ISSN 1440-771X

MONASH University

Department of Econometrics and Business Statistics

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

A comparison of ten principal component methods for forecasting mortality rates

Han Lin Shang, Rob J Hyndman and Heather Booth

April 2010

Working Paper 08/10

A comparison of ten principal component methods for forecasting mortality rates

Han Lin Shang
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: HanLin.Shang@buseco.monash.edu.au

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,
Australian National University, ACT 0200
Australia.
Email: Heather.Booth@anu.edu.au

9 April 2010

JEL classification: C14, C23

A comparison of ten principal component methods for forecasting mortality rates

Abstract

Using the age- and sex-specific data of 14 developed countries, we compare the short- to medium-term accuracy of ten principal component methods for forecasting mortality rates and life expectancy. These ten methods include the Lee-Carter method and many of its variants and extensions. For forecasting mortality rates, the weighted Hyndman-Ullah method provides the most accurate point forecasts, while the Lee-Miller method gives the best point forecast accuracy of life expectancy. Furthermore, the weighted Hyndman-Ullah method provides the most accurate interval forecasts of mortality rates, while the robust Hyndman-Ullah method provides the best interval forecast accuracy of life expectancy.

Keywords: mortality forecasting, life expectancy forecasting, principal component methods, Lee-Carter method, interval forecasts, forecasting time series

1 Introduction

In recent years, the rapid aging of the population has been a growing concern for governments and societies. In many developed countries, the concerns are concentrated on the sustainability of pensions and health and aged care systems, especially given increased longevity. This has resulted in a surge of interest among government policy makers and planners in accurately modeling and forecasting age-specific mortality rates. Any improvements in the forecast accuracy of mortality rates would be beneficial for policy decisions regarding the allocation of current and future resources. In particular, future mortality rates are of great interest to the insurance and pension industries.

Several authors have proposed new approaches for forecasting mortality rates and life expectancy using statistical modeling (see Booth 2006, Booth & Tickle 2008, for reviews). Of these, a significant milestone in demographic forecasting was the work of Lee & Carter (1992). They used a principal component method to extract a single time-varying index of the level of mortality rates, from which the forecasts are obtained using a random walk with drift. Since then, this method has been widely used for forecasting mortality rates in various countries, including Australia (Booth et al. 2002, De Jong & Tickle 2006), Austria (Carter & Prskawetz 2001), Belgium (Brouhns et al. 2002), Canada (Lee & Nault 1993), Chile (Lee & Rofman 1994), China (Lin 1995), Finland (Alho 1998), Japan (Wilmoth 1996), Norway (Keilman et al. 2002), Spain (Felipe et al. 2002, Debón et al. 2006), Sweden (Lundström & Qvist 2004, Tuljapurkar 2005), the U.K. (Renshaw & Haberman 2003*a*), the Nordic countries (Koissi et al. 2006), and the seven most economically developed nations (G-7) (Tuljapurkar et al. 2000). The strengths of the Lee-Carter (LC) method are its simplicity and robustness in situations where age-specific log mortality rates have linear trends (Booth et al. 2006). Although other methods have been developed (e.g., Renshaw & Haberman 2003*a*,*b*,*c*, Currie et al. 2004, Bongaarts 2005, Girosi & King 2008, Renshaw & Haberman 2006, Haberman & Renshaw 2008, Ediev 2008), the LC method is often considered as the benchmark method.

The fundamental principle of the LC method is the extrapolation of long-term trends in mortality rates. It is designed for long-term forecasting on the basis of a lengthy time series of historical data. However, structural changes in mortality patterns have occurred during the twentieth century, reducing the relevance of data from the distant past for the current forecasts (Booth et al. 2002, 2006). If a long fitting period cannot enhance the forecast accuracy, the heavy data demands of the LC method can be relaxed somewhat. The question of whether or not the length of the fitting period can dramatically affect the point forecast accuracy, had not been evaluated in great detail until the recent works of Carter & Prskawetz (2001), Booth et al. (2006), Wang (2007) and Wohlfart (2006).

Many people have proposed various modifications of the LC method. These include the LC method without adjustment, and the methods proposed by Tuljapurkar et al. (2000), Lee & Miller (2001) and Booth et al. (2002). The forecast accuracy of the LC method and its variants was first evaluated by Booth et al. (2005), and further studied by Booth et al. (2006). There have also been several extensions of the LC method. Of these, the extension proposed by Hyndman & Ullah (2007) has been receiving an increasing amount of attention in the fields of demography and statistics. Their method combined the ideas of nonparametric smoothing, functional principal component regression and functional data analysis, in order to forecast mortality and fertility rates. This method has been applied by Erbas et al. (2007) for forecasting breast cancer mortality rates in Australia. Furthermore, this method has been extended by Hyndman & Booth (2008) to improve the estimation of the variance and to include the forecasting of migration rates in Australia. Recently, Hyndman & Shang (2009) extended it to allow the weighting of more recent data.

In this article, we extend the results of Booth et al. (2006). First, we evaluate and compare the relative point forecast accuracy of the mortality rates and life expectancy from ten principal component methods, including methods proposed since the earlier comparison. Then, we evaluate the forecast uncertainty of mortality rates and life expectancy. This has been a recent focus, particularly for medium- to long-term forecasting (e.g., Alho 1997, Tayman et al. 1998, Lutz & Goldstein 2004, Alho & Spencer 2005, Brouhns et al. 2005, Koissi et al. 2006, Haberman & Renshaw 2008, Renshaw & Haberman 2008). To our knowledge, an empirical comparison of the forecast uncertainty estimates from these ten methods has never been undertaken.

This article is organized as follows. In Section 2, we briefly describe the ten mortality forecasting methods that are included in our comparisons, namely the LC method, the LC method without adjustment (LCnone), the Tuljapurkar-Li-Boe method (TLB), the Lee-Miller method (LM), the Booth-Maindonald-Smith method (BMS), the Hyndman-Ullah method (HU), the Hyndman-Ullah method using only data from 1950 onward (HU50), the robust Hyndman-Ullah method (HUrob), the robust Hyndman-Ullah method using only data from 1950 onward (HUrob50), and the weighted Hyndman-Ullah method (HUw). Section 3 presents the age- and sex-specific mortality rates of 14 developed countries. Using these 28 data sets, Section 4 compares the relative point forecast accuracy of the ten methods. Section 5 presents the methods for calculating interval forecasts, while Section 6 compares the relative interval forecast accuracy of the mortality rates and life expectancy from these ten methods. The discussion is presented in Section 7, along with some thoughts on how the comparisons might be further extended.

2 Review of mortality forecasting methods

In this section, we review the ten methods for forecasting mortality rates and life expectancy that are compared in the present study. We use the original notation of Lee & Carter (1992) and extend it as necessary for each method.

2.1 Lee-Carter (LC) method

In order to stabilize the high variance associated with high age-specific rates, it is necessary to transform the raw data by taking the natural logarithm. We denote by $m_{x,t}$ the observed mortality rate for age x in year t calculated as the number of deaths recorded among individuals aged x in calendar year t, divided by the corresponding exposure to risk.

The model structure proposed by Lee & Carter (1992) is given by

$$\ln(m_{x,t}) = a_x + b_x k_t + \varepsilon_{x,t}, \tag{1}$$

where a_x is the age pattern of the log mortality rates averaged across years; b_x is the first principal component reflecting relative change in the log mortality rate at each age; k_t is the first set of principal component scores by year t and measures the general level of the log mortality rates; and $\varepsilon_{x,t}$ is the residual at age x and year t.

The LC model in equation (1) is over-parametrized in the sense that the model structure is invariant under the following transformations:

$$\begin{aligned}
\{a_x, b_x, k_t\} &\mapsto \{a_x, b_x/c, ck_t\},\\
\{a_x, b_x, k_t\} &\mapsto \{a_x - cb_x, b_x, k_t + c\}.
\end{aligned}$$

In order to ensure the model's identifiability, Lee & Carter (1992) imposed two constraints, given as:

$$\sum_{t=1}^{n} k_t = 0, \qquad \sum_{x=x_1}^{x_p} b_x = 1,$$

where n is the number of years and p is the number of ages in the observed data set.

