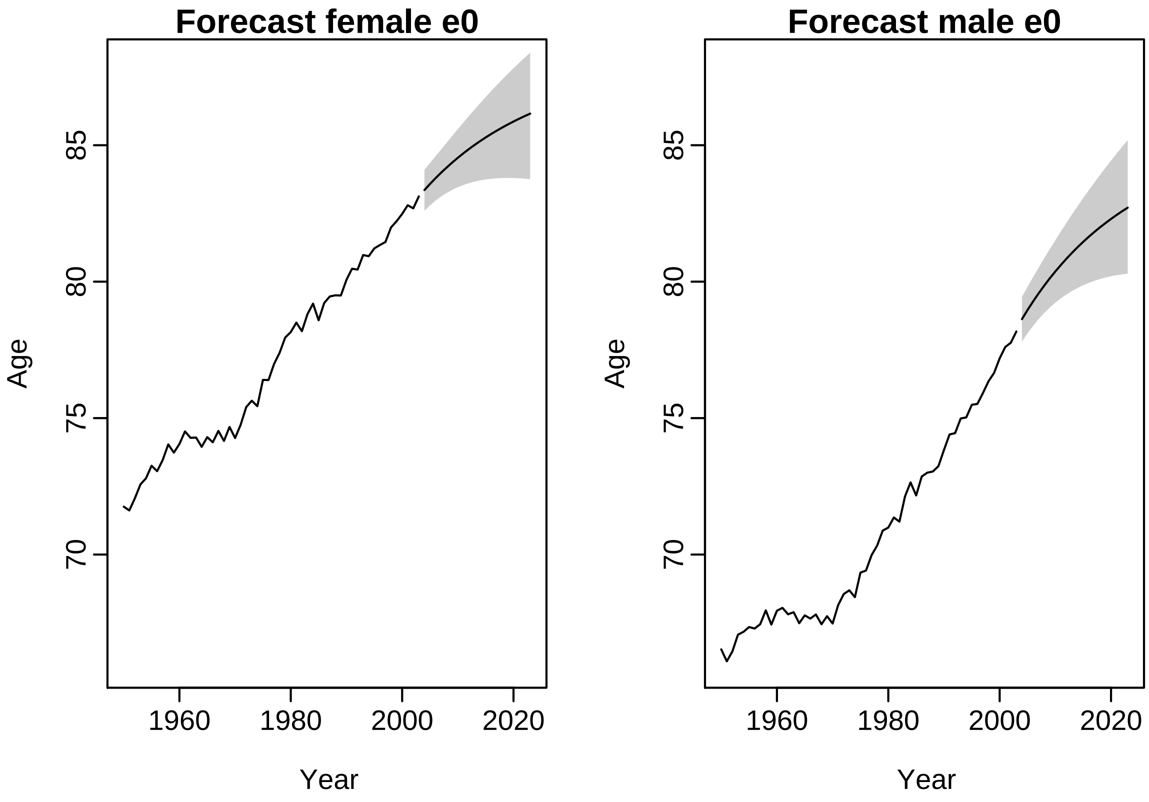

Figure 7: *Forecasts of e_0 (life expectancy at age 0) for 2004–2023.*

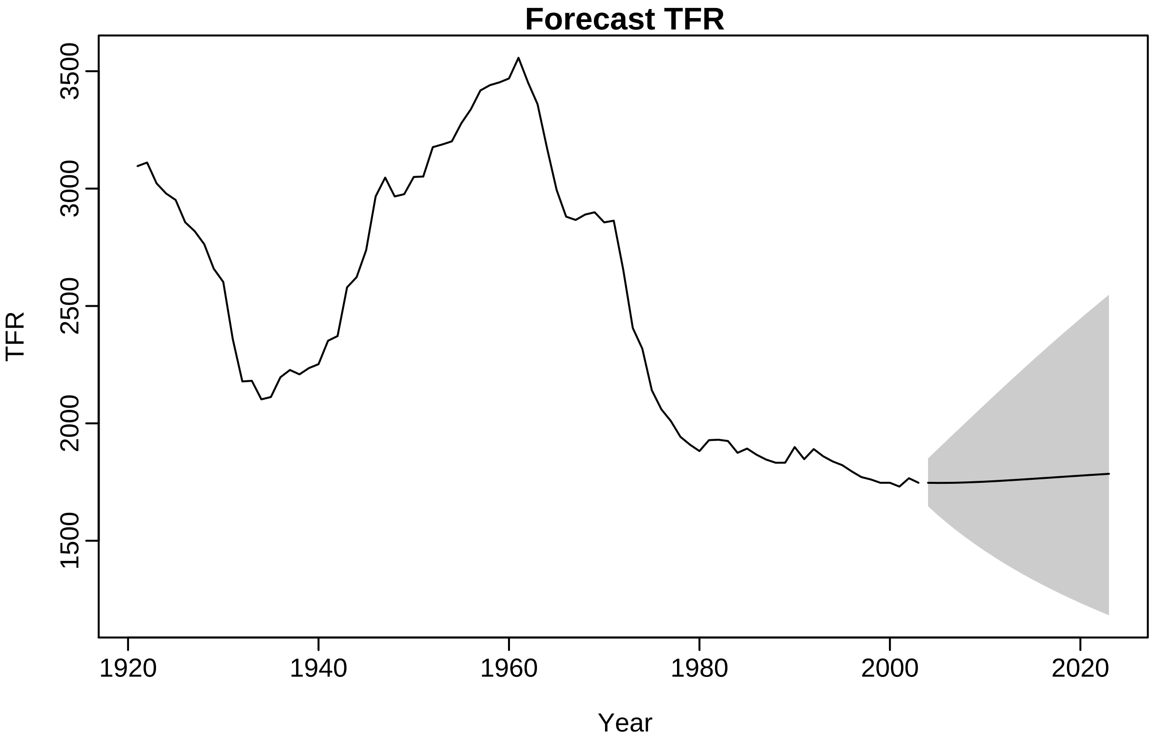

Figure 8: *Forecasts of total fertility rate (TFR) for 2004–2023.*

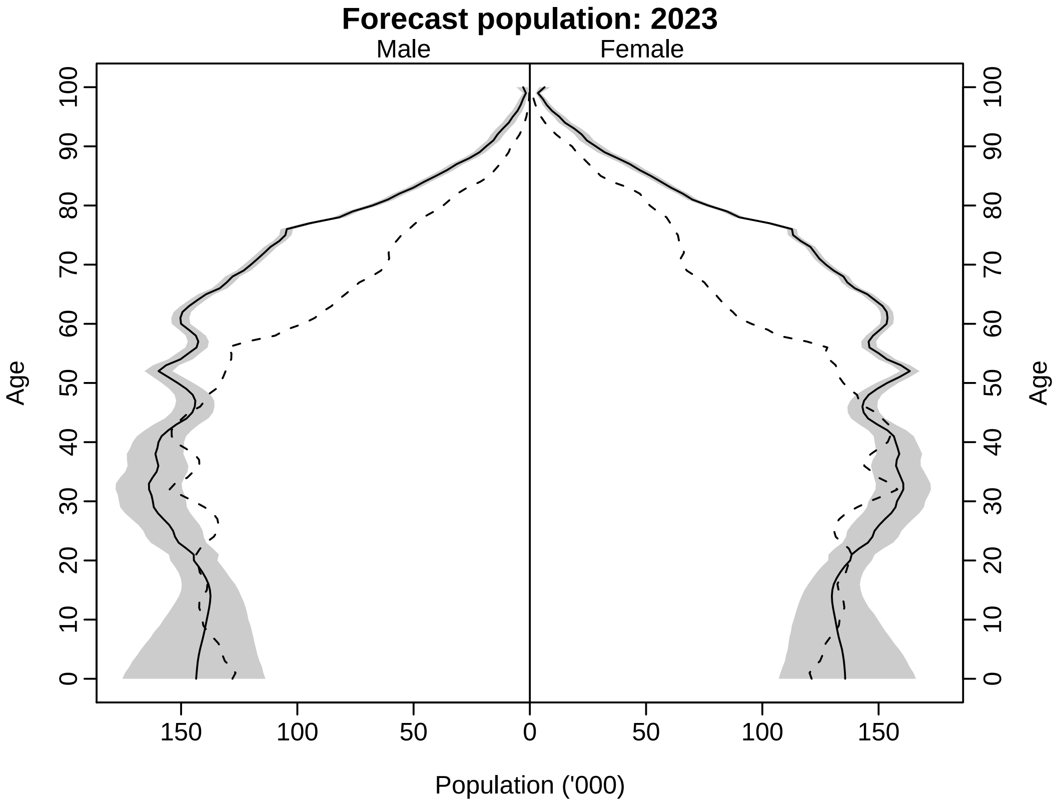

Figure 9: *Forecast population pyramid for 2023, along with 80% prediction intervals. The actual population pyramid for 2003 is shown using dashed lines.*

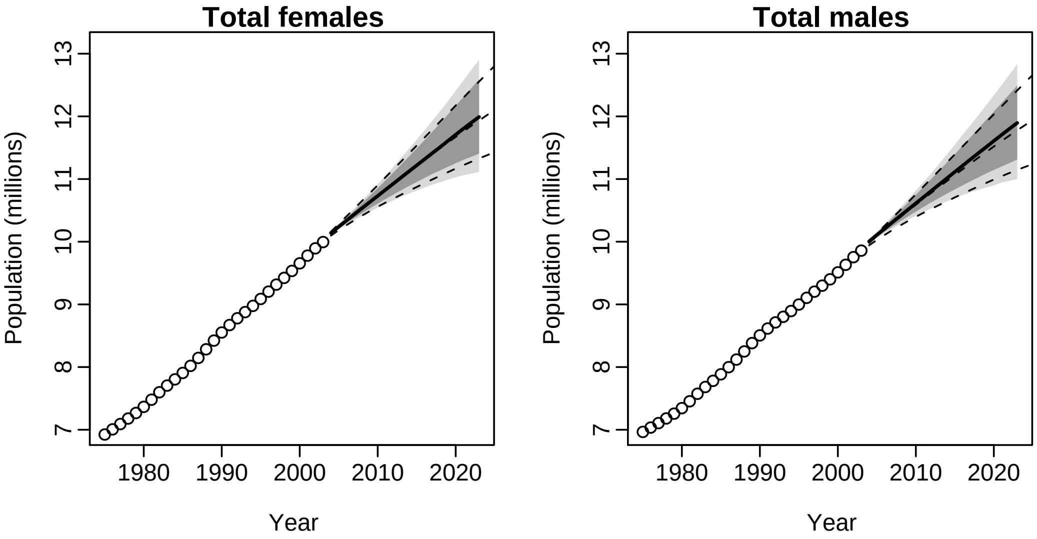

Figure 10: *Twenty-year forecasts of total population for each sex along with 80% and 95% prediction intervals. The dashed lines show the projections from Australian Bureau of Statistics (2003), series A, B and C.*

analysis shows that the probability of being above and below series A and series C is unequal, and thus the series are potentially misleading.

6 Comments and conclusions

The above analysis has demonstrated that functional data models can be successfully applied to forecasting mortality rates, fertility rates and migration numbers. This modelling framework is highly adaptable. The same basic model applies, allowing for the different characteristics of the three demographic components through the use of different transformations. The Box-Cox transformation has proved useful. Though this transformation is commonly used in statistics, it has rarely been applied in demographic modeling. For mortality, the preferred transform is the logarithm, which coincides with convention. For fertility, a slightly weaker transform ($\lambda = 0.2$) is preferred; this use of the Box-Cox transformation is a substantial improvement on previous research where it has been necessary to impose judgmental limits in order to constrain the prediction intervals to plausible values.

