A Bayesian forecasting model: predicting U.S. male mortality

CLAUDIA PEDROZA

Division of Biostatistics, University of Texas School of Public Health at Houston,
1200 Herman Pressler Drive, E831, Houston, TX 77030, USA
claudia.pedroza@uth.tmc.edu

SUMMARY

This article presents a Bayesian approach to forecast mortality rates. This approach formalizes the Lee–Carter method as a statistical model accounting for all sources of variability. Markov chain Monte Carlo methods are used to fit the model and to sample from the posterior predictive distribution. This paper also shows how multiple imputations can be readily incorporated into the model to handle missing data and presents some possible extensions to the model. The methodology is applied to U.S. male mortality data. Mortality rate forecasts are formed for the period 1990–1999 based on data from 1959–1989. These forecasts are compared to the actual observed values. Results from the forecasts show the Bayesian prediction intervals to be appropriately wider than those obtained from the Lee–Carter method, correctly incorporating all known sources of variability. An extension to the model is also presented and the resulting forecast variability appears better suited to the observed data.

Keywords: Bayesian prediction; Lee–Carter method; Missing data; Mortality forecasting; State-space model.

1. INTRODUCTION

Government agencies use mortality forecasts when planning and developing health policy. These agencies rely on mortality predictions in deciding the allocation of funds for government services. In the past, researchers have based many of these projections on life table methods and expert opinion. Recent interest centers on improving mortality forecasts, driven in part by an increasingly elderly population's effect on government policy (such as the U.S. Social Security). Private industries, such as insurance companies, also rely on these forecasts to plan for future programs. Thus, it is in the interest of public health to produce improved mortality forecasts.

Alho and Spencer (1990) among others argue that measures of uncertainty of mortality forecasts are also needed. Since a majority of the methods used to forecast were not probability based, calls for measures of uncertainty have gone unheeded. This has made the task of creating prediction intervals an almost impossible endeavor, leading to less informative procedures, e.g. reporting of a point forecast and a high and a low forecast. However, the high and low forecasts are worst and best case scenarios, not intervals expressing probabilistic uncertainty of the point forecast.

Over the past two decades, various methods have been developed to improve mortality forecasts. This paper focuses on an increasingly popular method of forecasting introduced by Lee and Carter (1992)

to model and forecast U.S. mortality. Widely used, their method has undergone various extensions and modifications (Carter and Lee, 1992; Lee and Tuljapurkar, 1994; Wilmoth, 1995; Carter, 1995). In their original paper, Lee and Carter discuss the computation of prediction error for the mortality forecasts. However, this prediction error does not account for the estimation error of the parameters involved in the model, instead only incorporating uncertainty from the forecast of the mortality index. The omission of the estimation error leads to prediction intervals which are too erroneously narrow.

Recently, Brouhns *et al.* (2002, 2005) implemented a bootstrap procedure for a Poisson log-bilinear formulation of the Lee–Carter model. This procedure accounts for all sources of variability in the Poisson model. Czado *et al.* (2005) present a Bayesian formulation of the log-bilinear Poisson which also incorporates the different sources of error in the model when forming forecasts and prediction intervals. A state-space model formulation of the Lee–Carter method is presented in Pedroza (2002), as well as various Bayesian time series models for forecasting mortality rates. Girosi and King (2005) present a Bayesian hierarchical modeling approach for predicting mortality rates which seeks to pool information from similar cross-sections (i.e. age groups, countries).

Here, the original Lee–Carter method is restated under a Bayesian framework. With this approach, missing data can be handled and sampling error is automatically incorporated within the model and its mortality forecasts. Extensions to the model can also be easily incorporated.

The outline of this paper is as follows: Section 2 presents a brief description of 20th century trends in mortality for developed countries, with the U.S. as an example. Section 3 describes Lee and Carter's method of forecasting. Section 4 presents the Lee–Carter method as a state-space model under a Bayesian framework and gives details on implementation and prediction. It also presents multiple imputation as a technique for handling missing data. In Section 5, extensions to the state-space model are presented. Section 6 applies the original Lee–Carter methodology and the Bayesian model to mortality data from U.S. males. Forecasts and prediction intervals from both methods are presented. Model diagnostics are presented and an extension to the Bayesian model is derived. Finally, Section 7 provides concluding remarks.

2. HISTORICAL MORTALITY DATA

Mortality rates for developed countries declined dramatically during the past century. Improvements in standards of living, sanitary conditions, hygiene, and medicine led to rapidly decreasing infant mortality rates in the early part of the century and consequently to first increasing and then decreasing mortality in older age groups in the latter part of the 20th century. Figure 1 shows log-mortality rates for the U.S. obtained from the Human Mortality Database (2002). By mortality rates, we mean the ratio of deaths to mid-year population size for a given interval of age and time (central mortality rates). Figure 1 shows the constant decline of mortality rates over the last four decades for all-cause mortality in the U.S. for males and females. This is the general pattern for developed countries as illustrated by the two graphs of England and Spain male mortality shown in the bottom panels of Figure 1. This figure also illustrates the 'arm-shaped' profile of log-mortality across age groups.

In the youngest ages, high mortality rates are attributable to perinatal conditions. Mortality decreases during childhood and then increases again to the 'accident hump' (traffic accidents and teenage violence) at the ages of 15–20. The rates then steadily increase across older age groups. Over the past five decades, mortality rates have declined for all age groups but at different rates. Age groups 0–10 have declined at the fastest rates, while the oldest ages have declined at a much slower pace. This holds true for both males and females in developed regions (Murray and Lopez, 1996).

Overall, females experience lower mortality rates than males. The decade of 1965–1975 marks the period with the widest gap between males and females at the oldest ages. However, in the past decade,
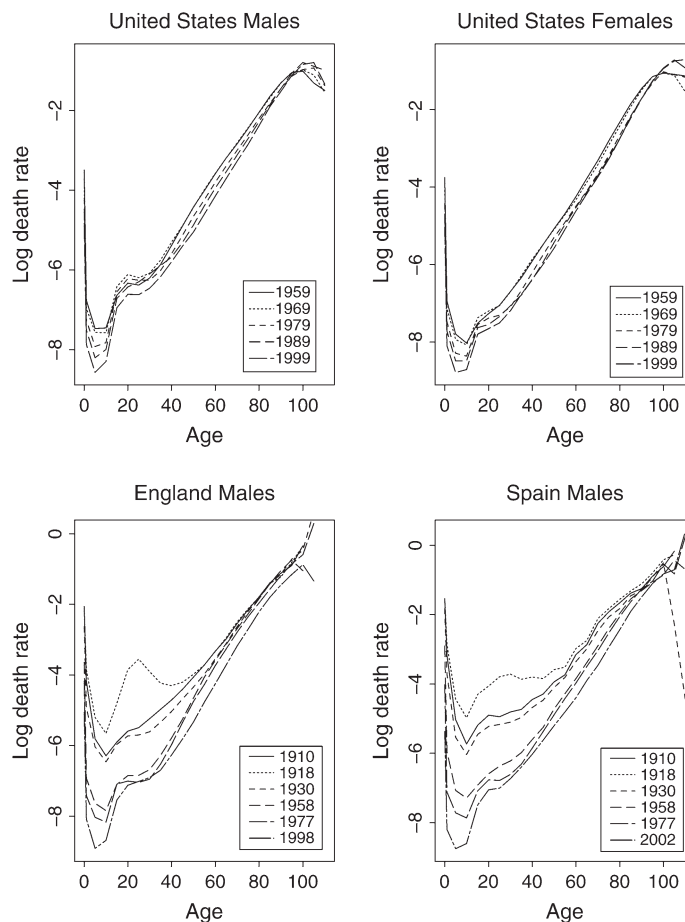
Fig. 1. Log-mortality rates for three countries. Notice the different rates of decline for different age groups over time.

the gap between males and females has steadily decreased. This is the general pattern for the developed countries, and in the past two decades, developing countries have exhibited the same pattern (see Murray and Lopez, 1996).

The calculation of mortality statistics in developing countries is plagued by missing or unreliable data from past years. Yet, mortality forecasts for these countries are critical since they are used to secure funds from international health-program organizations. Analyses based on data sets with missing data could lead to inaccurate estimation of mortality rates and in turn funds for health services. Thus, it is important that the forecasting method implement a systematic approach to the missing-data issue.

