Robust forecasting of mortality and fertility rates: a functional data approach

Rob J Hyndman* and Md Shahid Ullah

Department of Econometrics and Business Statistics,
Monash University, VIC 3800, Australia.

Abstract

A new method is proposed for forecasting age-specific mortality and fertility rates observed over time. This approach allows for smooth functions of age, is robust for outlying years due to wars and epidemics, and provides a modelling framework that is easily adapted to allow for constraints and other information. Ideas from functional data analysis, nonparametric smoothing and robust statistics are combined to form a methodology that is widely applicable to any functional time series data observed discretely and possibly with error. The model is a generalization of the Lee-Carter model commonly used in mortality and fertility forecasting. The methodology is applied to French mortality data and Australian fertility data, and the forecasts obtained are shown to be superior to those from the Lee-Carter method and several of its variants.

Key words: Fertility forecasting, functional data, mortality forecasting, nonparametric smoothing, principal components, robustness.

* Corresponding author: Professor Rob Hyndman, Department of Econometrics and Business Statistics, Monash University, VIC 3800, Australia. Telephone: +61 3 9905 2358. Fax: +61 3 9905 5474. Email: Rob.Hyndman@buseco.monash.edu.au

1 Introduction

In this paper, we propose a new approach to forecasting age-specific mortality and fertility rates that combines ideas from functional data analysis, nonparametric smoothing and robust statistics. While our methodology could be applied to forecasting functional time series in other contexts, we restrict our attention here to mortality and fertility forecasting.

There has been a surge of interest in this problem in the last few years, driven by the need for good forecasts to inform government policy and planning. Fundamental changes in welfare policy are taking place in many countries as a result of forecasts of an increasing elderly population. These age-specific population forecasts rely on age-specific forecasts of mortality and fertility rates. Therefore, any improvements in mortality and fertility forecasting have an immediate impact in guiding policy decisions regarding the allocation of current and future resources. Future mortality rates are also of great interest in the insurance and pensions industry, and fertility forecasts are of interest to governments in planning children's services.

Recently, several authors have proposed new approaches to mortality forecasting utilizing smoothing and statistical modelling. We add to this literature with another approach that differs from the existing proposals in that it treats the underlying process as functional, and provides estimation and forecasting that are robust to outliers. Our methodology also applies to the fertility forecasting problem, and to other areas where forecasts of functional data are required.

We review some of this literature below, but first we define the problem more precisely. Let $y_t(x)$ denote the log of the observed mortality or fertility rate for age x in year t. We assume there is an underlying smooth function $f_t(x)$ that we are observing with error and at discrete (and possibly sparse) points of x. Our observations are $\{x_i, y_t(x_i)\}$, $t = 1, \dots, n$, $i = 1, \dots, p$ where

$$y_t(x_i) = f_t(x_i) + \sigma_t(x_i)\varepsilon_{t,i}\,, \tag{1}$$

$\varepsilon_{t,i}$ is an iid standard normal random variable and $\sigma_t(x_i)$ allows the amount of noise to vary with x. Typically $\{x_1, \dots, x_p\}$ are single years of age ($x_1 = 0$, $x_2 = 1, \dots$) or denote 5-year age groups. We are interested in forecasting $y_t(x)$ for $x \in [x_1, x_p]$ and $t = n+1, \dots, n+h$.

Note that the data are not directly of a functional nature, but that we assume there are underlying functional time series which we are observing with error at discrete points.

Figure 1 shows an example of such data for France. (Age-specific mortality

Fig. 1. *Five years of data showing male death rates in France. Note the changing shape in the curve over time and the increased variance for high ages. The year 1918 was an outlier due to the Spanish flu pandemic and World War I.*

rates can be higher than one as they are traditionally computed as the total number of deaths of people aged x in year t divided by the estimated number of people of age x in the *middle* of the year. See Wilmoth, 2002.) The observational error is clear from the figure, as is the increasing variance for higher ages, especially for $x > 100$. It is also apparent that the dynamic behaviour of the underlying curves is relatively complicated with, for example, the 'bump' around 18–19 years higher relative to nearby ages in 1977 than in the other years plotted. Similarly, the general drop in mortality over time (largely due to improvements in hygiene, diet and medical interventions) is not uniform over either age or time. Furthermore, there are some 'outlier' years such as 1918; this particular year was unusual due to the large number of 15–50 year old deaths as a result of World War I and the effects of the Spanish flu pandemic. Clearly, any forecasts of these data will need to be able to model the complex dynamic behaviour and be robust to outlying years.

Our approach is a natural extension of methods for mortality and fertility forecasting that have evolved over the last two decades. An important milestone during that period was the publication of Lee and Carter (1992); they proposed a methodology for modelling and extrapolating long-term trends in mortality rates and used it to make forecasts of US mortality to 2065. The methodology has since become very widely used and there have been various extensions and modifications proposed (e.g., Lee and Miller, 2001; Booth

et al., 2002; Renshaw and Haberman, 2003; Wolf, 2004). The methodology has also been applied to fertility by Lee (1993) and others.

The Lee-Carter method involves using the first principal component of the log-mortality (or fertility) matrix with (i,t)th element $\{y_t(x_i)\}$. Our proposed methodology can be considered a successor to Lee and Carter in that it also involves a principal component decomposition of the mortality or fertility rates. It also draws on the approaches of Bozik and Bell (1987) and Bell and Monsell (1991).

However, we differ from these authors in several important respects. First, we use the functional data paradigm (Ramsay and Silverman, 2005) to frame our methodology. This immediately leads to the use of nonparametric smoothing to reduce some of the inherent randomness in the observed data. It also avoids problems associated with data grouped into age intervals. Second, we propose a robust version of principal components to avoid difficulties with outlying years.

Li and Chan (2005) provide an interesting discussion of the cause of the observed outliers and propose a robust version of the Lee-Carter methodology to cope with them. However, their method does not utilize the advantageous functional data paradigm to give smooth results, and they use only one principal component.

Our approach is summarized below.

(1) Smooth the data for each t using a nonparametric smoothing method to estimate $f_t(x)$ for $x \in [x_1, x_p]$ from $\{x_i, y_t(x_i)\}$, $i = 1, 2, \ldots, p$.
(2) Decompose the fitted curves via a basis function expansion using the following model:

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\, \phi_k(x) + e_t(x) \tag{2}$$

where $\mu(x)$ is a measure of location of $f_t(x)$, $\{\phi_k(x)\}$ is a set of orthonormal basis functions and $e_t(x) \sim \mathrm{N}(0, v(x))$.
(3) Fit univariate time series models to each of the coefficients $\{\beta_{t,k}\}$, $k = 1, \ldots, K$.
(4) Forecast the coefficients $\{\beta_{t,k}\}$, $k = 1, \ldots, K$, for $t = n+1, \ldots, n+h$ using the fitted time series models.
(5) Use the forecast coefficients with (2) to obtain forecasts of $f_t(x)$, $t = n+1, \ldots, n+h$. From (1), forecasts of $f_t(x)$ are also forecasts of $y_t(x)$.
(6) The estimated variances of the error terms in (2) and (1) are used to compute prediction intervals for the forecasts.

Because of the outlying years, we shall use robust estimation for the age func-

tions $\mu(x)$ and $\phi_k(x)$. The time series $\{\beta_{t,k}\}$ will not be estimated robustly, so that any outlying years will be modelled by outliers in the time series. This enables the outlying years to be clearly identified, and the time series models used in step 3 will be chosen to allow for outliers.