In addition, the LC method adjusts k_t by refitting the total number of deaths. This adjustment gives more weight to high rates, thus roughly counterbalancing the effect of using a log transformation of the mortality rates. In Sections 4 and 5, we also investigate the forecast accuracy of the LC model without adjustment, and the method proposed by Tuljapurkar et al. (2000). The method of Tuljapurkar et al. (2000) is a modification of the LC method without adjustment, and restricts the fitting period to 1950 onward only.

The adjusted k_t is then extrapolated using ARIMA models. Lee & Carter (1992) used a random walk with drift model, which can be expressed as:

$$k_t = k_{t-1} + d + e_t,$$

where d is known as the drift parameter and measures the average annual change in the series, and e_t is an uncorrelated error. It is notable that the random walk model with drift provides satisfactory results in many cases (Tuljapurkar et al. 2000, Lee & Miller 2001, Lazar & Denuit 2009). From this forecast of the principal component scores, the forecasted age-specific log mortality rates are obtained using the estimated age effects a_x and b_x in equation (1).

2.2 Lee-Miller (LM) method

The LM method is a variant of the LC method, and it differs from the LC method in three ways.

1. The fitting period begins in 1950.
2. The adjustment of k_t involves fitting to the life expectancy $e(0)$ in year t.
3. The jump-off rates are the actual rates in the jump-off year instead of the fitted rates.

In their evaluation of the LC method, Lee & Miller (2001) found a mismatch between fitted rates for the last year of the fitting period and actual rates in that year; this jump-off error amounted to 0.6 years in life expectancy for males and females combined (Lee & Miller 2001, p.539). This jump-off error was eliminated by using actual rates in the jump-off year.

In addition, the pattern of change in mortality rates was not constant over time, which is a strong assumption in the LC method. Consequently, the adjustment of historical principal component scores resulted in a large estimation error. To overcome this, Lee & Miller (2001) adopt 1950 as the commencing year of the fitting period due to different age patterns of change for 1900-1949 and 1950-1995. This solution to evolving age patterns of change has been adopted by Tuljapurkar et al. (2000).

In addition, the adjustment of k_t was done by fitting to observed life expectancy in year t, rather than by fitting to total deaths in year t. This has the advantage of eliminating the need for population data.

2.3 Booth-Maindonald-Smith (BMS) method

As a variant of the LC method, the BMS method differs from the LC method in three ways.

1. The fitting period is determined on the basis of statistical 'goodness of fit' criteria, under the assumption that the principal component scores $k_1, \dots, k_n$ are linear.
2. The adjustment of k_t involves fitting to the age distribution of deaths rather than the total number of deaths.
3. The jump-off rates are the fitted rates under this fitting regime.

A common feature of the LC method is the linearity of the best fitting time series model of the principal component scores, but Booth et al. (2002) found the linear time series to be compromised by structural change. By first assuming the linearity of principal component scores, the BMS method seeks to achieve the optimal 'goodness of fit' by selecting the optimal fitting period from all possible fitting periods. The optimal fitting period is determined based on the smallest ratio of the mean deviances of the fit of the underlying LC model to the overall linear fit. However, it can be quite computationally intensive, particularly for a long fitting period.

Instead of fitting to the total number of deaths, the BMS method fits to the age distribution of deaths using the Poisson distribution to model deaths, and using deviance statistics to measure the 'goodness of fit' (Booth et al. 2002). The jump-off rates are taken to be the fitted rates under this adjustment.

2.4 Hyndman-Ullah (HU) method

Using the functional data analysis technique of Ramsay & Silverman (2005), Hyndman & Ullah (2007) proposed a nonparametric method for modeling and forecasting log mortality rates. It extends the LC method in three ways.

1. The log mortality rates are first smoothed using penalized regression splines with a partial monotonic constraint. It is assumed that there is an underlying continuous and smooth function $f_t(x)$ that is observed with error at discrete ages. To emphasize that the age x is now considered as a continuous variable and incorporating the log transformation, we use $m_t(x)$ rather than $\ln m_{x,t}$ to represent the log mortality rates for age $x \in [x_1, x_p]$ in year t. Then, we can write

   $$m_t(x_i) = f_t(x_i) + \sigma_t(x_i)\varepsilon_{t,i}, \quad i = 1, \dots, p, \ t = 1, \dots, n \tag{2}$$

   where $m_t(x_i)$ denotes the log of the observed mortality rate for age x_i in year t; $\sigma_t(x_i)$ allows the amount of noise to vary with x_i in year t, thus rectifying the assumption of homoskedastic error in the LC model; and $\varepsilon_{t,i}$ is an independent and identically distributed standard normal random variable.

2. More than one principal component is used. Higher order terms of the principal component decomposition improve the LC model because these additional components capture non-random patterns, which are not explained by the first principal component (Booth et al. 2002, Renshaw & Haberman 2003*b*, Koissi et al. 2006). Using functional principal component analysis (FPCA), a set of curves is decomposed into orthogonal functional principal components and their uncorrelated principal component scores. That is,

   $$m_t(x) = a(x) + \sum_{j=1}^{J} b_j(x)k_{t,j} + e_t(x) + \sigma_t(x)\varepsilon_t, \tag{3}$$

   where the functional form of $m_t(x)$ can be reconstructed using a linear interpolation technique; $a(x)$ is the mean function, which can be estimated by $\hat{a}(x) = \frac{1}{n}\sum_{t=1}^{n} f_t(x)$; $\{b_1(x), \dots, b_J(x)\}$ is a set of the first J functional principal components; $\{k_{t,1}, \dots, k_{t,J}\}$ is a set of uncorrelated principal component scores satisfying $\sum_{j=1}^{J} k_{t,j}^2 < \infty$; $e_t(x)$ is the residual function with mean zero; and $J < n$ is the number of principal components used. Following Hyndman & Booth (2008), we chose $J = 6$, which should be larger than any of the components required. The conditions for the existence and uniqueness of $k_{t,j}$ are discussed by Cardot et al. (2003).

3. A wider range of univariate time series models may be used to forecast the principal component scores. By conditioning on the observed data $\mathcal{I} = \{m_t(x), t = 1, \ldots, n\}$ and the set of functional principal components $\boldsymbol{B} = \{b_1(x), \ldots, b_J(x)\}$, the h-step-ahead forecast of $m_{n+h}(x)$ can be obtained by:

$$\hat{m}_{n+h|n}(x) = \mathrm{E}[m_{n+h}(x)|\mathcal{I}, \boldsymbol{B}] = \hat{a}(x) + \sum_{j=1}^{J} b_j(x)\hat{k}_{n+h|n,j},$$

   where $\hat{k}_{n+h|n,j}$ denotes the h-step-ahead forecast of $k_{n+h,j}$ using a univariate time series model, such as an ARIMA model (Box et al. 2008), or an exponential smoothing state space model (Hyndman et al. 2008).

2.5 Weighted Hyndman-Ullah (HUw) method

The HUw method adopts the same smoothing technique as the HU method, but it uses geometrically decaying weights in the estimation of $a(x)$ and $b_j(x)$, thus allowing these quantities to be based more on recent data than on data from the distant past.

The HUw method differs from the HU method in three ways.

1. The weighted functional mean $a^*(x)$ is estimated using the weighted average

$$\hat{a}^*(x) = \sum_{t=1}^{n} w_t f_t(x), \tag{4}$$

   where $\{w_t = \lambda(1-\lambda)^{n-t}, t = 1, \ldots, n\}$ denotes a set of weights, and $0 < \lambda < 1$ denotes a geometrically decaying weight parameter. Hyndman & Shang (2009) describe how to estimate λ from data.

2. Using FPCA, a set of weighted curves $\{w_t[f_t(x) - \hat{a}^*(x)], t = 1, \ldots, n\}$ is decomposed into orthogonal weighted functional principal components and their uncorrelated principal component scores. That is,

$$m_t(x) = a^*(x) + \sum_{j=1}^{J} b_j^*(x)k_{t,j} + e_t(x) + \sigma_t(x)\varepsilon_t, \tag{5}$$

where $a^*(x)$ is the weighted functional mean, and $\{b_1^*(x),\dots,b_J^*(x)\}$ is a set of weighted functional principal components.