3. Lee–Carter Method

3.1 *Model*

Lee and Carter (1992) introduced a method for forecasting the age-specific log-mortality rates for the entire U.S. population. Let y_{it} be the logarithm of the central mortality rate for age group i, where $i = 1, \ldots, p$ and time $t = 1, \ldots, n$, and let $\mathbf{y}_t \equiv (y_{1t}, \ldots, y_{pt})'$. Lee and Carter's method consists of fitting

the following model to the log-mortality rates y_t:

$$y_t = \alpha + \beta \kappa_t + \varepsilon_t. \tag{3.1}$$

Here, α and β are vectors of unknown parameters, ε_t is a vector of error terms, and κ_t is an unobserved time series process. The parameters α, β, and κ are to be estimated. The vector α measures the general pattern of mortality for each age group, and β measures the relative rate of change of mortality at each age. The vector k_t is an index of the level of mortality. To ensure identifiability, Lee and Carter impose the following constraints:

$$\sum_i \beta_i = 1, \quad \sum_t \kappa_t = 0. \tag{3.2}$$

3.2 Estimation of parameters

Lee and Carter propose the use of singular value decomposition (SVD) to estimate the parameters in the model. Let Y be a $p \times n$ matrix of the log-mortality rates. The vector α is estimated by the average log-mortality over time for each age group. SVD is then applied to the mean corrected log-mortality rates $Y - \hat{\alpha}$:

$$Y - \hat{\alpha} = UDV',$$

where D is a diagonal matrix containing singular values and both U and V are orthogonal matrices. β is set equal to the first column of U, and the κ_t values are set equal to the product of the first column of V and the leading singular value d_1 along with the normalizations given in (3.2). This estimation amounts to using the first principal component of the log-mortality matrix, as first presented by Bozik and Bell (1987) and Bell and Monsell (1991).

3.3 Modeling and forecasting κ_t

After κ_t is estimated by SVD, one can make various adjustments to the estimated vector to produce a closer correspondence between estimated and observed death rates (see Lee, 2000, for further details). An appropriate univariate autoregressive integrated moving average (ARIMA) time series model forecasts the κ_t series. Lee and Carter proposed a random walk model with drift for κ_t, and most applications utilize the same approach. This model can be expressed as

$$\kappa_t = \kappa_{t-1} + \theta + \omega_t, \tag{3.3}$$

where θ is the drift parameter which models a linear trend and ω_t is an error term. The mean forecast value of κ_t is calculated as

$$\hat{\kappa}_{n+l} = \hat{\kappa}_n + l\hat{\theta},$$

for the lth step-ahead forecast. Finally, to obtain mean forecasts of the log-mortality rates, the mean forecasted values of κ_t are implemented along with the estimated values of α and β. The mean forecast of the log-mortality rate for year $n + l$ is

$$\hat{y}_{n+l} = \hat{\alpha} + \hat{\beta}\hat{\kappa}_{n+l}.$$

3.4 Prediction intervals

To calculate prediction intervals, Lee and Carter's method first expresses the squared standard error of the forecasted mortality index κ_{n+l} for the lth year-ahead forecast from model (3.3) as

$$\hat{\sigma}^2_{\kappa_{n+l}} = l^2 \hat{\sigma}^2_\theta + l \hat{\sigma}^2_\omega,$$

where $\hat{\sigma}_\theta^2$ is the squared standard error of θ and $\hat{\sigma}_\omega^2$ is the estimate of the variance of the error term ω. The Lee–Carter 95% prediction intervals for the log-mortality rates are computed as

$$\hat{y}_{n+l} \pm 1.96\hat{\boldsymbol{\beta}}\hat{\sigma}_{\kappa_{n+l}}.$$

As pointed out by Lee and Carter (1992, p. 669), this prediction interval ignores errors in estimating the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and the variance of the error term $\boldsymbol{\varepsilon}$. It only accounts for the error of forecasting κ_t. Their assumption is that the forecast error dominates all other errors. However, they do point out that for short-term forecasts their prediction intervals underestimate the 'true' forecast error.

The Lee–Carter method has several advantages. First, it is a parsimonious model which accounts for a large proportion of the variance in the log-mortality rates. Second, the Lee–Carter method uses time series methods to generate stochastic forecasts, and it provides probabilistic prediction intervals. The main criticism of the method is the assumption of a time-invariant age component, i.e. no age–time interaction. From historical data, it appears that this assumption is invalid. For example, the rate of decline in U.S. mortality during the 1990s is different than the one exhibited in the 1970s. Furthermore, infant mortality experienced the largest decrease in the early part of the century, whereas older age groups exhibited the largest decline in the latter part of the 20th century. Data from other developed countries also show different patterns for different age groups (Booth *et al.*, 2002). These violations raise important questions about the applicability of the Lee–Carter method for forecasting mortality in these countries.

4. BAYESIAN MODEL

Lee and Carter's method for forecasting mortality rates consists of (1) first fitting model (3.1) and (2) then forecasting the series κ_t using an appropriate time series model. As first presented in Pedroza (2002, p. 48), the Lee–Carter method can be reformulated as a state-space model:

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}\kappa_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{\text{iid}}{\sim} \text{N}_p(\boldsymbol{0}, \sigma_\varepsilon^2 I), \tag{4.1}$$

$$\kappa_t = \kappa_{t-1} + \theta + \omega_t, \quad \omega_t \overset{\text{iid}}{\sim} \text{N}(0, \sigma_\omega^2), \tag{4.2}$$

where $\boldsymbol{\varepsilon}_t$ and ω_t are assumed to be independent, and a random walk with drift model is assumed for the state vector. The multivariate normal model for the log-mortality rates provides a joint distribution for the p age groups at any given point in time. It assumes that the observations are independent across time, that is, the $\boldsymbol{y}_t$ are independent identically distributed (iid) with common variance σ_ε^2. In order to compare results from this model to those obtained from Lee and Carter's original model, the same constraints that Lee and Carter invoke, i.e. $\sum_i \beta_i = 1$ and $\sum_t \kappa_t = 0$, are used. Alternatively, any proper prior distribution on these parameters will lead to a valid posterior distribution. State-space models are extensively used in time series analysis (see, for example, Harvey, 1990; West and Harrison, 1997). Expressing the Lee–Carter method as a state-space model formalizes it as a statistical model. It also permits simultaneous estimation of all parameters, providing a systematic error assessment.

The Kalman filter can be used to estimate and forecast state-space models (Mehra, 1979; Harvey, 1990; Durbin and Koopman, 2001). Forecasting and handling of missing data in state-space models is straightforward with the Kalman filter. However, this algorithm does not incorporate uncertainty from unknown parameters and missing observations into either the forecasts or forecasting error. The forecast approach taken here follows the Bayesian forecasting methods of West and Harrison (1997) and makes use of simulation methods. The Kalman filter is incorporated into the fitting algorithm. Markov chain Monte Carlo (MCMC) methods are used to draw samples from the joint posterior distribution of the parameters and to form the posterior predictive distribution of the log-mortality rates. In particular, the

Gibbs sampler is used to fit the model and to forecast the rates. One advantage of the Bayesian approach is that uncertainty of the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and θ is incorporated in the estimation and forecasting of $\boldsymbol{\kappa}$.

Under a Bayesian paradigm, an investigator is able to formally incorporate prior information or knowledge about the problem at hand. This prior information is then combined with the observed data to form the posterior distribution of the parameters on which inference is based. Here, noninformative priors are used which lead to results comparable to those derived from Lee–Carter's original method. Standard flat prior distributions are assumed for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and θ along with noninformative priors for the variance parameters σ_ε^2 and σ_ω^2: $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \theta) \propto 1$, $p(\sigma_\varepsilon^2) \propto 1/\sigma_\varepsilon^2$, $p(\sigma_\omega^2) \propto 1/\sigma_\omega$. The initial prior distribution for the starting point κ_0 is specified as $N(a_0, Q_0)$, where a_0 and Q_0 are assumed to be known.

4.1 *Implementation*

To implement the Bayesian model, the joint posterior distribution of all unknown parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\kappa}$, θ, σ_ε^2, and σ_ω^2 needs to be obtained. However, obtaining posterior inferences analytically from such a high-dimensional distribution is complicated. Instead, MCMC methods are employed to obtain the posterior distribution of the parameters. More specifically, we use a Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) to draw samples from the joint posterior distribution. This algorithm consists of iteratively sampling from the conditional distribution of each of the parameters given assigned values to all the other parameters and the data. For the state-space model, the Gibbs sampler consists of two main steps: (1) drawing parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, θ, σ_ε^2, σ_ω^2 from their respective conditional distributions given fixed values for all other parameters and observed data and (2) simulating the state vector $\boldsymbol{\kappa}$.

One way to sample from the state vector is by using the single-state Gibbs sampler which consists of sampling from $p(\kappa_t | y, \boldsymbol{\kappa}^t, \phi)$, where $\boldsymbol{\kappa}^t$ is $\boldsymbol{\kappa}$ excluding κ_t and ϕ indexes all other parameters. However, this algorithm can be extremely inefficient. Frühwirth-Schnatter (1994) and Carter and Kohn (1994) independently developed methods for simulating the state vector based on the identity:

$$p(\kappa_1, \ldots, \kappa_T | y) = p(\kappa_T | y) p(\kappa_{T-1} | y, \kappa_T) \cdots p(\kappa_1 | y, \kappa_2, \ldots, \kappa_T). \tag{4.3}$$

De Jong and Shephard (1995) developed a similar method which samples the disturbance vectors instead of the state vectors. This method could also be used for the Lee–Carter model. Here we follow the approach of West and Harrison (1997) and use the identity given in (4.3) to sample the vector $\boldsymbol{\kappa}$.