At first it might be thought that model (2) should also have a separate additive time term $\beta_{t,0}$ along with the age term $\mu(x)$. Then the model would be similar to a two-way analysis of variance with $\mu(x)$ and $\beta_{t,0}$ representing the two main effects, and the interaction modelled as the sum of product terms. However, we omit the time term as it complicates the forecasting because it is then difficult to ensure $\beta_{t,0}$ is uncorrelated with $\beta_{t,k}$ for $k \geq 1$. We want to have $\beta_{t,k}$ uncorrelated with $\beta_{t,j}$ for all $j \neq k$, so that multivariate time series models are not required.

The nonparametric smoothing (step 1 above) can be undertaken using one of the existing smoothing methods (e.g., Ruppert et al., 2003; or Simonoff, 1996). We prefer constrained and weighted penalized regression splines for mortality and fertility data. The weighting takes care of the heterogeneity due to $\sigma_t(x)$ and we apply either a monotonic constraint (for mortality rates) or a concavity constraint (for fertility rates). These constraints and the smoothing methods used are described in more detail in Section 2.

The decomposition using an orthonormal basis (step 2) is achieved using functional principal components analysis based on Dauxois et al. (1982) and further developed by Ramsay and Dalzell (1991). We discuss the application of this approach in a time series context in Section 3, using a robust method for principal components to avoid problems with outlying observations.

Silverman (1996) discusses a smoothed version of principal component analysis for functional data, but we prefer smoothing the observed data first rather than smoothing the principal components directly, as it allows us to place relevant constraints on the smoothing more easily.

Locantore et al. (1999) also proposes a robust approach to functional principal component analysis, developed for analysing some ophthalmology data. Our approach for making the principal components decomposition robust is much more computationally efficient than their method (see Croux's contribution to the discussion of Locantore et al., 1999).

Valderrama et al. (2002) propose a special case of the above algorithm for forecasting seasonal data, where each curve $f_t(x)$ corresponds to one year of a seasonal time series and x denotes the season. They interpolate rather than smooth the observed data (step 1). Our approach extends their results in several ways including a more general framework allowing application to a broader class of problems, the use of robust estimation and forecasting, derivation of forecast variances and a proposed method for choosing an appropriate order.

The nonparametric functional data analysis discussed by Ferraty and Vieu (2004) differs from our approach in that they look at the problem with a scalar response variable and a functional covariate. Our problem involves a functional response variable observed at discrete values of the argument and at regular intervals in time.

Section 4 describes steps 3–6 including forecasting the coefficients $\{\beta_{t,k}\}$ and the construction of the forecasts and forecast intervals for $y_t(x)$. We apply the ideas to Australian fertility and French mortality data in Section 5. Finally, in Section 6 we explore connections between our proposed methodology and some of the methods that have been suggested in the demographic literature. We also show how the model can be easily extended to allow for cohort effects and explanatory variables.

There are two recent proposals for mortality forecasting that also incorporate nonparametric smoothing. De Jong and Tickle (2006) show how to combine spline smoothing and estimation via the Kalman filter to fit a generalized version of the Lee-Carter model. Currie et al. (2004) employ bivariate penalized B-splines to smooth the mortality surface in both the time and age direction. They then model the numbers of deaths using a Poisson distribution rather than the mortality rates that we choose to model. The forecasting is achieved by extrapolating the fitted surface in the time direction. They assume the observed data are independent, so any serial correlation in the time dimension is presumably soaked up by the smoothing procedure. Neither of these approaches is robust and so do not cater for unusual years. Nor can they easily incorporate qualitative constraints such as monotonicity or concavity on the smoothing. They are also more difficult to generalize for cohort effects and explanatory variables.

2 Constrained and weighted smoothing

Let $m_t(x)$ denote the observed mortality rate for age x in year t and define $N_t(x)$ to be the total population of age x at 30 June in year t. Then $m_t(x)$ is approximately binomially distributed with estimated variance $N_t^{-1}(x)m_t(x)[1 - m_t(x)]$. So the variance of $y_t(x) = \log[m_t(x)]$ is (via a Taylor approximation)

$$\hat{\sigma}_t^2(x) \approx [1 - m_t(x)]N_t^{-1}(x)m_t^{-1}(x). \tag{3}$$

We define weights equal to the inverse variances $w_t(x) = N_t(x)m_t(x)/[1 - m_t(x)]$ and use weighted penalized regression splines (Wood, 2003; He and Ng, 1999) to estimate the curve $f_t(x)$ in each year.

For fertility data, let $p_t(x)$ denote the observed fertility rate per thousand

women for mothers of age x in year t and $N_t(x)$ is the female resident population of age x at 30 June in year t. Then, using a similar approach,

$$\hat{\sigma}_t^2(x) \approx [1000 - p_t(x)]N_t^{-1}(x)p_t^{-1}(x). \tag{4}$$

We apply a qualitative constraint to obtain better estimates of $f_t(x)$, especially when $\sigma_t(x)$ is large. For mortality data, we assume that $f_t(x)$ is monotonically increasing for $x > c$ for some c (say 50 years). This monotonicity constraint allows us to reduce the noise in the estimated curves for high ages, and is not unreasonable for this application (the older you are, the more likely you are to die). We use a modified version of the approach described in Wood (1994) to implement the monotonicity constraint.

For fertility data, we constrain the fitted curves to be concave. Again, this seems reasonable and is satisfied for all the fertility data we have seen. He and Ng (1999) provide a method for implementing this constraint.

3 Robust functional principal components

There are many ways the basis functions (in step 2 above) could be chosen. However, the optimal orthonormal basis set is obtained via principal components (see Ramsay and Silverman, 2005, pp.151–152). For a given K, this gives the basis functions $\{\phi_k(x)\}$ which minimize the Mean Integrated Squared Error

$$\text{MISE} = \frac{1}{n}\sum_{t=1}^{n}\int e_t^2(x)\,dx.$$

Thus, this basis set provides the best fit to the estimated curves. It also enables more informative interpretations and gives coefficients that are uncorrelated, thus simplifying the forecasting process.

Because we seek a robust estimate of $\mu(x)$, we shall use the L_1-median of the estimated smooth curves $\{\hat{f}_1(x), \ldots, \hat{f}_n(x)\}$, given by

$$\hat{\mu}(x) = \underset{\theta(x)}{\text{argmin}} \sum_{t=1}^{n} \|\hat{f}_t(x) - \theta(x)\|$$

where $\|g(x)\| = \left(\int g^2(x)\,dx\right)^{1/2}$ denotes the norm of the function g. The algorithm of Hössjer and Croux (1995) can be used to compute $\hat{\mu}(x)$ on a fine grid of points. The median-adjusted data is denoted by $\hat{f}_t^*(x) = \hat{f}_t(x) - \hat{\mu}(x)$.

We now introduce two methods for obtaining robust principal components for $\{\hat{f}_t^*(x)\}$: one uses a weighted approach and the other is based on a projection

pursuit algorithm. Then we will combine the two methods to obtain an efficient and robust method for obtaining the basis functions $\{\phi_k(x)\}$.

3.1 Weighted principal components

Our aim is to find the functions $\phi_k(x)$ such that the variance of the scores

$$z_{t,k} = w_t \int \phi_k(x) \hat{f}_t^*(x)\, dx \tag{5}$$

is maximized subject to the constraints

$$\int \phi_k^2(x)\, dx = 1 \qquad \text{and} \qquad \int \phi_k(x)\phi_{k-1}(x)\, dx = 0 \qquad \text{if } k \geq 2. \tag{6}$$

These are defined iteratively for $k = 1, \ldots, K$ where $K \leq n - 1$. The weights w_t are chosen so that outlying observations receive low weight. (Exactly how this is done is explained in Section 3.3.)