3. By conditioning on the observed data $\boldsymbol{\mathcal{I}} = \{m_t(x), t = 1,\dots,n\}$ and the set of weighted functional principal components $\boldsymbol{B}^* = \{b_1^*(x),\dots,b_J^*(x)\}$, the h-step-ahead forecast of $m_{n+h}(x)$ can be obtained by:

$$\hat{m}_{n+h|n}(x) = \mathrm{E}[m_{n+h}(x)|\boldsymbol{\mathcal{I}},\boldsymbol{B}^*] = \hat{a}^*(x) + \sum_{j=1}^{J} b_j^*(x)\hat{k}_{n+h|n,j}.$$

3 Data sets and forecast accuracy measure

The data sets used in this study were taken from the Human Mortality Database (2009). Fourteen developed countries were selected, and thus 28 age- and sex-specific populations were obtained for all analyses. The fourteen countries selected all have reliable data series commencing before 1950. Note that it was desirable to use only countries that have data prior to 1950, in order to maintain full and consistent comparisons of the ten methods. The selected countries are tabulated in Table 1, along with the commencing dates used to define the fitting periods for each method. For the BMS method, commencing dates may differ between the sexes as a result of independently and objectively selecting the optimal fitting period; in some cases, the difference is very large. Though originally proposed in the context of a relatively short fitting period, the BMS method results in a long fitting period for some populations. In this study, a common end year of 2004 was used. Age is in single years and we restrict the age range to 0–89, in order to avoid fluctuation in older ages.

We divide each data set into a fitting period and a forecasting period. The forecasting period is initially set to be the last 30 years (i.e., 1975 to 2004), while the rest of the data are included in the fitting period. Using the data in the fitting period, we compute 1- to 30-step-ahead point forecasts, and determine the forecast errors by comparing the forecasts with the actual out-of-sample data. Then, we increase the fitting period by one year, and compute 1- to 29-step-ahead forecasts, and calculate the forecast errors through the rolling origin approach. This process is repeated until the fitting period covers the entire length of the data set. By using the *demography* package for R (Hyndman et al. 2009), we can calculate the point and interval forecasts of each method, and evaluate and compare their forecast accuracy.

Country	LC	LCnone	TLB	LM	BMS[f]	BMS[m]
Australia	1921	1921	1950	1950	1982	1980
Canada	1921	1921	1950	1950	1981	1973
Denmark	1835	1835	1950	1950	1971	1987
England	1841	1841	1950	1950	1975	1985
Finland	1878	1878	1950	1950	1987	1985
France	1816	1816	1950	1950	1968	1979
Iceland	1838	1838	1950	1950	1838	1838
Italy	1872	1872	1950	1950	1984	1982
Netherlands	1850	1850	1950	1950	1983	1977
Norway	1846	1846	1950	1950	1986	1986
Scotland	1855	1855	1950	1950	1855	1985
Spain	1908	1908	1950	1950	1936	1936
Sweden	1751	1751	1950	1950	1927	1987
Switzerland	1876	1876	1950	1950	1984	1987
Country	HU	HU50	HUrobust	HUrobust50	HUw	
Australia	1921	1950	1921	1950	1921	
Canada	1921	1950	1921	1950	1921	
Denmark	1835	1950	1835	1950	1835	
England	1841	1950	1841	1950	1841	
Finland	1878	1950	1878	1950	1878	
France	1816	1950	1816	1950	1816	
Iceland	1838	1950	1838	1950	1838	
Italy	1872	1950	1872	1950	1872	
Netherlands	1850	1950	1850	1950	1850	
Norway	1846	1950	1846	1950	1846	
Scotland	1855	1950	1855	1950	1855	
Spain	1908	1950	1908	1950	1908	
Sweden	1751	1950	1751	1950	1751	
Switzerland	1876	1950	1876	1950	1876	

Table 1: *Commencing year of the fitting period for each country and each method*

The forecast errors of log mortality rates are summarized using the mean absolute forecast error (MAFE). The MAFE is the average of absolute errors, |actual − forecast|, across different age groups, forecast horizons and forecast periods. It is a measure of forecast precision, or how close the forecasts were to the out-of-sample values, regardless of whether they were high or low.

4 Comparisons of the point forecasts

For simplicity, we will refer to the LC method and its variants (all based on a single principal component) as the LC methods, and to the HU method and its variants (all using a smoothing technique and several principal components) as the HU methods.

4.1 Forecast log mortality rates

Tables 2 and 3 provide summaries of the point forecast accuracy in the female and male data based on the MAFE averaged over age, forecast horizon and forecast period. The LC method performs the least well, but its variants and extensions provide substantial improvements in point forecast accuracy. Judging by the minimum MAFE averaged over countries, the HUw approach performs the best in both the female and male data.

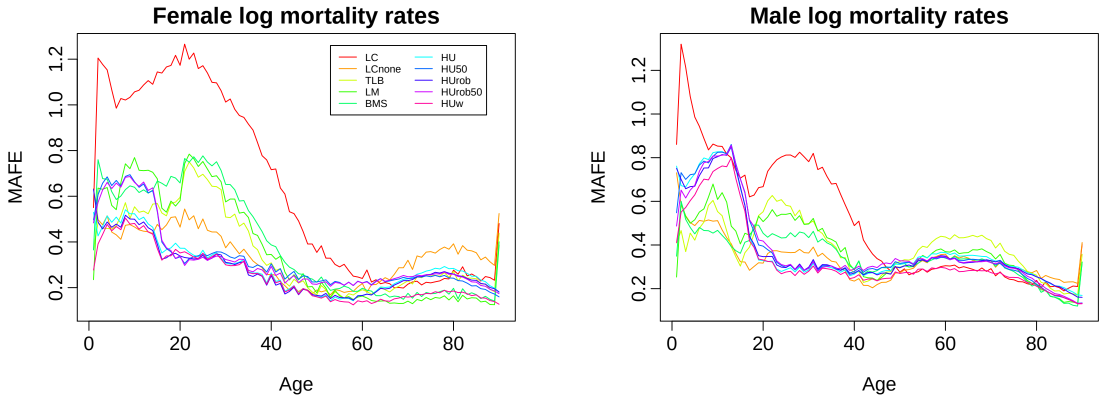
Country	LC	LCnone	TLB	LM	BMS	HU	HU50	HUrob	HUrob50	HUw
Australia	0.45	0.33	0.26	0.23	0.34	0.20	0.25	0.23	0.23	0.20
Canada	0.34	0.23	0.17	0.15	0.16	0.17	0.22	0.16	0.23	0.14
Denmark	0.58	0.26	0.27	0.29	0.29	0.23	0.28	0.25	0.29	0.20
England	0.33	0.19	0.27	0.26	0.21	0.17	0.26	0.20	0.27	0.17
Finland	0.82	0.40	0.52	0.50	0.48	0.33	0.42	0.28	0.39	0.29
France	0.44	0.31	0.21	0.20	0.22	0.33	0.27	0.30	0.28	0.23
Iceland	0.91	0.80	1.01	1.33	0.89	0.71	0.76	0.72	0.79	0.71
Italy	0.34	0.24	0.31	0.24	0.20	0.20	0.24	0.20	0.23	0.19
Netherlands	0.66	0.23	0.23	0.21	0.42	0.33	0.34	0.30	0.34	0.20
Norway	0.58	0.30	0.37	0.33	0.30	0.29	0.31	0.28	0.29	0.27
Scotland	1.30	0.29	0.40	0.40	0.63	0.33	0.45	0.34	0.49	0.30
Spain	0.63	0.31	0.41	0.39	0.59	0.25	0.37	0.22	0.34	0.23
Sweden	0.29	0.29	0.27	0.34	0.22	0.43	0.30	0.42	0.28	0.36
Switzerland	0.47	0.34	0.27	0.28	0.36	0.26	0.30	0.26	0.31	0.21
Mean	0.58	0.32	0.36	0.37	0.38	0.30	0.34	0.30	0.34	0.26

Table 2: *Accuracy of point forecasts of female log mortality rates, as measured by the MAFE*