The forecasts produced here are based on the first six principal components. For mortality, this is the main difference between our method and the Lee-Carter method (which additionally involves an adjustment). The extra principal components allow more accurate forecasting of age-specific rates (Hyndman and Ullah, 2007), though typically at least 90% of the variation is explained by the first component. The use of several components is more important for fertility, where the first component explains a smaller proportion of the variation (69% in the current example). The additional components may serve to incorporate changes in pattern that are relatively recent. For fertility, the emergence of relatively low rates in the early to mid-twenties is a case in point; this pattern has been forecast to evolve. The somewhat irregular age pattern in forecast male mortality also results from the combination of principal components reflecting recent trends; further research is needed to impose greater regularity. Nevertheless, the forecast age pattern is relatively smooth compared with the jagged age distribution of forecasts produced by the Lee-Carter method (Girosi and King, 2006).

For net migration, the numbers by age and sex estimated from the growth-balance equation include net errors in the recorded population and vital events. Comparison with total net recorded migration for Australia from 1989 indicates that such errors are small, especially in the very recent past. Further, errors in the age distribution arising from digit preference and other biases are removed by smoothing. Thus, while forecast net migration will include bias or systematic error not removed by smoothing, such error is unlikely to be significant. For purposes of forecasting population, the inclusion of such error will not bias the forecast if past errors can be assumed to be constant over time. In countries with high levels of illegal immigration, where the estimation of net age-sex-specific migration is highly preferable to the use of recorded data, this assumption may not be appropriate and care should be exercised in forecasting. For example, if population coverage has increased over time, the trend in net migration will be overestimated.

In the Australian case, the age-sex-specific net migration forecast is a simple constant. Though the level of migration has fluctuated over the last 30 years, there is no discernible trend; this can be attributed to the general stability of migration policies and economic factors. In countries where such factors have been less stable, for example Spain which has changed from a country of emigration to a country of immigration in the last few decades, the forecasting of net migration will undoubtedly be more difficult.

As noted above, the estimation of the uncertainty of forecast demographic processes presents considerable difficulties because estimates vary depending on the method used. In this paper, we have adjusted model-based estimates of the variance to ensure the model one-step forecast variance is equal to the historical empirical one-step forecast variance. In the Australian case, this calibration increases the model variance for some ages and decreases it for others. For mortality rates, the adjustment reduces variances except at very old ages; for fertility rates, the adjustment tends to increase variances between ages 25 and 35 and decrease variances between ages 35 and above. The adjustment has little effect on migration variances.

In population forecasting, several sources of uncertainty need to be taken into account. Our proposed method has accounted for all of the major sources of variation: the observational variation in the (Poisson) generation of births and deaths, the binomial variation in the sex ratio at birth, the observational variation in net migration, and the dynamic variation as the rates of mortality, fertility and migration change over time. This constitutes a more complete representation of uncertainty than most, if not all, other approaches.

In population forecasts, correlations among forecast errors in the demographic components affect the width of the prediction interval (Lee, 1999). The correlations are taken into account in our model through the smooth curves across age and the time series models handling temporal correlations. These are then naturally incorporated into the simulated sample paths of each demographic component. However, we have not taken account of any correlation between the demographic components. The assumed independence among the three demographic components is a reasonable first approximation for developed countries (e.g., Alho, 1992).

This application of functional data models to fully stochastic population forecasting is a first exercise using data that are of good quality and well behaved. Extensions to this research include extension to other countries with different population histories, especially with respect to migration. The dependencies between male and female mortality and between male and female migration also need to be addressed, through, for example, restriction to non-divergent futures (see for example Li and Lee, 2005). Research into the demographic interpretation of the main principal components with a view to constraining forecast coefficients may also improve forecasting accuracy. Other possible extensions include taking account of cohort effects, modeling fertility by parity and modeling interactions between components, such as (past) fertility and (current) immigration.

Although the methods might seem relatively complicated, their implementation is relatively easy using the demography package for R (Hyndman, 2006), provided the historical births, deaths and population data are available in the same form as is used in Human Mortality Database (2006). Instructions on the use of the package are provided at `www.robhyndman.info/Rlibrary/demography`.