We proceed in a similar manner to (Ramsay and Silverman, 2005, Chapter 8). Suppose that we can write each adjusted smoothed function $\hat{f}_t^*(x)$ in an alternative basis expansion

$$\hat{f}_t^*(x) = \sum_{j=1}^{m} a_{t,j}\, \xi_j(x)$$

and let $\boldsymbol{A} = (a_{t,j})$ denote the $n \times m$ matrix of coefficients. This basis expansion arises naturally if the computation of $\hat{f}_t(x)$ is achieved using regression splines. Now let $\boldsymbol{J}$ be the $m \times m$ matrix with (i,k)th element $J_{ik} = \int \xi_i(x)\, \xi_k(x)\, dx$ and find the Choleski decomposition $\boldsymbol{J} = \boldsymbol{U}'\boldsymbol{U}$. Then

$$\phi_k(x) = (\boldsymbol{U}^{-1}\boldsymbol{g}^{(k)})'\boldsymbol{\xi}(x)\,. \tag{7}$$

where $\boldsymbol{g}^{(k)}$ is the kth normalized eigenvector of $(\boldsymbol{U}^{-1})'\boldsymbol{J}\boldsymbol{S}\boldsymbol{J}'\boldsymbol{U}^{-1}$, $\boldsymbol{S} = (n-1)^{-1}\boldsymbol{A}'\boldsymbol{W}^2\boldsymbol{A}$, $\boldsymbol{W} = \text{diagonal}(w_1, \ldots, w_n)$, and $\boldsymbol{\xi}(x) = [\xi_1(x), \ldots, \xi_m(x)]'$. If $\boldsymbol{\Phi}$ is the $p \times (n-1)$ matrix with (i,k)th value $\phi_k(x_i)$, and $\boldsymbol{F}$ is the $n \times p$ matrix with (t,i)th element $\hat{f}_t^*(x_i)$ then $\hat{\beta}_{t,k}$ is the (t,k)th element of $\boldsymbol{B} = \boldsymbol{F}\boldsymbol{\Phi}$.

A computationally simpler procedure is to discretize the smooth functions $\hat{f}_t^*(x)$ on a fine grid of q equally spaced values $\{x_1^*, \ldots, x_q^*\}$ that span the interval $[x_1, x_p]$. Denote the resulting $n \times q$ matrix by $\boldsymbol{G}$ and let $\boldsymbol{G}^* = \boldsymbol{W}\boldsymbol{G}$ where $\boldsymbol{W} = \text{diagonal}(w_1, \ldots, w_n)$. Then the singular value decomposition of $\boldsymbol{G}^*$ gives $\boldsymbol{G}^* = \boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{V}'$ where $\hat{\phi}_k(x_j^*)$ is the (j,k)th element of $\boldsymbol{\Psi}$. If $\boldsymbol{B} = \boldsymbol{G}\boldsymbol{\Psi}$, then $\hat{\beta}_{t,k}$ is the (t,k)th element of $\boldsymbol{B}$. Other values of $\phi_k(x)$ can be computed using linear interpolation.

Jolliffe (2002) describes several proposals for obtaining robust principal components. However, most of these are based on robust estimates of the sample covariance matrix. In the context of functional data, this approach poses difficulties because the covariance matrix is of infinite dimension. Instead, we adopt the projection-pursuit approach of Li and Chen (1985).

Here, our aim is to find the functions $\phi_k(x)$ such that the dispersion of the scores (5) with equal weights is maximized subject to the constraints (6). We measure the dispersion by the function $S(z_{1,k}, \dots, z_{n,k})$. Note that if S is taken to be the sample variance, then this is identical to the definition of principal components given by Ramsay and Silverman (2005, p.149). Other measures of dispersion will lead to alternative decompositions analogous to the classical procedure. The remarkable properties of variance enable very rapid computation via singular value decomposition as described earlier. With other measures of dispersion, computation is more difficult, and the values of $\phi_k(x)$ must be obtained iteratively for $k = 1, 2, \dots$.

The algorithm of Li and Chen (1985) is as robust as the dispersion measure used. However, it is extremely slow and several authors (beginning with Croux and Ruiz-Gazen, 1996, 2005) have sought to find more computationally attractive approaches. To our knowledge, the best current solution is the RAPCA algorithm of Hubert et al. (2002) as it is more numerically stable and faster for high-dimensional data than the other proposals for obtaining projection-pursuit estimates of principal components. The RAPCA algorithm uses the first quartile of pairwise differences as the measure of dispersion; thus

$$S(z_{1,k}, \dots, z_{n,k}) = 2.2219 c_n \{|z_{i,k} - z_{j.k}|; i < j\}_{(\tau)}$$

where $\tau = \binom{\lfloor n/2 \rfloor + 1}{2}$ and c_n is a small-sample correction factor to make S an unbiased estimator.

The Li-Chen algorithm (and its derivatives, including RAPCA) are designed for discrete multivariate data, rather than functional data. However, we can use a fine discretization of the data to obtain approximate solutions. Applying the RAPCA algorithm to the matrix G gives functions $\phi_1(x), \dots, \phi_K(x)$ computed on the grid of points $x \in \{x_1^*, \dots, x_q^*\}$.

One drawback with this approach is that the RAPCA algorithm tends to give basis function estimates that have higher curvature than the original components $\{\hat{f}_t(x)\}$. Consequently, we seek a combination of the projection-pursuit approach with the weighted approach to principal component analysis, retaining the best features of each. This also has the advantage of higher efficiency.

3.3 Two-step algorithm for functional principal components

We propose a combination of the weighted principal component method and the RAPCA algorithm to give the following two-step procedure for robust functional principal component analysis:

(1) Use the RAPCA (projection pursuit) algorithm described in Section 3.2 to obtain initial (highly robust) values for $\{\hat{\beta}_{t,k}\}$ and $\{\phi_k(x)\}$ ($t = 1, \dots, n$; $k = 1, \dots, K$).
(2) Define the Integrated Squared Error for period t as

$$v_t = \int_x \left(\hat{f}_t^*(x) - \sum_{k=1}^{K} \hat{\beta}_{t,k}\phi_k(x) \right)^2 dx \,.$$

This provides a measure of the accuracy of the principal component approximation for year t. We then compute the weights $w_t = 1$ if $v_t < s + \lambda\sqrt{s}$ and 0 otherwise, where s is the median of $\{v_1, \dots, v_t\}$ and $\lambda > 0$ is a tuning parameter to control the degree of robustness.

Use these weights to obtain new estimates of $\{\hat{\beta}_{t,k}\}$ and $\{\phi_k(x)\}$ using the method described in Section 3.1.

The choice of weight function is motivated by assuming $e_t(x)$ is normally distributed for large enough K. Then v_t will have a χ^2 distribution and $\mathrm{E}(v_t) = \mathrm{Var}(v_t)/2$. Therefore, using a normal approximation (valid for large degrees of freedom), the probability that $v_t < s + \lambda\sqrt{s}$ is approximately $\Phi(\lambda/\sqrt{2})$ where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. This is also the efficiency of the algorithm compared to the classical approach when $w_t = 1$ for all t. For example, with $\lambda = 3$ the efficiency is $\Phi(3/\sqrt{2}) = 98.3\%$, and $\lambda = 4$ gives an efficiency of 99.8%.

If $\lambda \to \infty$, then $w_t = 1$ for all t and so the estimates in Step 3 are the same as those using the method described in Ramsay and Silverman (2005). On the other hand, the smaller the value of λ, the more outliers are omitted. For $\lambda > 0$, our algorithm retains the breakdown point of the RAPCA algorithm, namely 50%.