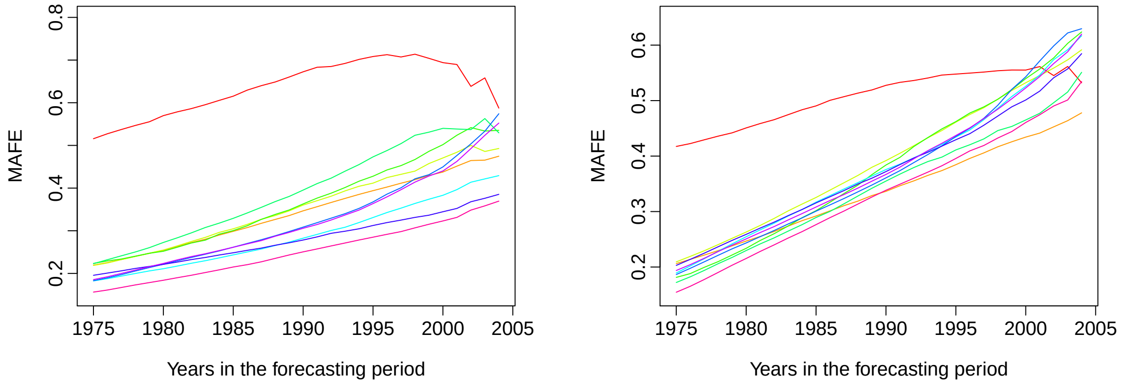
Country	LC	LCnone	TLB	LM	BMS	HU	HU50	HUrob	HUrob50	HUw
Australia	0.58	0.38	0.36	0.32	0.37	0.28	0.36	0.32	0.30	0.29
Canada	0.32	0.28	0.26	0.23	0.29	0.25	0.30	0.27	0.32	0.22
Denmark	0.25	0.23	0.33	0.32	0.32	0.23	0.35	0.25	0.37	0.21
England	0.46	0.33	0.30	0.26	0.28	0.35	0.30	0.35	0.31	0.32
Finland	0.49	0.39	0.56	0.52	0.34	0.47	0.41	0.46	0.42	0.40
France	0.40	0.37	0.24	0.22	0.28	0.42	0.33	0.43	0.30	0.38
Iceland	0.58	0.60	0.92	1.06	0.56	0.80	0.89	0.90	0.89	0.65
Italy	0.34	0.30	0.36	0.32	0.29	0.42	0.28	0.42	0.30	0.39
Netherlands	0.21	0.19	0.30	0.30	0.29	0.43	0.39	0.42	0.37	0.21
Norway	0.37	0.28	0.38	0.38	0.38	0.36	0.35	0.31	0.39	0.31
Scotland	1.65	0.39	0.47	0.42	0.38	0.46	0.50	0.46	0.51	0.42
Spain	0.44	0.28	0.37	0.36	0.41	0.39	0.40	0.24	0.38	0.24
Sweden	0.46	0.30	0.31	0.32	0.37	0.47	0.35	0.49	0.39	0.32
Switzerland	0.29	0.33	0.29	0.29	0.30	0.30	0.32	0.32	0.35	0.26
Mean	0.49	0.33	0.39	0.38	0.35	0.40	0.40	0.40	0.40	0.33

Table 3: *Accuracy of point forecasts of male log mortality rates, as measured by the MAFE*

Figures 1a and 1b show the MAFE for different methods averaged over countries and years in the forecasting period. The age patterns are similar in that larger errors occur at younger ages (less than 40) for all methods; errors of the LC method are particularly large at these ages. The HU methods tend to be more accurate than LC methods at ages 20–40, but at other ages there is



(a) *MAFE averaged over countries and years in the forecasting period* **(b)** *MAFE averaged over countries and years in the forecasting period*



(c) *MAFE averaged over countries and ages* **(d)** *MAFE averaged over countries and ages*

Figure 1: *MAFEs in log mortality rate forecasts by sex and method*

no consistency. Figures 1c and 1d show the MAFE for different methods averaged over countries and ages. As expected, all of the methods show that MAFE increases with the forecast horizon, but divergence between the MAFEs from different methods is occurring particularly for female mortality rates. The LC method starts from a high level in line with the large errors at younger ages, but finishes closer to other methods indicating better relative accuracy in the longer term.

From these results, we find that the effect of different fitting periods is a result of the different trends in the principal component scores. When forecasting log mortality rates, the MAFEs are consistently greater for methods with longer fitting periods. This can be observed from the significant improvement in point forecast accuracy of the LM method over the LC method and the HUw method over the HU method.

The effect of the adjustment in the LC method is seen to be relatively large. When fitted rates are used, any adjustment would worsen the forecasts of log mortality rates because the fit to the base model is no longer statistically optimal. The adjustment in the LC method ignores potentially large positive and negative errors by age as long as they cancel in total deaths; however, these errors do not cancel in the calculation of MAFE. The BMS adjustment produces smaller forecast errors because it takes the size of errors by age into account in fitting to the age distribution of deaths, while the LM method may achieve smaller forecast errors by virtue of the weighting implicit in the calculation of life expectancy, and by avoiding jump-off error.

Furthermore, the superior point forecast accuracy of the HUw method is attributable to its greater statistical sophistication. Compared with the HU50 and HUrob50 methods, the HUw method is a data-driven approach which discounts the effects of distant past data instead of eliminating them altogether by reducing the fitting period.

4.2 Forecast life expectancy

Tables 4 and 5 present the MAFEs for life expectancy forecasts for females and males. The LM method achieves the best point forecast accuracy among all of the methods. This is because the LM method adjusts the principal component score in a way that reproduces observed life expectancy. In addition, the LM method uses the actual jump-off rates rather than the fitted rates. When forecasting life expectancy, it is an advantage to use actual jump-off rates instead of fitted jump-off rates; the gain in forecast accuracy is substantial when the fitting period is long, or when an adjustment is used.

For life expectancy, the LC methods are often more accurate than the HU methods. This may be explained by a greater degree of cancelation of the generally larger errors at different ages in the mortality rates. Among the HU methods, the better performance of the HUw method can be explained by the weight given to the most recent data, which has a similar (but not as strong) effect as matching life expectancy in the jump-off year (as in the LM method).

It may be noted that the LC method performs much better for life expectancy than for log mortality rates. This is also a result of the cancelation of errors at different ages in the calculation of life expectancy.

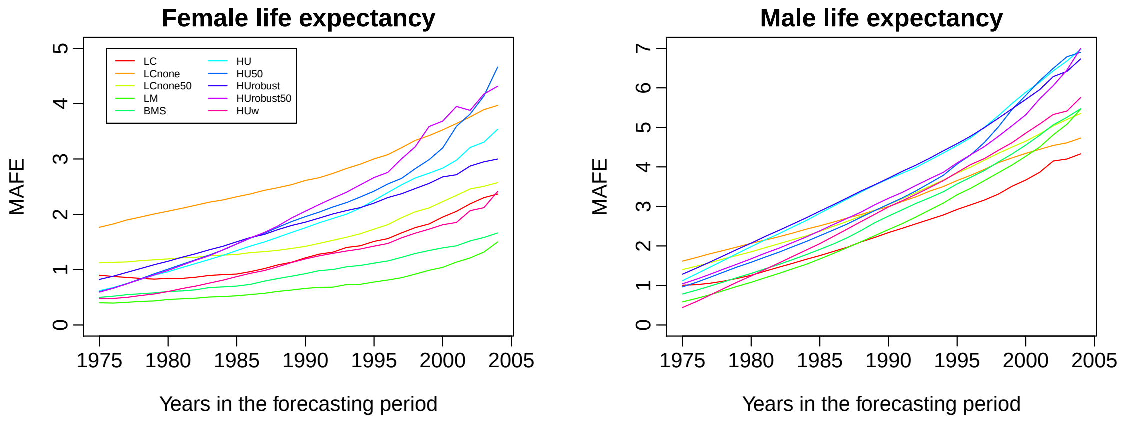
Figure 2 displays the MAFEs in life expectancy forecasts by forecasting period in the female and male data. For female life expectancy, there is considerable variation between methods, with the small MAFEs for LM and BMS methods increasing slowly over time, while others increase rapidly. For males, all methods increase rapidly.