Next we give the steps of the Gibbs sampler. Let Y^n be the observed data up to time n.

1. Run the Kalman filter with updating equations

$$\boldsymbol{\nu}_t = y_t - \boldsymbol{\alpha} - \boldsymbol{\beta} a_t, \quad Q_t = \boldsymbol{\beta} R_t \boldsymbol{\beta}' + \sigma_\varepsilon^2 I, \quad K_t = R_t \boldsymbol{\beta}' Q_t^{-1},$$
$$a_{t+1} = a_t + \theta + K_t \boldsymbol{\nu}_t, \quad R_{t+1} = R_t (1 - K_t \boldsymbol{\beta}) + \sigma_\omega^2,$$

for $t = 1, \ldots, n$, and store quantities a_t, Q_t, and R_t. Next, sample the state vector as follows (dependance on all other parameters is implicit):

   (a) Sample κ_n from $(\kappa_n | Y^n) \sim N(a_n, Q_n)$, then
   (b) for each $t = n - 1, n - 2, \ldots, 1, 0$, sample κ_t from $p(\kappa_t | \kappa_{t+1}, Y^n)$.

The conditional distribution for item (b) is

$$(\kappa_t | \kappa_{t+1}, Y^n) \sim N(h_t, H_t),$$

where

$$h_t = a_t + B_t(\kappa_{t+1} - a_{t+1}), \quad H_t = Q_t - B_t R_{t+1} B_t', \quad \text{and} \quad B_t = Q_t R_{t+1}^{-1}.$$

2. Draw σ_ε^2 from

$$\sigma_\varepsilon^2 | Y^n, \boldsymbol{\kappa}, \boldsymbol{\alpha}, \boldsymbol{\beta} \sim \text{Inv-Gamma}\left(\frac{pn}{2}, \frac{\sum_i \sum_t (y_{it} - \alpha_i - \beta_i \kappa_t)^2}{2}\right).$$

3. Draw $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by performing separate regressions for each age group of y_i on $\boldsymbol{\kappa}$. Letting $X = (\mathbf{1}, \boldsymbol{\kappa})$, where $\mathbf{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{\kappa}$ is the vector of κ_t values, the conditional distribution of α_i and β_i for age group i is

$$(\alpha_i, \beta_i) | Y^n, \boldsymbol{\kappa}, \sigma_\varepsilon^2 \sim \text{N}((X'X)^{-1}X'y, \sigma_\varepsilon^2 (X'X)^{-1}).$$

4. Draw θ from

$$\theta | \boldsymbol{\kappa}, \kappa_0, \sigma_\omega^2 \sim \text{N}\left(\frac{\kappa_n - \kappa_0}{n}, \frac{\sigma_\omega^2}{n}\right).$$

5. Draw σ_ω^2 from

$$\sigma_\omega^2 | \boldsymbol{\kappa}, \kappa_0, \theta \sim \text{Inv-Gamma}\left(\frac{n-1}{2}, \frac{\sum_{t=1}^n (\kappa_t - \kappa_{t-1} - \theta)^2}{2}\right).$$

Repeat steps 1 through 5 until convergence.

4.2   Prediction

In general, the posterior predictive distribution $(y_{n+1}|Y^n)$ for future observations can be expressed as

$$p(\boldsymbol{y}_{n+1}|Y^n) = \int p(\boldsymbol{y}_{n+1}|\phi, Y^n) p(\phi|Y^n) \mathrm{d}\phi = \int p(\boldsymbol{y}_{n+1}|\phi) p(\phi|Y^n) \mathrm{d}\phi,$$

where ϕ indicates all parameters in the model. The last equation holds because of the assumption that $\boldsymbol{y}_{n+1}$ and Y^n are conditionally independent given the parameters ϕ. The posterior predictive distribution can be obtained analytically or through simulation methods. Here, it is obtained by sampling iteratively from $p(\phi|Y^n)$ and $p(\boldsymbol{y}_{n+1}|\phi)$ to form the distribution.

Given the draws from the posterior distribution obtained from the Gibbs sampler, it is straightforward to forecast the log-mortality rates, $\boldsymbol{y}_t$. First, samples from the predictive distribution of κ_t are obtained from (4.2) and then samples from the predictive distribution of $\boldsymbol{y}_t$ are generated from (4.1).

These steps can be incorporated into the Gibbs sampler once convergence is attained as follows. Suppose we have $l = 1, \ldots, L$ draws of all the parameters from the Gibbs sampler and let $\varphi^{(l)}$ indicate the lth draw from the Gibbs sampler of the parameter φ. To forecast, we perform the following two steps:

1. Draw $\kappa_{n+1}^{(l)}$ from

$$\kappa_{n+1}^{(l)} \sim \text{N}\left(\kappa_n^{(l)} + \theta^{(l)}, \sigma_\omega^{2(l)}\right).$$

2. Draw $\boldsymbol{y}_{n+1}^{(l)}$ from

$$\boldsymbol{y}_{n+1}^{(l)} \sim \text{N}\left(\boldsymbol{\alpha}^{(l)} + \boldsymbol{\beta}^{(l)} \kappa_{n+1}^{(l)}, \sigma_\varepsilon^{2(l)} I\right).$$

Repeat steps 1 and 2 for $j = 1, \ldots, N$, where N is the number of step-ahead forecasts desired. The resulting samples of $\boldsymbol{y}$ form the predictive distribution of the mortality rates on which inferences are based.

4.3 *Missing data*

Mortality data can have missing observations across time and/or age groups. However, it is much more common to have incomplete time series data. Li *et al.* (2004) present a method for applying the original Lee–Carter model to forecast mortality for data sets with few observations over time and complete data across age groups. Their method accounts for extra variability arising from fewer time points when forecasting the $\boldsymbol{\kappa}$ parameter. However, their method does not account for the extra variability which will also be present in the $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters, and it is not meant to handle incomplete data across age groups.

To account for missing observations across time and/or age groups, we can use Rubin's (1987) method of multiple imputation. Briefly, multiple imputation replaces each missing value with m different values to create m complete data sets. Each complete data set is analyzed as if it was completely observed. The m results are then combined to produce a final inference. This final inference reflects the uncertainty due to the missing data. To impute the missing values, samples are drawn from the posterior predictive distribution of the missing values. Letting $\boldsymbol{y}_{\text{miss}}$ and $\boldsymbol{y}_{\text{obs}}$ be the vectors of missing and observed data, respectively, and assuming that the missing-data mechanism is ignorable (Little and Rubin, 1987), the posterior predictive distribution is

$$p(\boldsymbol{y}_{\text{miss}}|\boldsymbol{y}_{\text{obs}}) = \int p(\boldsymbol{y}_{\text{miss}}|\boldsymbol{y}_{\text{obs}}, \phi) p(\phi|\boldsymbol{y}_{\text{obs}}) \phi.$$

This is the conditional predictive distribution of $\boldsymbol{y}_{\text{miss}}$ given ϕ, averaged over the observed-data posterior of ϕ. Since the state-space model assumes iid observations over time and across age groups (the covariance matrix is assumed to be of the form $\sigma_{\varepsilon}^2 I$), the predictive distribution of the missing data given the parameters ϕ does not depend on the observed data, i.e. $p(\boldsymbol{y}_{\text{miss}}|\boldsymbol{y}_{\text{obs}}, \phi) = p(\boldsymbol{y}_{\text{miss}}|\phi)$. Thus, imputing the missing values is straightforward regardless of whether the missing data is across time, age groups, or both.

Here, we are mainly interested in computing estimates of the parameters in the model and in forming forecasts of the log-mortality rates. To estimate the parameters, we simulate random values of the parameters ϕ from the observed-data posterior distribution, $p(\phi|\boldsymbol{y}_{\text{obs}})$. This can be carried out through simulation techniques, in particular the Gibbs sampler. The Gibbs sampler in the presence of missing data consists of two main steps (Schafer, 1997): (1) imputation of the missing data given both the observed data and a sample of the parameters and (2) sampling of the parameters given the complete data. The first step imputes $\boldsymbol{y}_{\text{miss}}$ with a draw from $p(\boldsymbol{y}_{\text{miss}}|\boldsymbol{y}_{\text{obs}}, \boldsymbol{\kappa}, \phi)$ and the second steps draws the parameters from $p(\boldsymbol{\kappa}, \phi|\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{miss}})$. As mentioned above, the conditional distribution of $\boldsymbol{y}_{\text{miss}}$ given the parameters does not depend on $\boldsymbol{y}_{\text{obs}}$ regardless of whether the missing data is across time and/or age groups. Imputing the missing values consists of drawing from

$$p(\boldsymbol{y}_{\text{miss}}|\boldsymbol{\kappa}, \phi) \sim \text{N}(\boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{\kappa}, \sigma_{\varepsilon}^2 I).$$

Thus, in the presence of missing data, the Gibbs sampler has only one extra step. Handling missing data under the Bayesian paradigm can be incorporated seamlessly into the algorithm. The resulting forecasts of log-mortality rates account for the uncertainty due to missing data.