It may be argued that it would be better to use continuous weights rather than binary weights. However, for demographic data, outliers tend to be associated with specific, relatively short-term events such as wars and epidemics. Consequently, one wishes to identify and remove data associated with those years, but retain all data associated with other years. Therefore, we prefer to use binary weights for this application. Continuous weights (e.g., $w_t = v_t^{-1}$) may be better suited to other applications.

4 Forecasting

The two-step robust principal components method described in the previous section yields the decomposition

$$\hat{f}_t(x) = \hat{\mu}(x) + \sum_{k=1}^{K} \hat{\beta}_{t,k}\, \phi_k(x) + \hat{e}_t(x).$$

Thus we need to forecast $\beta_{t,k}$ for $k = 1, \ldots, K$ and $t = n+1, \ldots, n+h$. For $K > 1$ this is a multivariate time series problem. However, because of the way the basis functions $\phi_k(x)$ have been chosen, the coefficients $\hat{\beta}_{t,k}$ and $\hat{\beta}_{t,\ell}$ are uncorrelated for $k \neq \ell$. Therefore it is likely that univariate methods will be adequate for forecasting each series $\{\hat{\beta}_{t,k}\}$, for $k = 1, \ldots, K$. There may still be cross-correlations at non-zero lags, but these are likely to be small given the zero contemporaneous correlations. We explore this issue in the examples of Section 5.

To obtain forecasts of each coefficient, we use univariate robust ARIMA models (Chen and Liu, 1993) for each series of coefficients $\{\beta_{1,k}, \ldots, \beta_{n,k}\}$, $k = 1, \ldots, K$. In particular, for those time periods where there are outlying observations, the coefficients $\hat{\beta}_{t,k}$ are likely to be very different from the coefficients associated with times where $w_t = 1$. The method of Chen and Liu (1993) allows the fitted ARIMA models to contain outliers of various types so that these unusual observations do not contaminate the forecasts.

4.1 Forecasting the functions

Now combining (1) with (2) we obtain

$$y_t(x_i) = \mu(x_i) + \sum_{k=1}^{K} \beta_{t,k}\, \phi_k(x_i) + e_t(x_i) + \sigma_t(x_i)\varepsilon_{t,i}\,. \tag{8}$$

Then, conditioning on the observed data $\mathcal{I} = \{y_t(x_i);\ t = 1, \ldots, n;\ i = 1, \ldots, p\}$ and the set of basis functions $\mathbf{\Phi}$, we obtain h-step ahead forecasts of $y_{n+h}(x)$:

$$\hat{y}_{n,h}(x) = \mathrm{E}[y_{n+h}(x) \mid \mathcal{I}, \mathbf{\Phi}] = \hat{\mu}(x) + \sum_{k=1}^{K} \tilde{\beta}_{n,k,h}\, \hat{\phi}_k(x)\,. \tag{9}$$

where $\tilde{\beta}_{n,k,h}$ denotes the h-step ahead forecast of $\beta_{n+h,k}$ using the estimated time series $\hat{\beta}_{1,k}, \ldots, \hat{\beta}_{n,k}$.

The forecast variance also follows from (8). Because of the way the model has been constructed, each component is approximately orthogonal to the other

components and so the forecast variance can be approximated by the simple sum of component variances:

$$\zeta_{n,h}(x) = \text{Var}[y_{n+h}(x) \mid \mathcal{I}, \mathbf{\Phi}] \approx \hat{\sigma}^2_\mu(x) + \sum_{k=1}^{K} u_{n+h,k}\, \hat{\phi}^2_k(x) + v(x) + \sigma^2_{n+h}(x) \quad (10)$$

where $u_{n+h,k} = \text{Var}(\beta_{n+h,k} \mid \beta_{1,k}, \ldots, \beta_{n,k})$ can be obtained from the time series model, and $\hat{\sigma}^2_\mu(x)$ (the variance of the smooth estimate $\hat{\mu}(x)$) can be obtained from the smoothing method used. The observational error variance $\sigma^2_t(x)$ is computed using the approximations (3) or (4). The model error variance $v(x)$ is estimated by averaging $\hat{e}^2_t(x)$ for each x.

Then, assuming the various sources of error are all normally distributed, a $100(1-\alpha)\%$ prediction interval for $y_t(x)$ is constructed as $\hat{y}_{n,h}(x) \pm z_\alpha \sqrt{\zeta_{n,h}(x)}$, where z_α is the $1 - \alpha/2$ standard normal quantile. In practice, we have not detected any evidence of non-normality in our demographic applications. However, if the normal assumption is not justified, a bootstrap procedure could be used instead.

4.2 Forecast accuracy and order selection

Let $e_{n,h}(x) = y_{n+h}(x) - \hat{y}_{n,h}(x)$ denote the forecast error for (9). Then the **Integrated Squared Forecast Error** is defined as

$$\text{ISFE}_n(h) = \int_x e^2_{n,h}(x)\, dx.$$

To choose the order K of the model, we minimize the ISFE on a rolling holdout sample. That is, we fit the model to data up to time t and predict the next m periods to obtain $\text{ISFE}_t(h)$, $h = 1, \ldots, m$. Then we choose K to minimize $\sum_{t=N}^{n-h} \sum_{h=1}^{m} \text{ISFE}_t(h)$ where N is the minimum number of observations used to fit the model. In the applications in Section 5, we have used $N = 20$.

5 Applications

We demonstrate the methodology using two applications involving demographic data—age-specific fertility and age-specific mortality. In the first case we have $y_t(x_i) = \log(p_t(x_i))$ where $p_t(x_i)$ denotes the fertility rate for age x_i in year t. In the second case we have $y_t(x_i) = \log(m_t(x_i))$ where $m_t(x_i)$ denotes the mortality rate for age x_i in year t.

5.1 Fertility forecasting

Annual Australian fertility rates (1921–2000) for age groups {15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49} were obtained from the Australian Bureau of Statistics (Cat.No.3105.0.65.001, Table 38). These are defined as the number of live births during the calendar year, according to the age of the mother, per 1,000 of the female resident population of the same age at 30 June.

The observed data are shown as separate time series in Figure 2. We convert these to functional data by estimating a smooth curve through the observations, taking the centre of each age group as the point of interpolation. Several of these curves are shown in Figure 3. Note that we set fertility at ages 13 and 52 to be 0.005 for all years. While this is relatively arbitrary, it will be close to reality and helps stabilize the fitted curves. Of course, it makes a negligible difference to the fitted curves between ages 17 and 47.

The smooth curves $\hat{f}_t(x)$ were estimated using a weighted median smoothing B-spline, constrained to be concave (see He and Ng, 1999, for details). For these data, $\sigma_t^2(x)$ is obtained from (4) and we weight the smoothing B-splines using the inverse variances $w_t(x) = N_t(x)p_t(x)/[1000 - p_t(x)]$.

Figures 2 and 3 both reflect the changing social conditions affecting fertility. For example, there is an increase in fertility in all age groups around the end of World War II (1945), a rapid decrease in fertility during the 1970s due to the increasing use of contraceptive pills, and an increase in fertility at higher ages in more recent years caused by a delay in child-bearing while careers are established.

The order-selection procedure described in Section 4.2 led to a model with $K = 3$ basis functions. The robustness parameter was set to $\lambda = 3$ and the minimum number of observations used in fitting the models was $N = 20$.