Country	LC	LCnone	TLB	LM	BMS	HU	HU50	HUrob	HUrob50	HUw
Australia	1.41	3.03	1.95	1.28	1.86	1.42	1.70	1.24	1.60	1.19
Canada	0.39	1.54	0.55	0.56	0.43	1.76	1.35	1.55	1.88	1.54
Denmark	0.85	2.14	0.97	0.78	1.14	0.77	1.26	1.07	0.94	0.50
England	1.08	2.39	1.51	0.57	1.06	1.06	2.26	1.64	2.24	0.70
Finland	1.37	3.87	2.35	0.80	0.95	2.20	2.00	1.71	2.89	0.77
France	1.90	4.57	1.32	0.57	0.69	3.27	1.97	3.52	2.01	1.98
Iceland	1.65	1.50	3.69	0.89	1.49	2.32	3.32	1.88	3.23	2.12
Italy	1.94	3.16	1.51	0.92	1.18	2.12	2.19	2.82	1.96	1.61
Netherlands	1.28	1.77	0.82	0.95	1.62	1.94	2.63	1.53	3.47	0.79
Norway	0.86	2.36	1.64	0.54	0.67	1.00	1.89	1.29	1.70	0.54
Scotland	2.08	2.84	1.84	0.66	0.55	2.14	2.47	2.62	3.53	1.47
Spain	1.64	2.95	2.08	0.86	1.06	1.27	2.51	0.78	2.08	0.80
Sweden	1.23	2.70	1.51	0.36	0.24	2.99	1.67	2.94	1.44	2.07
Switzerland	0.80	2.98	0.66	0.26	0.44	1.38	1.48	1.33	1.23	0.55
Mean	1.32	2.70	1.60	0.71	0.96	1.83	2.05	1.85	2.16	1.19

Table 4: *Accuracy of point forecasts of female life expectancy, as measured by the MAFE*

Country	LC	LCnone	TLB	LM	BMS	HU	HU50	HUrob	HUrob50	HUw
Australia	3.00	4.84	4.66	3.66	4.26	2.95	3.71	3.43	2.44	3.29
Canada	2.53	3.65	3.05	2.40	3.02	2.53	2.48	2.86	2.60	2.06
Denmark	1.88	1.42	2.76	2.58	2.67	1.77	2.67	1.88	2.88	1.39
England	2.43	3.88	3.17	2.42	2.67	4.37	3.39	4.39	3.45	3.79
Finland	3.02	4.74	3.23	1.70	3.19	5.37	3.88	5.01	4.29	4.09
France	2.60	5.27	2.70	2.15	2.53	4.92	3.14	5.12	2.69	3.98
Iceland	1.78	1.86	5.39	4.56	1.92	4.67	5.69	6.06	5.61	3.47
Italy	3.01	3.30	3.18	2.47	2.84	5.00	2.76	5.24	3.28	4.25
Netherlands	1.36	1.05	3.07	0.30	0.29	0.43	0.39	0.42	0.37	3.00
Norway	2.93	1.96	3.43	3.48	3.76	3.15	3.03	2.57	3.69	2.51
Scotland	1.95	3.71	3.36	2.13	2.74	3.74	4.66	3.85	3.94	2.67
Spain	1.28	2.10	0.85	0.54	1.17	3.73	3.38	1.47	2.97	0.80
Sweden	3.62	2.16	2.80	2.75	3.26	4.28	2.96	4.65	3.46	3.48
Switzerland	2.01	3.15	2.20	2.08	2.15	2.59	2.33	2.75	2.81	1.78
Mean	2.39	3.08	3.13	2.37	2.60	3.54	3.18	3.55	3.18	2.90

Table 5: *Accuracy of point forecasts of male life expectancy, as measured by the MAFE*



(a) *MAFE averaged over countries* **(b)** *MAFE averaged over countries*

Figure 2: *MAFEs in life expectancy forecasts by sex and method averaged over countries*

5 Review of interval forecast methods

Prediction intervals are a valuable tool for assessing the probabilistic uncertainty associated with point forecasts. As was emphasized by Chatfield (1993, 2000), it is important to provide interval forecasts as well as point forecasts, so as to

1. assess future uncertainty levels;
2. enable different strategies to be planned for the range of possible outcomes indicated by the interval forecasts;
3. compare forecasts from different methods more thoroughly; and
4. explore different scenarios based on different assumptions.

In Section 5.1, we first briefly describe a parametric approach to the construction of prediction intervals for the LC methods. In Section 5.2, we review a parametric approach for the HU methods proposed by Hyndman & Ullah (2007). We then revisit a nonparametric bootstrap approach of Hyndman & Shang (2009) to the construction of prediction intervals for the HU methods in Section 5.3. Section 5.4 presents the comparisons of the interval forecast accuracy for both the log mortality rates and life expectancy.

5.1 LC methods

In Lee & Carter (1992), the LC method considers only the uncertainty in the innovations. Although Lee & Carter (1992, p.665) acknowledged that the inclusion of uncertainty in the drift would increase the standard error of their forecasts by 25% after 50 years (based on the period from 1900 to 1989), they did not include it. Booth et al. (2002) included the uncertainty in the drift, and compared the variances for the LC, LM and BMS methods with and without this additional uncertainty.

We consider two sources of uncertainty: errors in the parameter estimation of the LC models and forecast errors in the projected time series coefficients. Because of the orthogonality between the first principal component and the error term in equation (1), the overall forecast variance can be approximated by the sum of the two variances. Conditioning on the observed data $\mathcal{I}$ and the first principal component b_x, we obtained the overall forecast variance of $\ln m_{x,t}$,

$$\mathrm{Var}[\ln m_{x,t}] \approx b_x^2 u_{n+h|n} + v_x,$$

where b_x^2 is the variance of the first principal component; $u_{n+h|n} = \text{Var}(k_{n+h}|k_1,\dots,k_n)$ can be obtained from the time series model; and the model error variance v_x is estimated by averaging the residual squares $\{\varepsilon_{x,1}^2,\cdots,\varepsilon_{x,n}^2\}$ for each x.

5.2 HU methods — a parametric approach

The forecast variance follows from equation (3). Due to the orthogonality between the principal components and the error term, the overall forecast variance can be approximated by the sum of four variances. Conditioning on the observed data $\mathcal{I}$ and the set of fixed principal components $\boldsymbol{B} = \{b_1(x),\cdots,b_J(x)\}$, we obtained the overall forecast variance of $m_{n+h}(x)$,

$$\text{Var}[m_{n+h}(x)|\mathcal{I},\boldsymbol{B}] \approx \hat{\sigma}_a^2(x) + \sum_{j=1}^{J} b_j^2(x)u_{n+h|n,j} + v(x) + \sigma_{n+h}^2(x),$$

where $\hat{\sigma}_a^2(x)$ (the variance of the smooth estimate $\hat{a}(x)$) can be obtained from the smoothing method used in equation (2); $u_{n+h|n,j} = \text{Var}(k_{n+h,j}|k_{1,j},\dots,k_{n,j})$ can be obtained from the time series model; the model error variance $v(x)$ is estimated by averaging $\{e_1^2(x),\dots,e_n^2(x)\}$ for each x; and the observational error variance $\sigma_{n+h}^2(x)$ is computed from the data (Hyndman & Ullah 2007).

By assuming that each of the four sources of errors has a normal distribution and are uncorrelated, the $100(1-\alpha)\%$ prediction intervals of $m_{n+h}(x)$ are constructed as $\hat{m}_{n+h|n}(x) \pm z_\alpha\sqrt{\text{Var}[m_{n+h}(x)\,|\,\mathcal{I},\boldsymbol{B}]}$, where z_α is the $(1-\alpha/2)$ standard normal quantile. In practice, if normality is not justified, a nonparametric bootstrap procedure could be used instead.

5.3 HU and HUw methods — a nonparametric approach

Using univariate time series models, we can obtain one- or multi-step-ahead forecasts for the principal component scores $\{k_{1,j},\dots,k_{n,j}; j=1,\dots,J\}$, where J is the optimal number of principal components. Let the h-step-ahead forecast errors be given by $\hat{\xi}_{i,h,j} = k_{i,j} - \hat{k}_{i|i-h,j}$ for $i = J+h,\dots,n$. These can then be randomly sampled with replacement to give bootstrap samples of $k_{n+h,j}$, which can be expressed as

$$\hat{k}_{n+h|n,j}^{\ell} = \hat{k}_{n+h|n,j} + \hat{\xi}_{*,h,j}^{\ell}, \quad \text{for} \quad \ell = 1,\dots,L,$$

where $\hat{k}_{n+h|n,j}$ is the forecast of $k_{n+h,j}$; $\hat{\xi}_{*,h,j}^{\ell}$ is randomly sampled with replacement from $\{\hat{\xi}_{1,h,j},\dots,\hat{\xi}_{n,h,j}\}$; and L is the number of bootstrap replications.