Appendix: Population simulation equations

Notation

$$\begin{aligned}
P_t(x) &= \text{population of age } x \text{ on 1 January in year } t. \\
E_t(x) &= \text{mid-year (30 June) population} \\
D_t(x) &= \text{deaths at age } x \text{ in year } t \\
D_t(x,x+1) &= \text{cohort deaths in calendar year } t \text{ of persons aged } x \text{ at the beginning of year } t. \\
D_t(B,0) &= \text{cohort deaths in calendar year } t \text{ of births during year } t. \\
B_t(x) &= \text{births in year } t \text{ to women of age } x \\
G_t(x,x+1) &= \text{net migration for age } x \text{ at the beginning of year } t; \\
G_t(B,0) &= \text{net migration for births during year } t;
\end{aligned}$$

$$\begin{aligned}
&m_t(x) = D_t(x)/E_t(x) && = \text{central death rate for age } x \text{ in year } t; \\
&f_t(x) = B_t(x)/E_t^F(x) && = \text{fertility rate for mothers of age } x \text{ in year } t. \\
&r_t(x) && = \text{life table survivorship ratios computed from } m_t(x). \\
&R_t(x) && = \text{population adjusted for first half of migration}
\end{aligned}$$

$\rho =$ male:female sex-ratio at birth.

Use of $\bar{A}$ indicates an estimate of A.

Use of x^+ indicates the open-ended age group x and above.

Age groups: $x = 0, 1, 2, \ldots p-1, p^+$, where p is the lower limit of the (upper) open-ended age group.

Algorithm to get $P_{t+1}(x)$ given $P_t(x)$, $m_t(x)$, $f_t(x)$ and $G_t(x,x+1)$

$$\begin{aligned}
R_t(x) &= P_t(x) + G_t(x,x+1)/2, && x = 0,\ldots,p-2 \\
R_t(p-1) &= P_t(p-1) + G_t(p-1^+, p^+)/4 \\
R_t(p^+) &= P_t(p^+) + G_t(p-1^+, p^+)/4 \\
\bar{D}_t(x,x+1) &= r_t(x) R_t(x) && x = 0,\ldots,p-2 \\
\bar{D}_t(p-1^+, p^+) &= r_t(p-1^+)[R_t(p-1) + R_t(p^+)]
\end{aligned}$$

$$
\begin{aligned}
\bar{R}_{t+1}(x+1) &= R_t(x) - \bar{D}_t(x,x+1) && x=0,\dots,p-2\\
\bar{R}_{t+1}(p^+) &= R_t(p-1) + R_t(p^+) - \bar{D}_t(p-1^+,p^+)\\
\bar{E}_t(x) &= \left[R_t(x) + \bar{R}_{t+1}(x)\right]/2, && x=1,\dots,p-1,p^+\\
\bar{D}_t(x) &= m_t(x)\bar{E}_t(x) && x=1,\dots,p-1,p^+\\
D_t(x) &\sim \text{Poisson}(\bar{D}_t(x)) && x=1,\dots,p-1,p^+\\
D_t(x,x+1) &= \left[D_t(x) + D_t(x+1)\right]/2, && x=1,\dots,p-2\\
D_t(p-1^+,p^+) &= \Big[D_t(p-1)/2 + D_t(p^+)\Big],\\
R_{t+1}(x+1) &= R_t(x) - D_t(x,x+1) && x=1,\dots,p-2\\
R_{t+1}(p^+) &= R_t(p-1) + R_t(p^+) - D_t(p-1^+,p^+)\\
B_t(x) &\sim \text{Poisson}(f_t(x)[R_t(x)+R_{t+1}(x)]/2), && x=15,\dots,49\\
B_t &= \sum_{x=15}^{49} B_t(x)\\
B_t^M &\sim \text{Binomial}\left(B_t, \rho/(\rho+1)\right)\\
B_T^F &= B_t - B_t^M\\
R_t(B) &= B_t + G_t(B,0)/2\\
\bar{D}_t(B,0) &= r_t(B)R_t(B)\\
\bar{R}_{t+1}(0) &= R_t(B) - \bar{D}_t(B,0)\\
\bar{E}_t(0) &= \left[R_t(0) + \bar{R}_{t+1}(0)\right]/2\\
\bar{D}_t(0) &= m_t(0)\bar{E}_t(0)\\
D_t(0) &\sim \text{Poisson}(\bar{D}_t(0))\\
f_0 &= \bar{D}_t(B,0)/\bar{D}_t(0)\\
D_t(B,0) &= f_0 D_t(0)\\
D_t(0,1) &= (1-f_0)D_t(0) + D_t(1)/2\\
R_{t+1}(0) &= R_t(B) - D_t(B,0)\\
R_{t+1}(1) &= R_t(0) - D_t(0,1)\\
P_{t+1}(0) &= R_{t+1}(0) + G_t(B,0)/2\\
P_{t+1}(x+1) &= R_{t+1}(x+1) + G_t(x,x+1)/2, && x=0,\dots,p-1,p^+
\end{aligned}
$$