The fitted bases $\hat{\phi}_k(x)$ and associated coefficients $\hat{\beta}_{t,k}$ are shown in Figure 4. In this case, no points were identified as outliers, and so the fitted principal components are the same as those obtained using the procedure described in Ramsay and Silverman (2005). The basis functions explain 86.8%, 9.9% and 2.5% of the variation respectively, leaving only 0.8% unexplained.

From Figure 4, it is apparent that the basis functions are modelling the fertility rates of mothers in different age ranges: $\phi_1(x)$ models late-mothers in their 40s, $\phi_2(x)$ models young mothers in their late teens and 20s, and $\phi_3(x)$ models mothers in their 30s. The coefficients associated with each basis function demonstrate the social effects noted earlier. See, in particular, the increase in the coefficients around 1945, the decrease in the coefficients during the 1970s, and the increase in $\beta_{t,1}$ and $\beta_{t,3}$ and decrease in $\beta_{t,2}$ since 1980 reflecting the

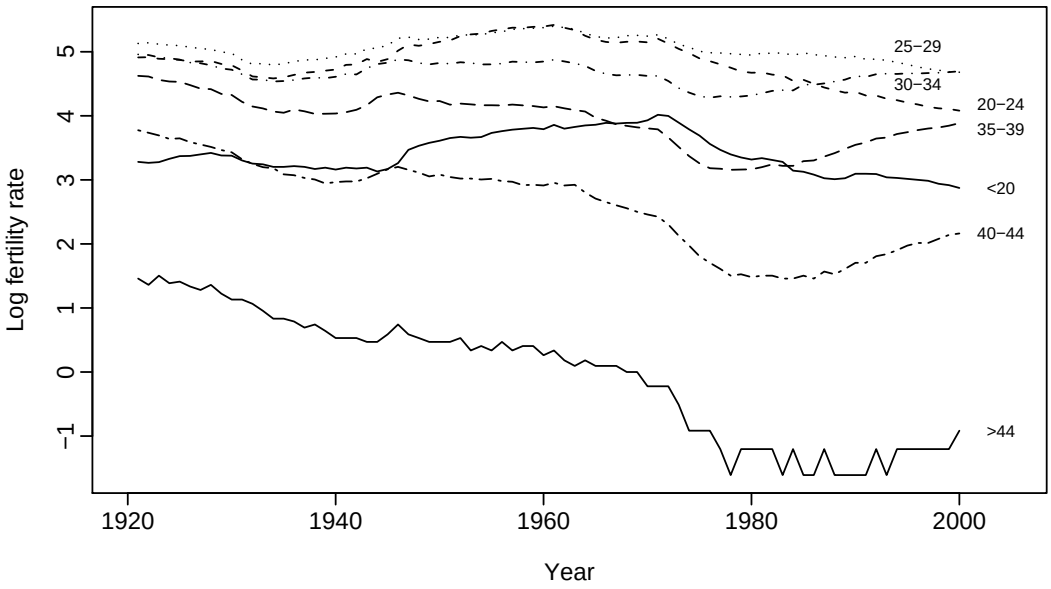

Fig. 2. *Log fertility rates viewed as time series.*

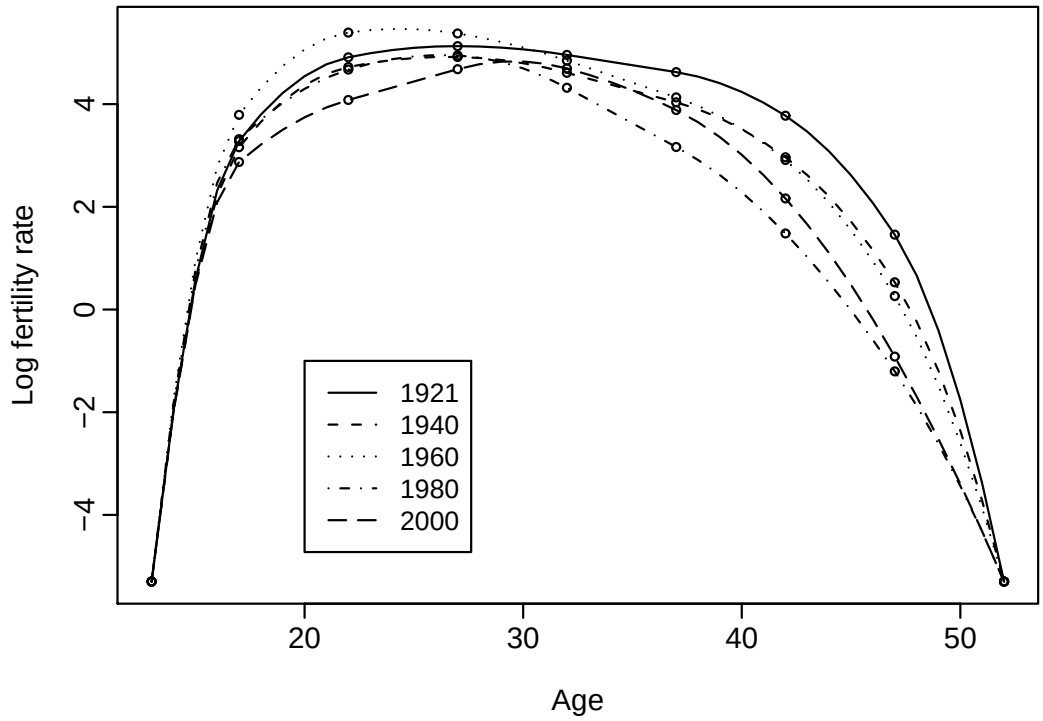

Fig. 3. *Log fertility rates viewed as functional data and calculated using median smoothing B-splines constrained to be concave.*

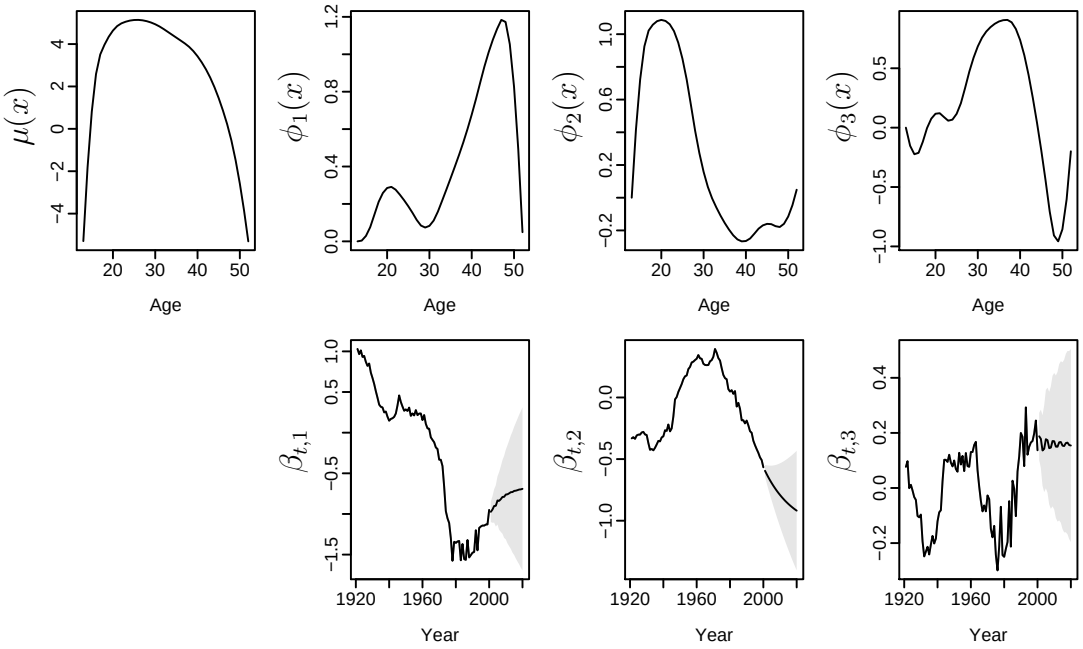

Fig. 4. *Basis functions and associated coefficients for the data shown in Figures 2 and 3. A decomposition of order $K = 3$ has been used. Forecasts of the coefficients are shown with 80% prediction intervals.*