By assuming that the first J principal components approximate the data relatively well, the model residual should contribute nothing but random noise. Consequently, we can bootstrap the model fit error $\hat{e}^{\ell}_{n+h}(x)$ in equation (3) by randomly sampling with replacement from the residual term $\{e_1(x), \dots, e_n(x)\}$. Similarly, we can also bootstrap the independent and identically distributed observational error $\hat{\varepsilon}^{\ell}_{n+h}$ in equation (3) by sampling with replacement from $\{\varepsilon_1, \dots, \varepsilon_n\}$.

Taking into account all possible sources of variability, and assuming that they are not correlated, we obtain L forecast variants of $m_{n+h}(x)$, which can be expressed as:

$$\hat{m}^{\ell}_{n+h|n}(x) = \hat{a}(x) + \sum_{j=1}^{J} b_j(x)\hat{k}^{\ell}_{n+h|n,j} + \hat{e}^{\ell}_{n+h}(x) + \sigma_{n+h}(x)\hat{\varepsilon}^{\ell}_{n+h,i}.$$

The $(1-\alpha)$ prediction intervals are the $\alpha/2$ and $1-\alpha/2$ quantiles of $\left\{\hat{m}^{1}_{n+h|n}(x), \dots, \hat{m}^{L}_{n+h|n}(x)\right\}$.

5.4 Evaluating interval forecast accuracy

The evaluation of the interval forecast accuracy is performed as follows. For data in the forecasting period, prediction intervals were computed at the 80% nominal coverage probability, and were tested to check whether the out-of-sample data fell within the constructed prediction intervals (Swanson & Beck 1994, Tayman et al. 2007). The empirical coverage probability is calculated as the ratio of the number of observations that fall into the calculated prediction intervals to the total number of observations in the forecasting period. Furthermore, we calculated the coverage probability deviance, which is the absolute difference between the nominal coverage probability and the empirical coverage probability, in order to evaluate the interval forecast accuracy of each method. With the nominal coverage probability of $1-\alpha$, the maximum coverage probability deviance is $1-\alpha$, when the empirical coverage probability is 0; the minimum coverage probability deviance is 0, when the empirical coverage probability is $1-\alpha$.

6 Comparisons of the interval forecasts

6.1 Forecast log mortality rates

Variances for the LC methods were calculated as described in Section 5.1. Variances for all HU methods were calculated as described in Section 5.2; and variances for the HU and HUw methods were also calculated by the nonparametric method described in Section 5.3. These are labeled as HU(NP) and HUw(NP). Because the HU and HUw methods give better interval forecast accuracy than others using the parametric approach, we further compare their interval forecast accuracy using the parametric approach with the nonparametric approach at the nominal coverage probability of 0.8.

The average coverage probability deviances for female and male log mortality rates are shown in Tables 6 and 7, respectively. It is clear from these results that the HUw(NP) method performs the best in both the female and the male data.

From these results, we find that the HU methods provide more accurate forecasts of uncertainty than the LC methods. This gain in coverage probability is attributable to the methods' greater statistical sophistication. There are two contributing factors for this. Firstly, the HU methods allow more principal components to be included in the model. These additional components

Country	LC	LCnone	TLB	LM	BMS	HU
Australia	0.57	0.68	0.63	0.45	0.54	0.26
Canada	0.62	0.67	0.54	0.60	0.63	0.28
Denmark	0.72	0.57	0.59	0.64	0.67	0.33
England	0.70	0.40	0.57	0.34	0.45	0.05
Finland	0.73	0.63	0.66	0.62	0.67	0.45
France	0.70	0.39	0.58	0.43	0.50	0.08
Iceland	0.40	0.54	0.56	0.67	0.48	0.31
Italy	0.57	0.27	0.57	0.40	0.47	0.00
Netherlands	0.60	0.39	0.44	0.50	0.52	0.19
Norway	0.74	0.57	0.63	0.64	0.65	0.32
Scotland	0.72	0.62	0.66	0.59	0.55	0.37
Spain	0.62	0.44	0.62	0.45	0.64	0.14
Sweden	0.64	0.35	0.57	0.55	0.64	0.28
Switzerland	0.67	0.59	0.59	0.65	0.66	0.40
Mean	0.64	0.51	0.59	0.54	0.58	0.25
Country	HU(NP)	HU50	HUrob	HUrob50	HUw	HUw(NP)
Australia	0.23	0.31	0.30	0.30	0.05	0.03
Canada	0.19	0.39	0.25	0.40	0.08	0.03
Denmark	0.31	0.47	0.36	0.49	0.20	0.09
England	0.01	0.35	0.12	0.36	0.10	0.11
Finland	0.36	0.53	0.42	0.50	0.32	0.25
France	0.12	0.43	0.17	0.42	0.14	0.13
Iceland	0.31	0.47	0.30	0.46	0.30	0.31
Italy	0.01	0.38	0.07	0.38	0.14	0.14
Netherlands	0.18	0.43	0.22	0.39	0.16	0.12
Norway	0.25	0.43	0.39	0.38	0.13	0.08
Scotland	0.30	0.45	0.40	0.48	0.31	0.20
Spain	0.10	0.46	0.12	0.48	0.09	0.10
Sweden	0.19	0.56	0.28	0.49	0.05	0.01
Switzerland	0.36	0.52	0.42	0.51	0.38	0.34
Mean	0.21	0.44	0.27	0.43	**0.18**	**0.14**

Table 6: *Coverage probability deviances of forecast female log mortality rates by method*

Country	LC	LCnone	TLB	LM	BMS	HU
Australia	0.71	0.71	0.72	0.51	0.61	0.23
Canada	0.74	0.67	0.69	0.64	0.71	0.39
Denmark	0.58	0.59	0.75	0.74	0.76	0.22
England	0.64	0.47	0.66	0.46	0.60	0.07
Finland	0.65	0.54	0.60	0.42	0.67	0.23
France	0.68	0.41	0.67	0.56	0.64	0.12
Iceland	0.42	0.49	0.59	0.71	0.50	0.29
Italy	0.47	0.43	0.64	0.43	0.63	0.06
Netherlands	0.21	0.27	0.63	0.56	0.53	0.17
Norway	0.58	0.58	0.71	0.71	0.75	0.22
Scotland	0.64	0.65	0.69	0.64	0.54	0.45
Spain	0.57	0.44	0.60	0.40	0.63	0.18
Sweden	0.39	0.36	0.67	0.66	0.71	0.31
Switzerland	0.59	0.58	0.65	0.68	0.68	0.44
Mean	0.56	0.51	0.66	0.58	0.64	0.24
Country	HU(NP)	HU50	HUrob	HUrob50	HUw	HUw(NP)
Australia	0.13	0.31	0.26	0.32	0.13	0.07
Canada	0.34	0.45	0.44	0.46	0.21	0.13
Denmark	0.19	0.47	0.24	0.47	0.16	0.05
England	0.02	0.36	0.01	0.40	0.14	0.16
Finland	0.22	0.51	0.18	0.51	0.16	0.06
France	0.04	0.42	0.15	0.38	0.13	0.17
Iceland	0.28	0.48	0.29	0.47	0.27	0.26
Italy	0.04	0.33	0.04	0.38	0.16	0.18
Netherlands	0.19	0.54	0.18	0.52	0.12	0.11
Norway	0.20	0.40	0.23	0.44	0.21	0.19
Scotland	0.39	0.59	0.43	0.58	0.35	0.20
Spain	0.15	0.45	0.13	0.44	0.13	0.13
Sweden	0.11	0.63	0.40	0.66	0.03	0.07
Switzerland	0.42	0.55	0.44	0.53	0.28	0.21
Mean	0.19	0.46	0.24	0.47	0.18	0.14

Table 7: *Coverage probability deviances of forecast male log mortality rates by method*

capture non-random patterns, which are not explained by the first principal component (Booth et al. 2002, Renshaw & Haberman 2003*a*). Secondly, the HU methods use a nonparametric smoothing technique, namely penalized regression splines with the partial monotonic constraint, to smooth the noisy log mortality rates.

Of the HU methods, the HU50 and HUrob50 methods perform least well. This indicates that restricting the fitting period to 1950 onward does not provide a more accurate estimate of the forecast uncertainty. The prediction intervals using data from 1950 onward are too narrow, because of the under-estimated variability. Furthermore, the nonparametric approach of constructing prediction intervals outperforms the parametric approach in both the HU method and the HUw method.