5. EXTENSIONS

As noted before, the Lee–Carter method assumes no age–time interaction. As a demonstration of the flexibility afforded by the Bayesian approach, we could let each age group follow its own random walk with different rates of decline. This level of flexibility is not required. Alternatively, the age groups can be clustered so that the youngest age groups follow one random walk, the middle age groups follow

another, and the oldest age groups follow a third different random walk. By adopting a Bayesian approach
to the Lee–Carter method, an array of changes and/or extensions to the original model can be easily
implemented. In this section, the notation of the model is generalized and some extensions are briefly
discussed.

The state-space model can be generalized as follows:

$$y_t = \alpha + Z\mu_t + \varepsilon_t, \quad \varepsilon_t \sim N(\mathbf{0}, \sigma_\varepsilon^2 I),$$

$$\mu_t = G\mu_{t-1} + \psi_t, \quad \psi_t \sim N(\mathbf{0}, \Psi). \tag{5.1}$$

Here, Z and G are the design and state matrices, respectively, which can include unknown parameters.
For the state-space model in Section 4, $Z = (\boldsymbol{\beta}, \mathbf{0})$, $\mu_t = (\kappa_t, \theta)'$, $\psi_t = (\omega_t, \eta_t)'$, $\Psi = \text{diag}(\sigma_\omega^2, 0)$, and

$$G = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}. \tag{5.2}$$

One way to generalize the Lee–Carter model is to change the form of the state equation. Thus, a simple
random walk with drift model can be replaced by different ARIMA models. For example, an autoregres-
sive process of order 1 [AR(1)] process could be used, placing more weight on the latest observations
when forming forecasts. The only thing that would differ from the original model would be the matrix G:

$$G = \begin{bmatrix} \phi & 1 \\ 0 & 1 \end{bmatrix}. \tag{5.3}$$

Another way to extend the Lee–Carter model is to allow the drift parameter to vary over time. From
empirical data, it appears that the rate of decrease of mortality has changed over the last century, sug-
gesting that a model with time-dependent drift would be a plausible model. In this case, $\mu_t = (\kappa_t, \theta_t)'$,
$\psi_t = (\omega_t, \eta_t)'$, G is as in (5.2), and $\Psi = \text{diag}(\sigma_\omega^2, \sigma_\eta^2)$. Another extension would be to have both an
AR(1) process for the κ_t and a time-dependent drift. The model would then be as above with the matrix
G equal to that of (5.3).

The correlation between age groups could also be modeled by changing the first-level covariance
matrix to a nondiagonal form. From historical data, older age groups exhibit more variability due to the
small size of that population. This could be readily incorporated into the model by letting different age
groups or clusters of the groups have different variances. Modeling of the variance–covariance matrix
would complicate the implementation of the model but could potentially allow for a more realistic model
of mortality rates. Barnard *et al.* (2000) provide a detail discussion on the modeling of covariance matrices.
Below, we give details of the Gibbs sampler for a model that assumes different variances for different age
groups but no correlation between the groups. Details of the fitting algorithm for the other extensions
discussed are given in the Appendix.

An extension to the Lee–Carter model which assumes different variance for the different age groups
can be expressed in state-space form as

$$y_t = \alpha + \beta\kappa_t + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(\mathbf{0}, \Sigma),$$

$$\kappa_t = \kappa_{t-1} + \theta + \omega_t, \quad \omega_t \overset{iid}{\sim} N(0, \sigma_\omega^2), \tag{5.4}$$

where Σ is assumed to be a diagonal matrix with up to p different σ_j^2 variance parameters along the
diagonal. Thus, if we grouped the age groups into three clusters of young, middle, and old, then we would
have three σ_j^2 parameters. Assuming that the variance parameters σ_j^2 are independent a priori, steps 1–3

of the Gibbs sampler change as follows:

1. Run the Kalman filter with updating equations

$$\boldsymbol{v}_t = \boldsymbol{y}_t - \boldsymbol{\alpha} - \boldsymbol{\beta} a_t, \quad Q_t = \boldsymbol{\beta} R_t \boldsymbol{\beta}' + \Sigma, \quad K_t = R_t \boldsymbol{\beta}' Q_t^{-1},$$

$$a_{t+1} = a_t + \theta + K_t \boldsymbol{v}_t, \quad R_{t+1} = R_t(1 - K_t \boldsymbol{\beta}) + \sigma_\omega^2,$$

   Note that the only change occurs in the equation for Q_t. Sampling of the state vector $\boldsymbol{\kappa}$ is carried out as described in Section 4.1.
2. For each cluster of age groups j, draw σ_j^2 from

$$\sigma_j^2 | Y^n, \boldsymbol{\kappa}, \boldsymbol{\alpha}, \boldsymbol{\beta} \sim \text{Inv-Gamma}\left(\frac{p_j n}{2}, \frac{\sum_{i \in j} \sum_t (y_{it} - \alpha_i - \beta_i \kappa_t)^2}{2} \right).$$

   where p_j is the number of age groups in cluster j.
3. Draw $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by performing separate regressions for each age group of $\boldsymbol{y}_i$ on $\boldsymbol{\kappa}$. Letting $X = (\mathbf{1}, \boldsymbol{\kappa})$, where $\mathbf{1}$ is an $n \times 1$ vector of ones and $\boldsymbol{\kappa}$ is the vector of κ_t values, the conditional distribution of α_i and β_i for age group i belonging to cluster j is

$$(\alpha_i, \beta_i) | Y^n, \boldsymbol{\kappa}, \sigma_j^2 \sim \text{N}((X'X)^{-1} X' \boldsymbol{y}, \sigma_j^2 (X'X)^{-1}).$$

Steps 4 and 5 of the Gibbs sampler are unchanged.

The prediction of future values of $\boldsymbol{y}_t$ is replaced by drawing from the following normal distribution

$$\boldsymbol{y}_{n+1} \sim \text{N}(\boldsymbol{\alpha} + \boldsymbol{\beta} \kappa_{n+1}, \Sigma).$$

If missing data are present across time, the only change in the imputation of missing observations is on the distribution used. Namely, the covariance matrix changes and the missing values are drawn from the multivariate normal distribution

$$p(\boldsymbol{y}_{\text{miss}} | \boldsymbol{\kappa}, \phi) \sim \text{N}(\boldsymbol{\alpha} + \boldsymbol{\beta} \boldsymbol{\kappa}, \Sigma).$$

If missing data are present across age groups, the imputations will consist of sampling from a conditional normal distribution. The methods for multivariate data presented in Schafer (1997) can be used to carry out the imputations.

6. APPLICATION

We analyze a data set for U.S. male all-cause mortality to demonstrate the Bayesian model for forecasting mortality. The data set was obtained from the Human Mortality Database (2002). This data set consists of annual age-specific death rates for years 1959–1999. The age groups are $0, 1$–$4, 5$–$9, \ldots, 105$–$109, 110+$. Figure 1, top left panel, shows a plot of the log-death rates for 5 years. To illustrate the methodology presented here and to assess the predictive quality of the models, we perform out-of-sample forecast for the last 10 years of the data. That is, we use the data from 1959 to 1989 to fit the model and then forecast for 10 years ahead. We then compare the forecasts to the observed values.

We fitted and forecasted log-mortality rates using both the original Lee–Carter method and the Bayesian model presented in Section 4. For the Bayesian model, we used normal distribution with a mean of 5 and variance of 10 as a prior distribution for the starting point of the state process, κ_0, based on results from previous studies. We consider a variance of 10 to be large enough to consider this a 'vague'

prior so as to lead to comparable results to those derived from Lee–Carter's original method. However, different choices of the prior distribution, in particular different values for the variance, lead to virtually identical results. Here, we summarize the results by presenting the out-of-sample forecasts and prediction intervals from both models.