References

Alho, J. M. (1992) Population forecasting theory, methods and assessments of accuracy: the magnitude of error due to different vital processes in population forecasts, *International Journal of Forecasting*, **8**, 301–314.

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

Alho, J. M. and B. D. Spencer (1997) The practical specification of the expected error of population forecasts, *Journal of Official Statistics*, **13**(3), 203–225.

Australian Bureau of Statistics (2003) *Population projections, Australia: 2002–2101*, Catalogue No. 3222.0, Australian Bureau of Statistics, Canberra.

Betz, F. and O. Lipps (2004) Stochastic population projection for Germany based on the QS-approach to modelling age-specific fertility rates, Tech. Rep. 59-2004, Mannheim Research Institute for the Economics of Aging.

Booth, H. (2006) Demographic forecasting: 1980 to 2005 in review, *International Journal of Forecasting*, **22**(3), 547–581.

Booth, H., R. J. Hyndman, L. Tickle and P. de Jong (2006) Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions, *Demographic Research*, **15**(9), 289–310.

Booth, H., J. Maindonald and L. Smith (2002) Applying Lee-Carter under conditions of variable mortality decline, *Population Studies*, **56**(3), 325–336.

Booth, H., L. Tickle and L. Smith (2005) Evaluation of the variants of the Lee-Carter method of forecasting mortality: a multi-country comparison, *New Zealand Population Review*, **31**(1), 13–34.

Box, G. E. P. and D. R. Cox (1964) An analysis of transformations, *Journal of the Royal Statistical Society, Series B*, **26**(2), 211–252.

Bozik, J. E. and W. R. Bell (1987) Forecasting age-specific fertility using principal components, in *Proceedings of the American Statistical Association, Social Statistics Section*, pp. 396–401, San Francisco, California.

Brillinger, D. R. (1986) The natural variability of vital rates and associated statistics, *Biometrics*, **42**, 693–734.

Brouhns, N., M. Denuit and J. K. Vermunt (2002) A Poisson log-bilinear regression approach to the construction of projected lifetables, *Insurance: Mathematics and Economics*, **31**(3), 373–393.

Chandola, T., D. A. Coleman and R. W. Hiorns (1999) Recent European fertility patterns: fitting curves to distorted distributions, *Population Studies*, **53**, 317– 329.

Congdon, P. (1993) Statistical graduation in local demographic analysis and projection, *Journal of the Royal Statistical Society, Series A*, **156**, 237–270.

De Beer, J. (1997) The effect of uncertainty of migration on national population forecasts: the case of the Netherlands, *Journal of Official Statistics*, **13**, 227–243.

De Jong, P. and L. Tickle (2006) Extending Lee-Carter mortality forecasting, *Mathematical Population Studies*, **13**(1), 1–18.

Erbas, B., R. J. Hyndman and D. M. Gertig (2007) Forecasting age-specific breast cancer mortality using functional data models, *Statistics in Medicine*, **26**(2), 458–470.

Forfar, D. O. and D. M. Smith (1987) The changing shape of English life tables, *Transactions of the Faculty of Actuaries*, **40**, 98–133.

George, M. V. (1994) *Population projections for Canada, provinces and territories: 1993–2016*, Statistics Canada, Ottawa.