6.2 Forecast life expectancy

The variance of life expectancy was calculated from the variance of mortality rates (Chiang 1984), and prediction intervals for life expectancy were compared. As shown in Tables 8 and 9, the average coverage probability deviances of the female life expectancy and male life expectancy are calculated. The results show that the HUrob method performs the best when both the female and male data are considered.

Again, the HU methods have smaller coverage probability deviances than the LC methods on average. Of the HU methods, the HU50 and HUrob50 methods perform least well, but not in the female data. While the poorer performance for male data might be expected from the similar result for log mortality rates, the good performance for female data is unexpected.

Country	LC	LCnone	TLB	LM	BMS	HU	HU(NP)	HU50	HUrob	HUrob50	HUw	HUw(NP)
Australia	0.17	0.80	0.67	0.00	0.20	0.03	0.03	0.17	0.13	0.03	0.20	0.20
Canada	0.07	0.80	0.27	0.13	0.33	0.20	0.10	0.40	0.30	0.33	0.13	0.17
Denmark	0.60	0.80	0.00	0.27	0.43	0.20	0.20	0.10	0.20	0.07	0.20	0.20
England	0.47	0.77	0.80	0.00	0.07	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Finland	0.30	0.80	0.80	0.53	0.43	0.03	0.03	0.30	0.03	0.17	0.10	0.10
France	0.27	0.80	0.70	0.20	0.13	0.20	0.20	0.03	0.13	0.03	0.20	0.20
Iceland	0.07	0.20	0.23	0.33	0.13	0.20	0.20	0.03	0.20	0.03	0.20	0.20
Italy	0.07	0.77	0.80	0.10	0.13	0.20	0.20	0.07	0.03	0.13	0.20	0.20
Netherlands	0.40	0.80	0.10	0.10	0.20	0.17	0.20	0.07	0.17	0.07	0.17	0.20
Norway	0.57	0.80	0.77	0.43	0.53	0.20	0.20	0.03	0.10	0.13	0.20	0.20
Scotland	0.80	0.80	0.80	0.33	0.37	0.03	0.20	0.30	0.07	0.33	0.13	0.20
Spain	0.50	0.80	0.80	0.40	0.40	0.10	0.20	0.00	0.03	0.10	0.20	0.20
Sweden	0.13	0.77	0.57	0.37	0.33	0.20	0.20	0.23	0.17	0.03	0.20	0.20
Switzerland	0.03	0.80	0.50	0.53	0.47	0.13	0.17	0.13	0.07	0.13	0.17	0.20
Mean	0.32	0.75	0.56	0.27	0.30	0.15	0.17	0.15	0.13	0.13	0.18	0.19

Table 8: *Coverage probability deviances of forecast female life expectancy by method*

Country	LC	LCnone	TLB	LM	BMS	HU
Australia	0.40	0.80	0.80	0.07	0.27	0.07
Canada	0.13	0.80	0.80	0.70	0.70	0.03
Denmark	0.40	0.17	0.73	0.63	0.73	0.10
England	0.40	0.70	0.80	0.33	0.40	0.20
Finland	0.03	0.80	0.80	0.37	0.60	0.07
France	0.07	0.80	0.80	0.67	0.73	0.20
Iceland	0.17	0.10	0.43	0.53	0.17	0.20
Italy	0.07	0.57	0.80	0.17	0.17	0.17
Netherlands	0.07	0.03	0.80	0.70	0.70	0.07
Norway	0.43	0.40	0.70	0.67	0.80	0.07
Scotland	0.70	0.80	0.80	0.60	0.27	0.70
Spain	0.37	0.80	0.80	0.27	0.60	0.07
Sweden	0.30	0.07	0.73	0.70	0.73	0.20
Switzerland	0.03	0.70	0.70	0.67	0.67	0.40
Mean	0.26	0.54	0.75	0.51	0.54	0.18

Country	HU(NP)	HU50	HUrob	HUrob50	HUw	HUw(NP)
Australia	0.13	0.07	0.17	0.10	0.20	0.20
Canada	0.13	0.20	0.33	0.23	0.17	0.20
Denmark	0.03	0.30	0.07	0.33	0.20	0.20
England	0.13	0.10	0.10	0.23	0.20	0.20
Finland	0.07	0.57	0.07	0.57	0.10	0.20
France	0.20	0.27	0.07	0.07	0.20	0.20
Iceland	0.20	0.40	0.20	0.40	0.20	0.20
Italy	0.13	0.17	0.17	0.17	0.20	0.20
Netherlands	0.10	0.63	0.00	0.60	0.20	0.17
Norway	0.17	0.23	0.03	0.37	0.20	0.03
Scotland	0.47	0.80	0.60	0.77	0.20	0.20
Spain	0.10	0.33	0.00	0.23	0.20	0.20
Sweden	0.20	0.70	0.00	0.63	0.20	0.20
Switzerland	0.13	0.53	0.40	0.50	0.20	0.20
Mean	0.16	0.38	0.16	0.37	0.19	0.19

Table 9: *Coverage probability deviances of forecast male life expectancy by method*

Further, it remains unclear whether or not the nonparametric approach of constructing prediction intervals outperforms the parametric approach. In the female data, the parametric approach provides slightly smaller coverage probability deviances than the nonparametric approach for both the HU and HUw methods. In the male data, however, the parametric approach has slightly larger coverage probability deviances than the nonparametric approach for both the HU and HUw methods.

7 Discussion

There has been a large number of studies on forecasting mortality rates and life expectancy. However, there has been limited attention given to comparing the various proposed forecasting methods, with the exceptions of Booth et al. (2002, 2005, 2006). This article has conducted comparative evaluations of point and interval forecasts for log mortality rates and life expectancy based on ten principal component methods and 28 data sets. The methods include the Lee-Carter method and four LC variants, and the Hyndman-Ullah method, itself an extension of Lee-Carter, and four HU variants.

The comparison of point forecasts is an expansion of Booth et al. (2006). Our work differs from Booth et al. (2006) by expanding the number of LC variants and by including the recently-developed variants of the HU method. The finding that the LC method performs least well for forecasting log mortality rates is in keeping with Booth et al. (2005, 2006). Among the HU methods, the HUw method performs best for forecasting log mortality rates. This can be attributed to the fact that the HUw method discounts the effect of past data in the estimation of the principal components. The LM method performs the best for forecasting life expectancy in both the female and male data; this finding supports the findings of Booth et al. (2005, 2006).

Although Booth et al. (2002) compared the LC, LM and BMS methods based on standard errors and prediction intervals, they did not compare interval forecast accuracy. This article is among the first to systematically compare interval forecast accuracy based on coverage probability deviances. The use of coverage probability deviances allows comparisons of interval forecast accuracy for each method by measuring the differences between the empirical coverage probabilities and the nominal coverage probability. For forecasting log mortality rates, the HUw method produces the most accurate coverage probability in both the female and male data, particularly when the variance is estimated by the nonparametric approach. The nonparametric approach of estimating variance does not assume normality that may not be satisfied in practice. For forecasting life expectancy, the HUrob method produces the most accurate coverage probability in both the female and male data. The better interval forecast accuracy of the HU methods over the LC methods may be due to the increased forecast uncertainty when the four sources of error are considered (Alho & Spencer 2005, p. 255). In comparison with the prediction intervals of LC methods, the prediction intervals of HU methods are generally wider.

The conclusions from this comparative analysis are based on country averages. They are broad generalizations and do not necessarily hold in all situations. Indeed, examples can be seen in the tables of methods that perform relatively poorly on average but perform well for a particular country, and conversely of methods that perform relatively well on average but perform poorly for a particular country. Further research is needed to understand the circumstances in which this occurs.

Implementation of the aforementioned methods is straightforward using the readily-available *demography* package of Hyndman et al. (2009), provided that historical mortality rates are available in a complete matrix format.

8 Acknowledgments

The first author would like to acknowledge the financial support from the postgraduate publication award of Monash University.

References

Alho, J. (1998), A stochastic forecast of the population of Finland, Reviews 1998/4, Statistics Finland, Helsinki.