For the Bayesian model, we ran three parallel chains of the Gibbs sampler with different overdispersed starting values for 2000 iterations. Convergence of the chains was checked using the Gelman–Rubin convergence diagnostic statistic (Gelman and Rubin, 1992) implemented in the CODA package for R (Plummer *et al.*, 2005). This statistic is computed for each scalar estimate of interest. A statistic less than 1.1 suggests convergence. The Gelman–Rubin statistic was computed for all parameters in the model, and all were less than 1.01, indicating convergence. Trace plots for a representative subset of the parameters are presented in Figure 2. These plots show that the three chains mix well together. A final sample for each parameter was obtained by selecting the last 1000 values of each of the three chains. Results presented here are based on the combined 3000 draws from the posterior distribution.

An appropriate concern would be computational cost. Each of the analyses (implementation of the model and prediction of the log-mortality rates) presented here took 5 min or less to run in R (R Development Core Team, 2005) on a personal computer with 3.2 GHz and 1 GB of memory. However, computational efficiency was not optimized.

Estimates of the parameters α and β obtained from the Lee–Carter method and the Bayesian model are shown in Table 1. These sets of estimates are almost identical for all age groups.

6.1 *Display of the forecasts*

For each of the $p = 24$ age groups, we present a graph of the observed log-mortality rates (solid line) for years 1990–1999 along with the point estimate of the forecast (dashed line) and 95% prediction intervals (dotted lines) (Figure 3). For the Bayesian model, the point estimate is the mean of the posterior predictive distribution. Figure 3 displays the results for a subset of five representative age groups from both the Lee–Carter method (left panel) and the Bayesian model (middle panel). The two features to focus on in these graphs are the accuracy of the point estimates and the coverage of the observed values by the prediction intervals. Comparing first the point forecasts, it appears that forecasts from the Bayesian and Lee–Carter method are very close. We calculated the mean squared error (MSE) and averaged it over the age groups:

$$\text{MSE} = \sum_{i} \sum_{t} (y_{it} - \hat{y}_{it})^2 / pn,$$

where $\hat{y}_{it}$ indicates either the Bayesian or Lee–Carter forecast for age group i and year t. The average MSE was 0.0098 for the Bayesian forecasts and 0.0104 for the Lee–Carter ones, indicating a very close agreement between the two sets of forecasts. There does not appear to be a clear pattern of over- or underestimation from these two methods.

Comparison of the prediction intervals yields very different results from the two methods. The Lee–Carter intervals for age groups 15–40 are quite narrow. The reason being that for these age groups the β parameters are very close to zero (Table 1), indicating that the relative rate of decline of mortality is close to zero. In other words, mortality for these age groups has seen a very small decline relative to other age groups. Given its role in the prediction interval, the magnitude of the β parameter makes the intervals on these age groups quite small. Overall, the prediction intervals obtained from the Bayesian model are wider than the Lee–Carter intervals. This is not a surprising result given the fact that this model incorporates all sources of variation in the model when creating forecasts and intervals. The majority of the Bayesian intervals cover the observed rates except for the oldest age groups which exhibit the most variability.

The 1-year-ahead forecast standard errors yield the starkest difference between the Bayesian and Lee–Carter model. For age group 25, the Lee–Carter error of 0.003 is 25 times smaller than the Bayesian
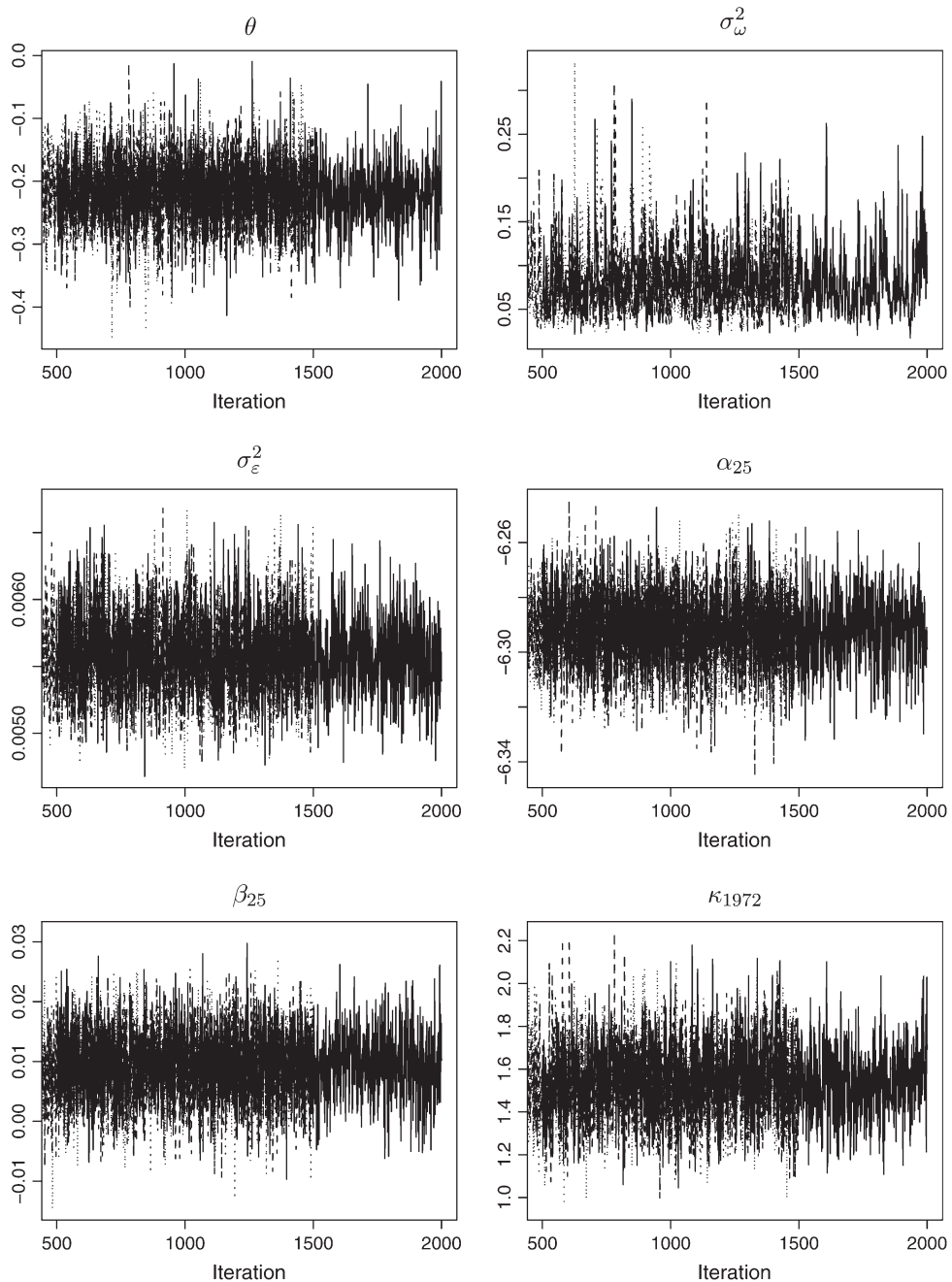
Fig. 2. Trace plots of three separate chains for a subset of parameters in the Bayesian model.

estimate of 0.078. The closest error estimates are for the youngest age group with the Lee–Carter estimate being approximately half the value of the Bayesian error of 0.098. Looking at the forecast standard errors for the 10-year-ahead forecasts, we see a closer agreement from the two methods. The biggest discrepancy is still in the age group 25, with a Lee–Carter error of 0.011 and a Bayesian error of 0.086.

Table 1. *Parameter estimates for the Lee–Carter (LC) and Bayesian models*

Age group	α			β		
	LC	Bayesian	Extended	LC	Bayesian	Extended
0	−3.9834	−3.9834	−3.9834	0.1543	0.1548	0.1520
1–4	−7.1189	−7.1190	−7.1190	0.1063	0.1063	0.1042
5–9	−7.7961	−7.7964	−7.7961	0.1079	0.1079	0.1065
10–14	−7.7148	−7.7151	−7.7148	0.0796	0.0794	0.0783
15–19	−6.5999	−6.5997	−6.6000	0.0296	0.0285	0.0295
20–24	−6.2755	−6.2756	−6.2761	0.0248	0.0238	0.0249
25–29	−6.2920	−6.2917	−6.2919	0.0101	0.0093	0.0100
30–34	−6.1792	−6.1793	−6.1791	0.0137	0.0131	0.0134
35–39	−5.8984	−5.8983	−5.8985	0.0412	0.0406	0.0408
40–44	−5.4982	−5.4983	−5.4981	0.0665	0.0661	0.0659
45–49	−5.0396	−5.0396	−5.0397	0.0729	0.0726	0.0724
50–54	−4.5676	−4.5678	−4.5677	0.0714	0.0715	0.0712
55–59	−4.1285	−4.1284	−4.1287	0.0641	0.0640	0.0638
60–64	−3.7038	−3.7037	−3.7037	0.0592	0.0592	0.0591
65–69	−3.3011	−3.3012	−3.3010	0.0577	0.0576	0.0575
70–74	−2.9102	−2.9102	−2.9103	0.0469	0.0468	0.0469
75–79	−2.5287	−2.5286	−2.5283	0.0426	0.0426	0.0426
80–84	−2.1288	−2.1288	−2.1282	0.0344	0.0343	0.0342
85–89	−1.7311	−1.7310	−1.7305	0.0326	0.0326	0.0323
90–94	−1.3700	−1.3703	−1.3702	0.0258	0.0258	0.0258
95–99	−1.0612	−1.0612	−1.0613	0.0117	0.0117	0.0115
100–104	−0.9447	−0.9446	−0.9447	−0.0238	−0.0243	−0.0242
105–109	−1.0464	−1.0462	−1.0464	−0.0618	−0.0616	−0.0597
110+	−1.2933	−1.2932	−1.2935	−0.0679	−0.0622	−0.0589

The Lee–Carter method gives an error of 0.175 for age group 0 compared to the Bayesian error of 0.181 for this group. Thus, for long-term forecasts, it appears that the Lee–Carter forecasting error starts to approach the Bayesian error, but for most age groups, it is still less than 75% of the error obtained from the Bayesian model.