George, M. V. and J. Perreault (1992) Methods of external migration projections and forecasts, in N. Keilman and H. Cruijsen (eds.) *National population forecasting in industrialized countries*, pp. 87–103, Swets and Zeitlinger, Amsterdam.

Girosi, F. and G. King (2006) *Demographic forecasting*, Cambridge University Press, Cambridge.

Heligman, L. and J. H. Pollard (1980) The age pattern of mortality, *Journal of the Institute of Actuaries*, **107**, 49–80.

Hilderink, H., N. Van der Gaag, L. Van Wissen, R. Jennissen, A. Román, J. Salt, J. Clarke and C. Pinkerton (2002) Analysis and forecasting of international migration by major groups. Part III, working papers and studies 3/2002/E/n17, Eurostat, The Hague.

Hoem, J. M., D. Madsen, J. L. Nielsen, E. M. Ohlsen, H. O. Hansen and B. Rennermalm (1981) Experiments in modelling recent Danish fertility curves, *Demography*, **18**(2), 231–244.

Human Mortality Database (2006) University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany), downloaded on 1 May 2006. **URL:** *www.mortality.org*

Hyndman, R. J. (2006) *demography: Forecasting mortality and fertility data*, R package version 0.97. **URL:** *http://www.robhyndman.info/Rlibrary/demography*

Hyndman, R. J., A. B. Koehler, R. D. Snyder and S. Grose (2002) A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting*, **18**(3), 439–454.

Hyndman, R. J. and M. S. Ullah (2007) Robust forecasting of mortality and fertility rates: a functional data approach, *Computational Statistics & Data Analysis*, **51**, 4942–4956.

Keilman, N. (1997) Ex-post errors in official population forecasts in industrialized countries, *Journal of Official Statistics*, **13**(3), 245–277.

Keilman, N. (2001) Uncertain population forecasts, *Nature*, **412**, 490–491.

Keilman, N. and D. Q. Pham (2000) Predictive intervals for age-specific fertility, *European Journal of Population*, **16**, 41–66.

Keilman, N. and D. Q. Pham (2004) Empirical errors and predicted errors in fertility, mortality and migration forecasts in the European economic area, Discussion Papers 386, Research Department of Statistics Norway. **URL:** *http://www.ssb.no/publikasjoner/DP/pdf/dp386.pdf*

Keilman, N., D. Q. Pham and A. Hetland (2002) Why population forecasts should be probabilistic — illustrated by the case of Norway, *Demographic Research*, **6**(Article 15).

Khoo, S.-E. and P. McDonald (2002) Adjusting for change of status in international migration: demographic implications, *International Migration*, **40**(4), 103–123.

Knudsen, C., R. McNown and A. Rogers (1993) Forecasting fertility: an application of time series methods to parameterized model schedules, *Social Science Research*, **22**, 1 –23.

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

Lee, R. D. (1993) Modeling and forecasting the time series of U.S. fertility: age distribution, range, and ultimate level, *International Journal of Forecasting*, **9**, 187–202.

Lee, R. D. (1999) Probabilistic approaches to population forecasting, in W. Lutz, J. W. Vaupel and D. A. Ahlburg (eds.) *Frontiers of population forecasting*, Supplement to *Population and development review*, **24**, pp. 156–190, Population Council, New York.

Lee, R. D. and L. R. Carter (1992) Modeling and forecasting U.S. mortality, *Journal of the American Statistical Association*, **87**, 659–675.

Lee, R. D. and T. Miller (2001) Evaluating the performance of the Lee-Carter method for forecasting mortality, *Demography*, **38**(4), 537–549.

Lee, R. D. and S. Tuljapurkar (1994) Stochastic population forecasts for the United States: Beyond high, medium, and low, *Journal of the American Statistical Association*, **89**, 1175–1189.

Li, N. and R. D. Lee (2005) Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method, *Demography*, **42**(3), 575–594.

Lutz, W., W. Sanderson, S. Scherbov and A. Goujon (1996) World population scenarios for the 21st century, in W. Lutz (ed.) *The future population of the world: what can we assume today?*, pp. 361–396, Earthscan, London.

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.