Alho, J. M. (1997), 'Scenarios, uncertainty and conditional forecasts of the world population', *Journal of the Royal Statistical Society: Series A* **160**(1), 71–85.

Alho, J. M. & Spencer, B. D. (2005), *Statistical demography and forecasting*, Springer, New York.

Bongaarts, J. (2005), 'Long-range trends in adult mortality: models and projection methods', *Demography* **42**(1), 23–49.

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

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

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

Booth, H. & Tickle, L. (2008), 'Mortality modelling and forecasting: a review of methods', *Annals of Actuarial Science* **3**(1-2), 3–43.

Booth, H., Tickle, L. & Smith, L. (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., Jenkins, G. M. & Reinsel, G. C. (2008), *Time series analysis: forecasting and control*, 4th edn, Wiley, Hoboken, New Jersey.

Brouhns, N., Denuit, M. & Van Keilegom, I. (2005), 'Bootstrapping the Poisson log-bilinear model for mortality forecasting', *Scandinavian Actuarial Journal* **2005**(3), 212–224.

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

Cardot, H., Ferraty, F. & Sarda, P. (2003), 'Spline estimators for the functional linear model', *Statistica Sinica* **13**(3), 571–591.

Carter, L. R. & Prskawetz, A. (2001), Examining structural shifts in mortality using the Lee-Carter method, Working paper 2001-007, Max Planck Institute for Demographic Research.
URL: `http://www.demogr.mpg.de/Papers/Working/wp-2001-007.pdf`

Chatfield, C. (1993), 'Calculating interval forecasts', *Journal of Business & Economic Statistics* **11**(2), 121–135.

Chatfield, C. (2000), *Time-series forecasting*, Chapman & Hall/CRC, Boca Raton, Florida.

Chiang, C. L. (1984), *The life table and its applications*, R.E. Krieger Pub. Co., Malabar, Florida.

Currie, I. D., Durban, M. & Eilers, P. H. C. (2004), 'Smoothing and forecasting mortality rates', *Statistical Modelling* **4**(4), 279–298.

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

Debón, A., Montes, F. & Sala, R. (2006), 'A comparison of models for dynamic life tables. Application to mortality data from the Valencia Region (Spain)', *Lifetime Data Analysis* **12**(2), 223–244.

Ediev, D. M. (2008), Extrapolative projections of mortality: towards a more consistent method, Working paper 3/2008, Vienna Institute of Demography.
URL: `http://www.oeaw.ac.at/vid/download/WP2008_03.pdf`

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

Felipe, A., Guillén, M. & Pérez-Marín, A. M. (2002), 'Recent mortality trends in the Spanish population', *British Actuarial Journal* **8**(4), 757–786.

Girosi, F. & King, G. (2008), *Demographic forecasting*, Princeton University Press, Princeton.

Haberman, S. & Renshaw, A. (2008), 'Mortality, longevity and experiments with the Lee-Carter model', *Lifetime Data Analysis* **14**(3), 286–315.

Human Mortality Database (2009), *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).* Accessed at 15/4/2009.
URL: `http://www.mortality.org/`

Hyndman, R. J. & Booth, H. (2008), 'Stochastic population forecasts using functional data models for mortality, fertility and migration', *International Journal of Forecasting* **24**(3), 323–342.

Hyndman, R. J., Booth, H., Tickle, L. & Maindonald, J. (2009), *demography: Forecasting mortality and fertility data*. R package version 0.995.
URL: `http://robjhyndman.com/software/demography`

Hyndman, R. J., Koehler, A. B., Ord, J. K. & Snyder, R. D. (2008), *Forecasting with exponential smoothing: the state space approach*, Springer, Berlin.

Hyndman, R. J. & Shang, H. L. (2009), 'Forecasting functional time series (with discussion)', *Journal of the Korean Statistical Society* **38**(3), 199–221.

Hyndman, R. J. & Ullah, M. S. (2007), 'Robust forecasting of mortality and fertility rates: a functional data approach', *Computational Statistics & Data Analysis* **51**(10), 4942–4956.

Keilman, N., Pham, D. & Hetland, A. (2002), 'Why population forecasts should be probabilistic—illustrated by the case of Norway', *Demographic Research* **6**, 409–454.

Koissi, M.-C., Shapiro, A. F. & Högnäs, G. (2006), 'Evaluating and extending the Lee-Carter model for mortality forecasting: bootstrap confidence interval', *Insurance: Mathematics and Economics* **38**(1), 1–20.

Lazar, D. & Denuit, M. M. (2009), 'A multivariate time series approach to projected life tables', *Applied Stochastic Models in Business and Industry* **25**(6), 806–823.

Lee, R. D. & Carter, L. R. (1992), 'Modeling and forecasting U.S. mortality', *Journal of the American Statistical Association* **87**(419), 659–671.

Lee, R. D. & Miller, T. (2001), 'Evaluating the performance of the Lee-Carter method for forecasting mortality', *Demography* **38**(4), 537–549.

Lee, R. D. & Nault, F. (1993), Modeling and forecasting provincial mortality in Canada, *in* 'World Congress of the International Union for the Scientific Study of Population', Montreal, Canada.

Lee, R. D. & Rofman, R. (1994), 'Modeling and projecting mortality in Chile', *Notas de Poblacion* **22**(59), 183–213.

Lin, J. (1995), 'Changing kinship structure and its implications for old-age support in urban and rural China', *Population Studies* **49**(1), 127–145.

Lundström, H. & Qvist, J. (2004), 'Mortality forecasting and trend shifts: an application of the Lee-Carter model to Swedish mortality data', *International Statistical Review/Revue Internationale de Statistique* **72**(1), 37–50.

Lutz, W. & Goldstein, J. R. (2004), 'Introduction: how to deal with uncertainty in population forecasting?', *International Statistical Review/Revue Internationale de Statistique* **72**(1), 1–4.

Ramsay, J. O. & Silverman, B. W. (2005), *Functional data analysis*, 2nd edn, Springer, New York.

Renshaw, A. E. & Haberman, S. (2003*a*), 'Lee-Carter mortality forecasting: a parallel generalized linear modelling approach for England and Wales mortality projections', *Journal of the Royal Statistical Society: Series C* **52**(1), 119–137.

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

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

Renshaw, A. E. & Haberman, S. (2006), 'A cohort-based extension to the Lee-Carter model for mortality reduction factors', *Insurance: Mathematics and Economics* **38**(3), 556–570.

Renshaw, A. E. & Haberman, S. (2008), 'On simulation-based approaches to risk measurement in mortality with specific reference to Poisson Lee-Carter modelling', *Insurance: Mathematics and Economics* **42**(2), 797–816.

Swanson, D. A. & Beck, D. M. (1994), 'A new short-term county population projection method', *Journal of Economic and Social Measurement* **20**(1), 25–50.

Tayman, J., Schafer, E. & Carter, L. (1998), 'The role of population size in the determination and prediction of population forecast errors: an evaluation using confidence intervals for subcounty areas', *Population Research and Policy Review* **17**(1), 1–20.

Tayman, J., Smith, S. K. & Lin, J. (2007), 'Precision, bias, and uncertainty for state population forecasts: an exploratory analysis of time series models', *Population Research and Policy Review* **26**(3), 347–369.

Tuljapurkar, S. (2005), Stochastic forecasts of mortality, population, and pension systems, *in* N. Keilman, ed., 'Perspectives on Mortality Forecasting', Vol. II: Probabilistic Models, Swedish Social Insurance Agency, Stockholm, pp. 65–77.

Tuljapurkar, S., Li, N. & Boe, C. (2000), 'A universal pattern of mortality decline in the G7 countries', *Nature* **405**(6788), 789–792.

Wang, J. Z. (2007), Fitting and forecasting mortality for Sweden: applying the Lee-Carter model, Master's thesis, Stockholm University.
URL: *http://www2.math.su.se/matstat/reports/serieb/2007/rep1/report.pdf*

Wilmoth, J. R. (1996), Mortality projections for Japan: a comparison of four methods, *in* G. Caselli & A. D. Lopez, eds, 'Health and Mortality among Elderly Populations', Clarendon Press, Oxford, pp. 266–287.

Wohlfart, P. (2006), Mortality predictions for longevity analysis and annuity valuation, Master's thesis, Stockholm University.
URL: *http://www2.math.su.se/matstat/reports/serieb/2006/rep11/report.pdf*