The assumption made by Lee and Carter that the main source of error would come from the forecasting error of the mortality index κ_t yields prediction intervals which are too narrow for short-term forecasts, specially for young adults and older age groups. When we account for all sources of estimation error, we calculate prediction intervals which are wider and reflect the true prediction error in the forecasted mortality rates.

6.2 Model diagnostics

From the plots of the forecasted log-mortality rates, it appears that for some of the age groups (e.g. 50 and 90), the prediction intervals are too wide. One assumption of the fitted model is that all age groups have the same variability. However, plots of the observed data from 1959 to 1989 show that some age groups have more variability than others. In particular, the oldest age group appears to be more variable than all the other groups as would be expected. To test the assumption of homogeneity across age groups, we can look at posterior predictive diagnostics (Gelman *et al*., 1996). This approach consists of replicating data from the model and then comparing it to the observed data. Any systematic differences
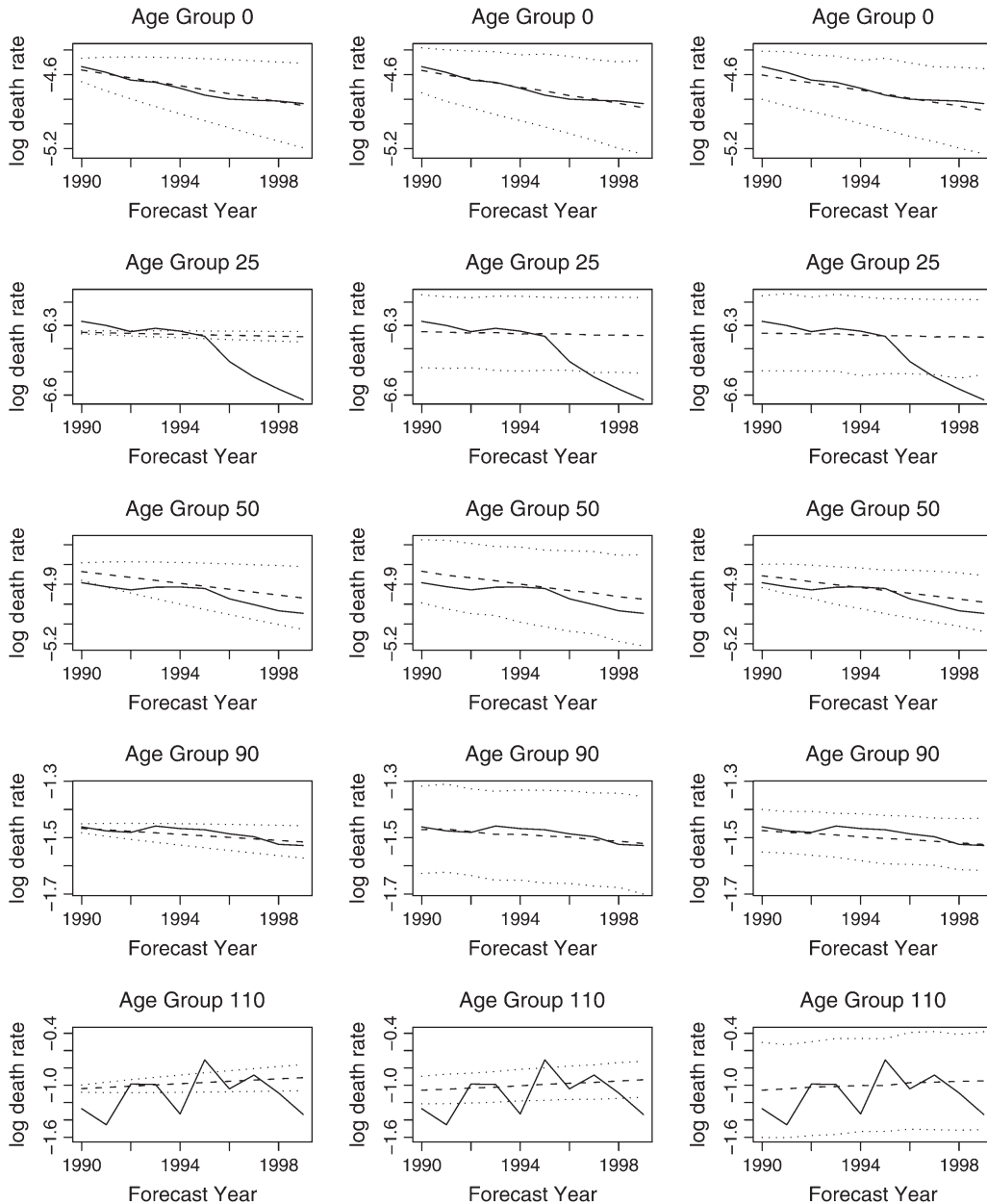
Fig. 3. Observed (solid line) and forecasted (dashed line) log-mortality rates with 95% prediction intervals for U.S. males for Lee–Carter's method (left panel), Bayesian model (middle), and extended Bayesian model (right).

between the replicated and observed data could indicate a misfit in the model. Differences or discrepancies between the model and data can be measured by a test quantity or discrepancy measure. This measure can depend on either just data (i.e. test statistic) or both data and parameters. The posterior predictive check is the comparison between realized values of the measure $D(y, \phi^*)$ obtained from the observed data y and simulated parameters ϕ and the predictive measures $D(y^*, \phi^*)$ obtained from simulated data and

parameters. If the actual data are plausible, then the posterior distribution of $D(y, \phi^*)$ will be similar to that of $D(y^*, \phi^*)$.

As a discrepancy measure here, we calculate the average squared residual for each age group as

$$e_i^2 = 1/n \sum_t (y_{it} - \alpha_i - \beta_i \kappa_t)^2.$$

Let $e_i^2(y^*; \phi^*)$ be the value of the discrepancy measure for the simulated data y^* and simulated parameters ϕ^* for age group i and $e_i^2(y; \phi^*)$ the value of the measure for the actual data and simulated parameters. This provides an estimate of the regression variance. The approximate posterior distributions of $(e_i^2(y; \phi^*), e_i^2(y^*; \phi^*))$ based on the same 500 draws of the parameters and simulated data are shown in Figure 4 (left panel) for a subset of the age groups.

If the posterior distributions of the discrepancy for the simulated data and the actual data are similar, then we should see large portions of the distribution of the discrepancies lie both above and below the diagonal line $(e_i^2(y; \phi^*) = e_i^2(y^*; \phi^*))$. For age groups 0 and 25, this seems to be the case, indicating that the observed variability is similar to the expected variability under the Bayesian model. However, for age groups 50 and 90, almost all the points lie above the diagonal, indicating that the variability under the model is much larger than the observed one for these two age groups. In other words, we overestimate the variability for these age groups. In contrast, the plot for the oldest age group 110 shows a clump of points all well below the diagonal, indicating that the observed variability is much larger than that expected under the model. These results indicate a lack of fit of the model to the observed data. From these graphs, we can conclude that there is evidence of heterogeneous variances for different age groups.

The posterior predictive diagnostics seem to indicate that a model with different variances for different age groups may be more appropriate. To keep the model as simple as possible, age groups may be grouped into clusters which exhibit similar variances. To help construct these clusters, we examined the approximate posterior distributions of the average squared residuals (i.e. approximate estimates of the regression variances) obtained from the actual data. The means of these distributions for the age group 24 are shown in Table 2. To try to capture the largest differences between the age groups, we decided on a model with 10 different variance parameters. We clustered the age groups as follows: 0, 1, 5–10,15–35, 40, 45–65, 70–80, 85–100, 105, 110. We fitted this model and again calculated the discrepancy measures for both actual and simulated data. This analysis took between 5 and 7 min to run in R. The resulting posterior distributions are shown in Figure 4 (right panel). Here we see that the values for the discrepancies are much more evenly distributed above and below the diagonal line, indicating a better fit to the observed data.