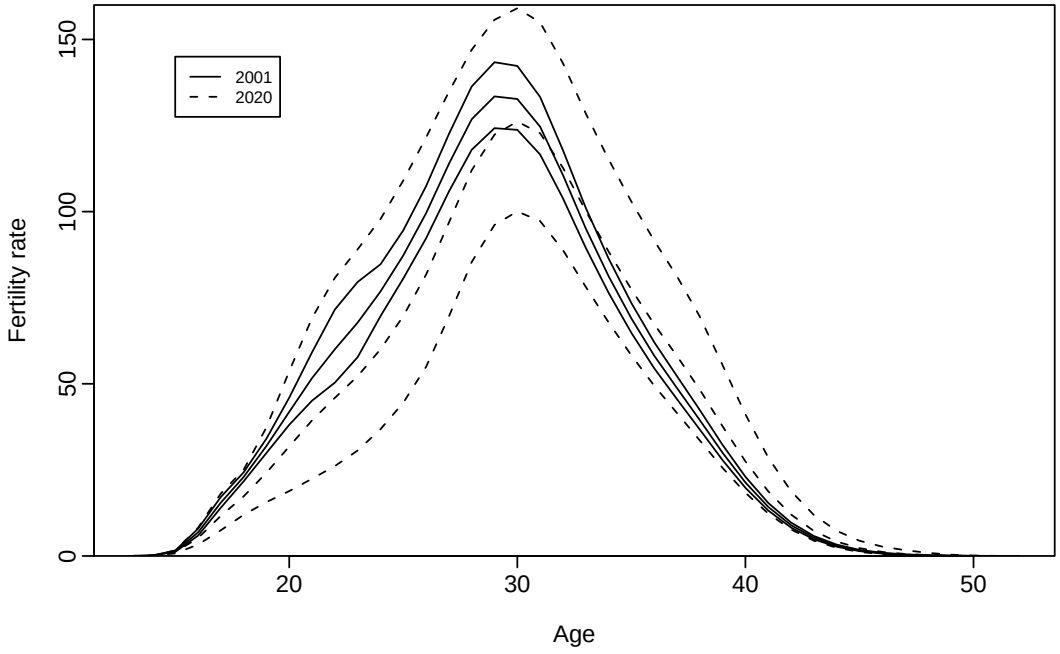

Fig. 5. *Forecasts of fertility rates for 2001 and 2020, along with 80% prediction intervals.*

shift to later ages for giving birth.

Twenty-year forecasts of the coefficients based on robust ARIMA models are shown in Figure 4. Again, no outliers were detected. The grey shaded regions are 80% prediction intervals.

Combining the forecast coefficients with the estimated basis functions yields forecasts of the fertility curves for 2001–2020. The forecasts for 2001 and 2020 are shown in Figure 5 along with 80% prediction intervals computed using the variance given by (10). Forecasts for the intervening years lie between these two years. Clearly, the greatest forecast change is a continuing decrease in fertility rates for ages 17–30. A small increase is forecast for ages 30 and over.

5.2 Mortality forecasting

Annual French mortality rates (1899–2001) for single year of age were obtained from the Human Mortality Database. These are simply the ratio of death counts to population exposure in the relevant interval of age and time. Some of the data were shown in Figure 1.

For these data, we estimate $f_t(x)$ using penalized regression splines (Wood, 2003) as they allow monotonicity constraints to be imposed relatively easily. These are useful in our mortality application where we assume that $f_t(x)$ is monotonic for $x > c$ for some c (say 50 years). This monotonicity constraint allows us to avoid some of the noise in the estimated curves for high ages, and is not unreasonable for this application (the older you are, the more likely you are to die). We use a modified version of the approach described in Wood (1994) to implement the monotonicity constraint.

The observational variance, $\sigma_t^2(x)$, is obtained from (3). The order-selection procedure described in Section 4.2 led to a model with $K = 4$ basis functions. The forecast methodology used in these computations was the same as in the Australian fertility example.

The fitted bases $\hat{\phi}_k(x)$ and associated coefficients $\hat{\beta}_{t,k}$ are shown in Figure 6. In this case, several years were identified as outliers, namely 1914–1919 and 1940–1945 and 1960. Obviously the first two periods are largely due to war deaths changing the mortality patterns for those years, although the Spanish flu also had a substantial impact in 1918. The other outlier, in 1960, appears to be due to slightly unusual mortality rates for teenagers, although the cause of this is unknown. The basis functions explain 95.6%, 2.0%, 1.1% and 0.5% of the variation respectively, leaving only 0.7% unexplained.

From Figure 6, it is apparent that the basis functions are modelling different

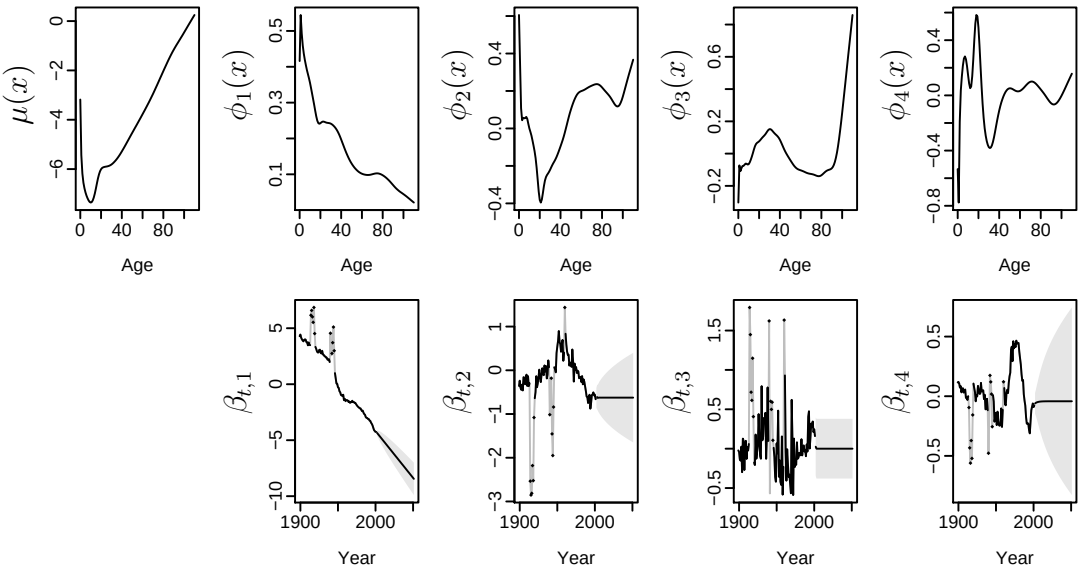

Fig. 6. *Basis functions and associated coefficients for the French male mortality data. A decomposition of order $K = 4$ has been used. Outlying years are connected with grey lines; other years are connected with black lines. Fifty-year forecasts of the coefficients are also shown. The grey shaded regions are 80% prediction intervals.*

movements in mortality rates: $\phi_1(x)$ primarily models mortality changes in children; $\phi_2(x)$ models the difference between young adults and those over 60; and $\phi_3(x)$ models the very old. The final basis function, $\phi_4(x)$ is more complex and we do not attempt to interpret it. The mortality rates for children have dropped over the whole data period, while the difference between young adults and those over 60 has only been falling since about 1960, and has levelled off in the last decade of data. This phenomenon is captured by the time series model in projecting no change to this coefficient over the next fifty years. The coefficients were forecast using robust ARIMA models with innovation outliers. The outliers in the time series coefficients coincided with the outliers identified in the functional data.