McDonald, J. (1984) The emergence of countercyclical US fertility: A reassessment of the evidence, *Journal of Macroeconomics*, **5**, 421–436.

McNown, R. and A. Rogers (1989) Forecasting mortality: A parameterized time series approach, *Demography*, **26**(4), 645–660.

McNown, R., A. Rogers and J. Little (1995) Simplicity and complexity in extrapolative population forecasting models, *Mathematical Population Studies*, **5**(3), 235–257.

Miller, R. B. (1986) A bivariate model for total fertility rate and mean age of childbearing, *Insurance: Mathematics and Economics*, **5**, 133–140.

Miller, T. (2003) California's uncertain population future, technical appendix, in *Special Report: the growth and aging of California's population: demographic and fiscal projections, characteristics and service needs* by R. D. Lee, T. Miller and R. Edwards, Center for the Economics and Demography of Aging. CEDA Papers: Paper 2003-0002CL, URL: *http://repositories.cdlib.org/iber/ceda/papers/2003-0002CL/*.

Miller, T. and R. D. Lee (2004) A probabilistic forecast of net migration to the United States. Report I in *Stochastic infinite horizon forecasts for social security and related studies*, Working paper 10917, National Bureau of Economic Research. **URL:** *http://www.nber.org/papers/w10917*

Ortega, J. A. and P. Poncela (2005) Joint forecasts of southern European fertility rates with non-stationary dynamic factor models, *International Journal of Forecasting*, **21**(3), 539–550.

Pollard, J. H. (1987) Projection of age-specific mortality rates, *Population Bulletin of the United Nations*, **21-22**, 55–69.

Preston, S. H., P. Heuveline and M. Guillot (2001) *Demography: measuring and modelling population processes*, Blackwell, Oxford.

Renshaw, A. E. and S. Haberman (2003a) On the forecasting of mortality reduction factors, *Insurance: Mathematics and Economics*, **32**(3), 379–401.

Renshaw, A. E. and S. Haberman (2003b) Lee-Carter mortality forecasting with age-specific enhancement, *Insurance: Mathematics and Economics*, **33**(2), 255–272.

Renshaw, A. E. and S. Haberman (2003c) Lee-Carter mortality forecasting: a parallel generalized linear modelling approach for England and Wales mortality projections, *Applied Statistics*, **52**(1), 119–137.

Rogers, A. (1990) Requiem for the net migrant, *Geographical Analysis*, **22**, 283–300.

Rogers, A. and L. Castro (1981) Model migration schedules, Research Report 81-30, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Rogers, A., L. J. Castro and M. Lea (2005) Model migration schedules: Three alternative linear parameter estimation methods, *Mathematical Population Studies*, **12**, 17–38.

Rogers, A. and J. S. Little (1994) Parameterizing age patterns of demographic rates with the multiexponential model schedule, *Mathematical Population Studies*, **4**, 175–195.

Saboia, J. L. M. (1977) Autoregressive Integrated Moving Average (ARIMA) models for birth forecasting, *Journal of the American Statistical Association*, **72**, 264–270.

Sykes, Z. M. (1969) Some stochastic versions of the matrix model for population dynamics, *Journal of the American Statistical Association*, **44**, 111–130.

Thompson, P. A., W. R. Bell, J. F. Long and R. B. Miller (1989) Multivariate time series projections of parameterized age-specific fertility, *Journal of the American Statistical Association*, **84**(407), 689–699.

Tuljapurkar, S., N. Li and C. Boe (2000) A universal pattern of mortality decline in the G7 countries, *Nature*, **405**, 789–792.

United Nations (2001) Replacement migration: is it a solution to declining and aging populations?, No. ST/ESA/SER.A/206, United Nations, New York.

Wilmoth, J. R. (1993) Computational methods for fitting and extrapolating the Lee-Carter model of mortality change, Technical report, Department of Demography, University of California, Berkeley. **URL:** *http://www.demog.berkeley.edu/ jrw/Papers/LCtech.pdf*

Wilson, T. and M. Bell (2004) Australia’s uncertain demographic future, *Demographic Research*, **11**(8), 195–234.