Plots of the forecasts from the extended model are shown in Figure 3 (right panel). Here we see the same pattern as for the simulated in-sample data. Prediction intervals for the younger age groups do not differ much from the ones produced by the original model. For the middle groups (i.e. 50–90) the intervals are narrower, and for the oldest group, the interval is much wider and covers the observed values. In general, the prediction intervals are much tighter for age groups which have historically shown smaller variability. The prediction intervals are improved, although the point estimates of the forecast do not change by much.

7. CONCLUSIONS

This paper has presented a Bayesian approach to the Lee–Carter method for forecasting mortality rates. The Bayesian formulation of the model incorporates all sources of variation in the model when forming forecasts. The resulting prediction intervals are much wider than those obtained by Lee–Carter, and they reflect more accurately the forecasting error associated with the model. Furthermore, the Bayesian model can easily handle missing data both in the time series and across age groups and incorporate the uncertainty
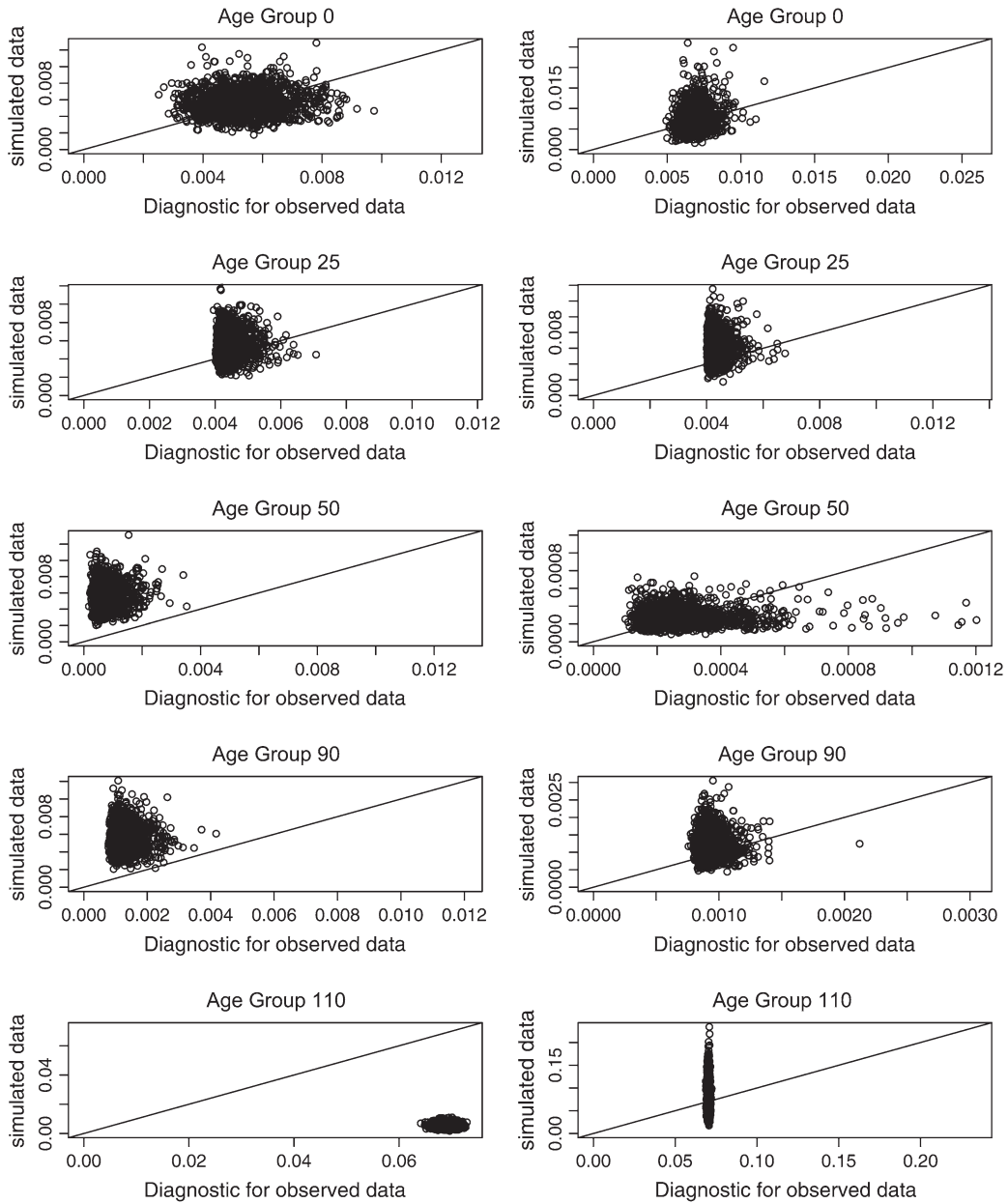
Fig. 4. Scatter plot of the joint posterior distribution of the averaged squared residual e_i^2 for age group i evaluated with observed data and simulated parameters vs. simulated data and parameters. Left panel corresponds to data simulated from the model assuming homogeneity of variance across age groups and the right corresponds to data simulated assuming heterogeneity.

associated with it. This aspect of the model is important when working with data from countries where vital records are incomplete or unreliable.

We compared results from the Bayesian model and the Lee–Carter model for U.S. male data and showed that confidence interval widths are different. For some age groups, the Lee–Carter intervals do

Table 2. *Approximate estimates of the regression variance for homogenous Bayesian model*

Age group	Estimated variance
0	0.0054
1–4	0.0030
5–9	0.0012
10–14	0.0014
15–19	0.0095
20–24	0.0080
25–29	0.0045
30–34	0.0043
35–39	0.0047
40–44	0.0018
45–49	0.0010
50–54	0.0008
55–59	0.0010
60–64	0.0008
65–69	0.0007
70–74	0.0008
75–79	0.0008
80–84	0.0007
85–89	0.0010
90–94	0.0013
95–99	0.0013
100–104	0.0021
105–109	0.0089
110+	0.0690

not cover the observed values. In contrast, the Bayesian intervals cover the observed values and reflect the true uncertainty of these forecasts. As an extension, we also fitted a model with heterogeneous variance for the age groups. This extension models the observed variability for different age groups and provides more realistic confidence intervals for the estimated log-mortality rates as well as prediction intervals for future observations. Another possible extension to the model would be to allow age groups to have different rates of decline. We could also let each age group have its own time-dependent drift. This model would allow each age group to have different rates of decline and also allow these rates to vary over time. Also note that different prior specifications for the model parameters can be used to incorporate prior information.

We realize that implementation of the Bayesian model is more complex. However, we believe that it is important for forecasts and their estimated errors to reflect the underlying uncertainty of the model and data. Furthermore, a Bayesian analysis allows the investigator to test the validity of the model through the use of posterior predictive checks as demonstrated in Section 6.2. We refer the reader to Gelman *et al*. (1995, 1996) for a more in-depth discussion of model checking in a Bayesian setting.

A second disadvantage of the Bayesian approach is in the specification of prior information. If true prior information is known, it should be incorporated into the Bayesian model. However, it is not always easy to quantify prior information in terms of a prior distribution. Careful consideration needs to be given to the prior specifications, and sensitivity analysis should be conducted to check how robust the forecasts are to the choice of the prior distribution.

One aspect of the Bayesian approach taken here is worthy of discussion. We implicitly used a squared-error loss when determining the Bayesian estimator for the forecasts. This loss function penalizes equally underestimation and overestimation. However, suppose we want forecasts for a developing country to

decide funding for a public health program. In this case, underestimation would be a far worse loss than overestimation. Here, a linear loss would be more appropriate. The Bayes estimate of the mortality forecasts would then be an appropriate percentile of the posterior predictive distribution. For example, if we deemed underestimation to be three times worse than overestimation, then we would use the 75th percentile as the estimate. For a different country, the estimate could possibly be another percentile, depending on the intended use of the forecasts. Tailoring a model to a specific data set is one of the most important advantages of a Bayesian analysis. A better fitting prediction model can lead to more realistic forecasts and prediction intervals of mortality rates.

APPENDIX

A.1   *The Gibbs sampler for extended models*

The general model presented in (5.1) can be fitted by the Gibbs sampler. Assuming noninformative priors for the parameters, the steps of the Gibbs sampler are as follows:

Gibbs sampler for general model.