6 Connections with other approaches

There has been many recent studies on mortality forecasting, the vast majority based on the Lee-Carter (LC) method which combines a parsimonious demographic model with statistical time series analysis (Lee and Carter, 1992; Carter and Lee, 1992; Lee et al., 1995). The LC method uses raw mortality rates rather than smoothed functions, and conventional PCA rather than functional PCA. It also uses only one principal component. In practical implementations, the LC method seems to always utilize a random walk with drift

for forecasting the coefficients, although the original paper allowed for other possibilities.

The modelling framework we propose here is a generalization of the LC method. The LC method is obtained if we set $K = 1$, $\lambda = \infty$, and under-smooth the data so the fitted functions $f_t(x)$ interpolate the observations $y_t(x)$. A random walk with drift is also a special case of an ARIMA model.

Lee and Miller (2001) proposed a modification to the LC method in which the coefficient series is adjusted so the fitted life expectancy is equal to the observed life expectancy in each year. They also adjust the forecasts by the amount $y_n(x) - \hat{\mu}(x) - \hat{\beta}_{n,1}\phi_1(x)$ in an attempt to reduce bias. We conjecture that this bias is due to the model having order too low ($K = 1$) and is reduced by simply increasing K.

Booth et al. (2002) modified the LC method to adjust the coefficients using age-at-death distribution, and to determine the optimal fitting period in order to address non-linearity in the time component. For Australian data, this modified version results in greater forecast life expectancy and a fifty percent reduction in forecast error than LC method.

Our approach also bears some similarities to the work of Bozik and Bell (1987), Bell and Monsell (1991) and Bell (1997). These authors also use a principal component decomposition, but do not allow for outliers and do not use non-parametric smoothing. They suggest using a multivariate time series model for all coefficients and do not exploit the orthogonality of the coefficient series.

We compare our results with those obtained using the methods of Lee and Carter (1992), Lee and Miller (2001) and Booth et al. (2002). The methods are applied to the French mortality data for years 1899–m and we forecast years $m + 1, \dots, \min(2001, m + 20)$. The forecasts are compared with actual values and we average the MSE (on the log scale) over $m = 1959, \dots, 2000$. The results are shown in Figure 7. Our approach and the Lee-Miller approach are barely distinguishable for forecasts of 7-steps-ahead or less. For longer forecast horizons, the advantage of the additional basis functions becomes apparent. The big difference between the Lee-Miller method and the Lee-Carter method is primarily due to the bias reduction.

Booth et al. (2005) compared out-of-sample forecast performance for the LC, Lee-Miller and Booth-Maindonald-Smith methods on 20 sex-specific populations, and found that the two latter methods were both substantially better than the original LC method, but that there was little difference between these methods. The relatively poor performance of the Booth-Maindonald-Smith method in our comparison is due to the much longer evaluation period (1959–2000) than was considered in Booth et al. (2005). Forecasts early in this period were particularly poor due to the presence of substantial outliers

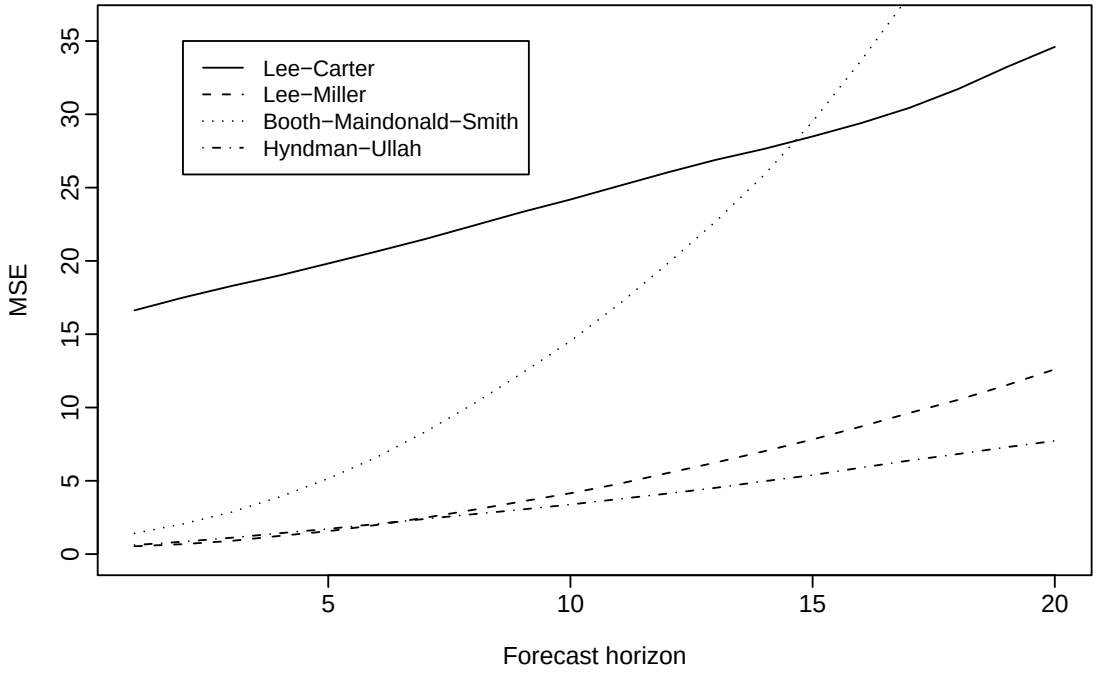

Fig. 7. *Comparative forecast accuracy of four methods showing out-of-sample MSE of log mortality. Fitting period: 1899–m where* $m = 1959, \ldots, 2000$*. Forecast period:* $m + 1, \ldots, \min(2001, m + 20)$*.*

in the fitting period.

Booth et al. (2006) extended the results of Booth et al. (2005) and compared the out-of-sample forecast performance on the same 20 sex-specific populations (but with slightly revised data) for the LC, Lee-Miller and Booth-Maindonald-Smith methods as well as the method proposed in this paper and that of De Jong and Tickle (2006). Of the five methods compared, forecasts of log-death rates based on the method in this paper were about the same as those from the De Jong-Tickle method, and both were better than the other three methods.

The advantage of our proposed method is that it is applicable to other applications as well as mortality forecasting, and it is relatively easy to generalize it allow for cohort effects, explanatory variables and multiple populations. We give the flavour of some possible extensions in the next section.

7 Extensions

In many situations, there will be multiple functional time series to be forecast, and these will have related dynamics. Consequently, it is useful to consider several extensions of our basic model.

Suppose we observe $y_{t,j}(x)$ for each of $j = 1, \dots, M$ groups. For example, j may denote sex in which case $M = 2$. Alternatively, j may denote a geographical region within a country. Then the following models are of interest:

$$y_{t,j}(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_{t,j}(x) \tag{11}$$

$$y_{t,j}(x) = \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_{t,j}(x) \tag{12}$$

$$y_{t,j}(x) = \mu_j(x) + \sum_{k=1}^{K} \beta_{t,j,k}\phi_k(x) + e_{t,j}(x) \tag{13}$$

$$y_{t,j}(x) = \mu_j(x) + \sum_{k=1}^{K} \beta_{t,j,k}\phi_{k,j}(x) + e_{t,j}(x) \tag{14}$$

These represent models of successively more degrees of freedom. The first (11) assumes no difference between the groups; (12) assumes the groups differ by an amount depending only on x and not t; (13) assumes the same basis functions apply to all groups, but the coefficients differ between groups; and (14) allows completely different models for each group.

More complicated variations are possible. For example,

$$y_{t,j}(x) = \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + \sum_{\ell=1}^{L} \gamma_{t,j,\ell}\psi_{\ell,j}(x) + e_{t,j}(x) \tag{15}$$

If j denotes sex, than a model of type (15) may be of interest where each of the coefficient series $\gamma_{t,j,\ell}$ is constrained to be a stationary process. This guarantees that the difference in mortality rates

$$y_{t,1}(x) - y_{t,2}(x) = \mu_1(x) - \mu_2(x) + \sum_{\ell=1}^{L} [\gamma_{t,1,\ell}\psi_{\ell,1}(x) - \gamma_{t,2,\ell}\psi_{\ell,2}(x)] + (e_{t,1}(x) - e_{t,2}(x))$$

is non-divergent.

We plan to develop methods to estimate and forecast these extended models in future papers. Statistical tests for the differences between these models are also of interest as they will enable the dynamics of related functional time series to be studied.

8 Conclusions

We have introduced a new approach to forecasting functional time series data. While we have focussed on its applications in demography, we have also used

the method in epidemiology (Erbas et al., 2006) and finance with equal success. It is suitable for any situation where multiple time series are observed and where the observations in each period can be considered as arising from an underlying smooth curve.

We have demonstrated the method on fertility and mortality data, and shown that it achieves better forecasting results than other approaches to mortality forecasting. This superior performance arises for several reasons: (1) we allow more complex dynamics than other methods by setting $K > 1$, thus allowing higher order terms to be included; (2) nonparametric smoothing reduces the observational noise; (3) the use of robust methods avoids problems of outlying years, especially around the world wars. It has the added advantage of providing interesting historical interpretations of dynamic changes by separating out the effects of several orthogonal components. We can also produce prediction intervals for our forecasts, taking into account all sources of variation.

In our implementation of the methodology, we have proposed a simple procedure for order selection and a simple automated procedure for forecasting the individual coefficients. This makes it very easy to apply and use. An R package which implements the methodology is available from the first author.

References

Bell, W. R., 1997. Comparing and assessing time series methods for forecasting age-specific fertility and mortality rates. Journal of Official Statistics 13 (3), 279–303.

Bell, W. R., Monsell, B., 1991. Using principal components in time series modelling and forecasting of age-specific mortality rates. In: Proceedings of the American Statistical Association, Social Statistics Section. pp. 154–159.

Booth, H., Hyndman, R. J., Tickle, L., de Jong, P., 2006. Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. Working paper, Department of Econometrics and Business Statistics, Monash University. `http://www.robhyndman.info/papers/LCcomparison.htm`.

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., 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.

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

Carter, L. R., Lee, R. D., 1992. Modelling and forecasting US sex differentials in mortality. International Journal of Forecasting 8 (3), 393–411.

Chen, C., Liu, L.-M., 1993. Joint estimation of model parameters and outlier effects in time series. Journal of the American Statistical Association 88, 284–297.

Croux, C., Ruiz-Gazen, A., 1996. A fast algorithm for robust principal components based on projection pursuit. In: Prat, A. (Ed.), COMPSTAT: Proceedings in Computational Statistics. Physica-Verlag, Heidelberg, pp. 211–216.

Croux, C., Ruiz-Gazen, A., 2005. High breakdown estimators for principal components: The projection-pursuit approach revisited. Journal of Multivariate Analysis 95 (1), 206–226.

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

Dauxois, J., Pousse, A., Romain, Y., 1982. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. Journal of Multivariate Analysis 12, 136–154.

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

Erbas, B., Hyndman, R. J., Gertig, D. M., 2006. Forecasting age-specific breast cancer mortality using functional data models. Statistics in Medicine. To appear.

Ferraty, F., Vieu, P., 2004. Nonparametric models for functional data, with application in regression, time-series prediction and curve discrimination. Nonparametric Statistics 16 (1–2), 111–125.

He, X., Ng, P., 1999. COBS: qualitatively constrained smoothing via linear

programming. Computational Statistics 14, 315–337.
Hössjer, O., Croux, C., 1995. Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. Nonparametric Statistics 4, 293–308.
Hubert, M., Rousseeuw, P. J., Verboven, S., 2002. A fast method of robust principal components with applications to chemometrics. Chemometrics and Intelligent Laboratory Systems 60, 101–111.
Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at `www.mortality.org` (data downloaded on 1 May 2006).
Jolliffe, I. T., 2002. Principal component analysis, 2nd Edition. Springer-Verlag, New York.
Lee, R. D., 1993. Modeling and forecasting the time series of U.S. fertility: age distribution, range, and ultimate level. International Journal of Forecasting 9, 187–202.
Lee, R. D., Carter, L. R., 1992. Modeling and forecasting U.S. mortality. Journal of the American Statistical Association 87, 659–675.
Lee, R. D., Carter, L. R., Tuljapurkar, S., 1995. Disaggregation in population forecasting: Do we need it? And how to do it simply? Mathematical Population Studies 5 (3), 217–234.
Lee, R. D., Miller, T., 2001. Evaluating the performance of the Lee-Carter method for forecasting mortality. Demography 38 (4), 537–549.
Li, G., Chen, Z., 1985. Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. Journal of the American Statistical Association 80 (391), 759–766.
Li, S.-H., Chan, W.-S., 2005. Outlier analysis and mortality forecasting: the United Kingdom and Scandinavian countries. Scandinavian Actuarial Journal 3, 187–211.
Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., Cohen, K. L., 1999. Robust principal component analysis for functional data. Sociedad de Estadística e Investigación Operativa Test 8 (1), 1–73.
Ramsay, J. O., Dalzell, C. J., 1991. Some tools for functional data analysis. Journal of the Royal Statistical Society, Series B 53 (3), 539–572.
Ramsay, J. O., Silverman, B. W., 2005. Functional data analysis, 2nd Edition. Springer-Verlag, New York.
Renshaw, A. E., Haberman, S., 2003. Lee-Carter mortality forecasting: a parallel generalized linear modelling approach for England and Wales mortality projections. Applied Statistics 52 (1), 119–137.
Ruppert, D., Wand, M. P., Carroll, R. J., 2003. Semiparametric regression. Cambridge University Press, New York.
Silverman, B. W., 1996. Smoothed functional principal components analysis by choice of norm. Annals of Statistics 24, 1–24.
Simonoff, J. S., 1996. Smoothing methods in statistics. Springer-Verlag, New York.
Valderrama, M. J., Ocaña, F. A., Aguilera, A. M., 2002. Forecasting PC-

ARIMA models for functional data. In: Härdle, W., Rönz, B. (Eds.), Proceedings in Computational Statistics. pp. 25–36.

Wilmoth, J. R., 2002. Methods Protocol for the Human Mortality Database. Revised 1 October 2002. Downloaded on 18 July 2003. http://www.mortality.org/Public/Docs/MethodsProtocol.pdf.

Wolf, D. A., 2004. Another variation on the Lee-Carter model. Paper presented at the annual meeting of the Population Association of America, April 2004.

Wood, S. N., 1994. Monotonic smoothing splines fitted by cross validation. SIAM J. Sci Comput 15 (5), 1126–1133.

Wood, S. N., 2003. Thin plate regression splines. Journal of the Royal Statistical Society, Series B 65 (1), 95–114.