1. Run the Kalman filter with the following updating equations

$$\boldsymbol{v}_t = \boldsymbol{y}_t - \boldsymbol{\alpha} - Z\boldsymbol{a}_t, \quad Q_t = ZR_tZ' + \sigma_\varepsilon^2 I,$$

$$K_t = GR_tZ'Q_t^{-1}, \quad L_t = G - K_tZ,$$

$$\boldsymbol{a}_{t+1} = G\boldsymbol{a}_t + K_t\boldsymbol{v}_t, \quad R_{t+1} = GR_tL_t' + \Psi.$$

Store quantities $\boldsymbol{a}_t$, Q_t, and R_t for $t = 1, \ldots, n$. Next, sample the state vector as follows:

(a) Sample $\boldsymbol{\mu}_n$ from $(\boldsymbol{\mu}_n|Y^n) \sim N(\boldsymbol{a}_n, Q_n)$, then
(b) for each $t = n - 1, n - 2, \ldots, 1, 0$, sample $\boldsymbol{\mu}_t$ from $p(\boldsymbol{\mu}_t|\boldsymbol{\mu}_{t+1}, Y^n)$.

The conditional distribution for item (b) is

$$(\boldsymbol{\mu}_t|\boldsymbol{\mu}_{t+1}, Y^n) \sim N(\boldsymbol{h}_t, H_t),$$

where

$$\boldsymbol{h}_t = \boldsymbol{a}_t + B_t(\boldsymbol{\mu}_{t+1} - \boldsymbol{a}_{t+1}), \quad H_t = Q_t - B_tR_{t+1}B_t', \quad \text{and} \quad B_t = Q_tG'R_{t+1}^{-1}.$$

2. Draw σ_ε^2 from Inv-Gamma$(pn/2, \sum_t(\boldsymbol{y}_t - \boldsymbol{\alpha} - Z\boldsymbol{\mu}_t)'(\boldsymbol{y}_t - \boldsymbol{\alpha} - Z\boldsymbol{\mu}_t)/2)$.
3. Draw $\boldsymbol{\alpha}$ and Z by performing separate regressions of $\boldsymbol{y}_i$ on $\boldsymbol{\mu}_t$ for each age group. Letting $X = (\boldsymbol{1}, \boldsymbol{\mu})$, the conditional distribution of α_i and Z_i for age group i is

$$(\alpha_i, Z_i)|Y^n, \boldsymbol{\mu}, \sigma_\varepsilon^2 \sim N((X'X)^{-1}X'\boldsymbol{y}, \sigma_\varepsilon^2(X'X)^{-1}).$$

4. Draw Ψ from its conditional distribution. Assuming the prior

$$p(\Psi) \propto |\Psi|^{-(d+1)/2},$$

where d is the dimension of Ψ, then the conditional distribution is an inverted-Wishart distribution:

$$p(\Psi|\boldsymbol{\mu}) \sim \mathrm{IW}_d(n, S^{-1}),$$

where $S = \sum_t (\boldsymbol{\mu}_t - G\boldsymbol{\mu}_{t-1})(\boldsymbol{\mu}_t - G\boldsymbol{\mu}_{t-1})'$.

Gibbs sampler for AR(1) model. In the case with an AR(1) process in the state equation, the model becomes

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}\kappa_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{\text{iid}}{\sim} \mathrm{N}(\boldsymbol{0}, \sigma_\varepsilon^2 I),$$

$$\kappa_t = \phi\kappa_{t-1} + \theta + \omega_t, \quad \omega_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_\omega^2).$$

The Gibbs sampler has an added step for drawing the parameter ϕ. The steps for this model are as follows:

1. Run the Kalman filter using the equations for the general model with the matrix G as in (5.2).
2. Draw σ_ε^2 from Inv-Gamma$(pn/2, \sum_i \sum_t (y_{it} - \alpha_i - \beta_i\kappa_t)^2/2)$.
3. Draw each α_i and β_i from $\mathrm{N}((X'X)^{-1}X'\boldsymbol{y}_i, \sigma_\varepsilon^2(X'X)^{-1})$, where $X = (\mathbf{1}, \boldsymbol{\kappa})$.
4. Draw σ_ω^2 from

$$\sigma_\omega^2|\boldsymbol{\kappa}, \kappa_0, \theta, \phi \sim \text{Inv-Gamma}\left(\frac{n-1}{2}, \frac{\sum_{t=1}^n (\kappa_t - \phi\kappa_{t-1} - \theta)^2}{2}\right).$$

5. Draw ϕ from

$$\phi|\boldsymbol{\kappa}, \theta, \sigma_\omega^2, Y^n \sim \mathrm{N}\left(\frac{\sum_{t=1}^n \kappa_{t-1}(\kappa_t - \theta)}{\sum_{t=1}^n \kappa_{t-1}^2}, \frac{\sigma_\omega^2}{\sum_{t=1}^n \kappa_{t-1}^2}\right).$$

References

Alho, J. M. and Spencer, B. D. (1990). Error models for official mortality forecasts. *Journal of the American Statistical Association* **85**, 609–616.

Barnard, J., McCulloch, R. and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10**, 1281–1311.

Bell, W. and Monsell, B. (1991). *Using principal components in time series modeling and forecasting of age-specific mortality rates*, ASA Proceedings of the Social Statistics Section. Alexandria, VA: American Statistical Association, pp. 154–159.

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

Bozik, J. E. and Bell, W. R. (1987). *Forecasting age specific fertility using principal components*, ASA Proceedings of the Social Statistics Section. Alexandria, VA: American Statistical Association, pp. 396–401.

Brouhns, N., Denuit, M. and Van Keilegom, I. (2005). Bootstrapping the Poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal* 212–224.

Brouhns, N., Denuit, M. and Vermunt, J. K. (2002). Measuring the longevity risk in mortality projections. *Bulletin of the Swiss Association of Actuaries*, 105–130.

CARTER, C. K. AND KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–553.

CARTER, L. (1995). Forecasting U.S. mortality: a comparison of Box-Jenkins ARIMA and structural time series models. *The Sociological Quarterly* **37**, 127–144.

CARTER, L. R. AND LEE, R. D. (1992). Modeling and forecasting US sex differentials in mortality. *International Journal of Forecasting* **8**, 393–411.

CZADO, C., DELWARDE, A. AND DENUIT, M. (2005). Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics & Economics* **36**, 260–284.

DE JONG, P. AND SHEPHARD, N. (1995). The simulation smoother for time series models. *Biometrika* **82**, 339–350.

DURBIN, J. AND KOOPMAN, S. J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.

FRÜHWIRTH-SCHNATTER, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis* **15**, 183–202.

GELFAND, A. E. AND SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

GELMAN, A., CARLIN, J. B., STERN, H. S. AND RUBIN, D. B. (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

GELMAN, A., MENG, X.-L. AND STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (discussion 760–807). *Statistica Sinica* **6**, 733–760.

GELMAN, A. AND RUBIN, D. B. (1992). Inference from iterative simulations using multiple sequences (with discussion). *Statistical Science* **7**, 457–472.

GEMAN, S. AND GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

GIROSI, F. AND KING, G. (2005). *Demographic Forecasting*. Monograph (unpublished).

HARVEY, A. C. (1990). *Forecasting, Structural Time Series Models, and the Kalman Filter*. Cambridge: Cambridge University Press.

HUMAN MORTALITY DATABASE. (2002). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at http://www.mortality.org or http://www.humanmortality.de (accessed on April 1, 2005).

LEE, R. (2000). The Lee-Carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal* **4**, 80–93.

LEE, R. D. AND CARTER, L. R. (1992). Modeling and forecasting U.S. mortality (discussion 671–675). *Journal of the American Statistical Association* **87**, 659–671.

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

LI, N., LEE, R. AND TULJAPURKAR, S. (2004). Using the Lee-Carter method to forecast mortality for populations with limited data. *International Statistical Review* **72**, 19–36.

LITTLE, R. J. A. AND RUBIN, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

MEHRA, R. K. (1979). Kalman filters and their applications to forecasting. In Makridakis, S. and Wheelwright, S. C. (eds), *Forecasting. Studies in the Management Sciences*, Volume 12. New York: Elsevier, pp. 75–94.

MURRAY, C. J. AND LOPEZ, A. D. (1996). *The Global Burden of Disease*. Cambridge, MA: Harvard University Press.

PEDROZA, C. (2002). Bayesian hierarchical time series modeling of mortality rates, *Ph.D. Thesis*, Harvard University, Cambridge, MA.

PLUMMER, M., BEST, N., COWLES, K. AND VINES, K. (2005). *Coda: Output Analysis and Diagnostics for MCMC*. R package version 0.9-2. Available at http://www-fis.iarc.fr/coda/.

R DEVELOPMENT CORE TEAM. (2005). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at http://www.R-project.org.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

WEST, M. AND HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer.

WILMOTH, J. R. (1995). Are mortality projections always more pessimistic when disaggregated by cause of death? *Mathematical Population Studies* **5**, 293–319.