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

Rob J Hyndman and Md Shahid Ullah
*Monash University, Business & Economic Forecasting Unit*
*Clayton VIC 3135, Australia*
*Rob.Hyndman@buseco.monash.edu.au*

## Notation and background

We propose a new method for robust forecasting of age-specific mortality and fertility rates. To illustrate our methodology, we use annual Australian fertility rates (1921–2000) for five-year age groups (15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49). The data were obtained from the Australian Bureau of Statistics and are shown as separate time series in Figure 1 (left). We convert these to functional data by estimating a smooth curve through the observations for each year, taking the centre of each age group as the point of interpolation. Several of these curves are shown in Figure 1 (right). Note that we set fertility at ages 13 and 52 to be 0.001 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. Our aim is to forecast future curves.

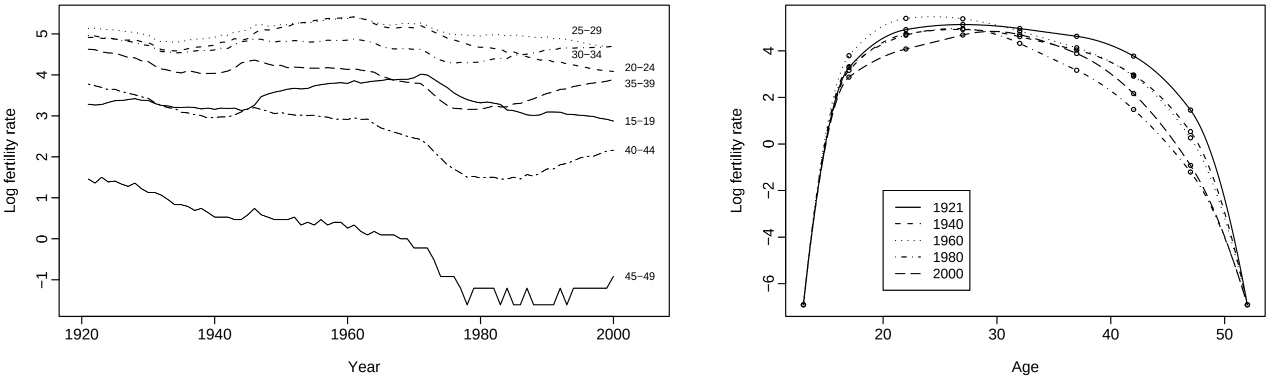

**Figure 1:** *Left: Log fertility rates per thousand women (Australia, 1921–2000) viewed as time series. Right: Log fertility rates viewed as functional data and calculated using median smoothing B-splines constrained to be concave.*

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 mortality function $f_t(x)$ that we are observing with error. Thus, we observe the functional time series $\{x_i, y_t(x_i)\}$, $t = 1, \dots, n$, $i = 1, \dots, p$ where

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

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

We can compute the observational variance, $\sigma_t^2(x)$, as follows. 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) $\sigma_t^2(x) \approx [1 - m_t(x)]N_t^{-1}(x)m_t^{-1}(x)$. Similarly, let $p_t(x)$ denote the observed fertility rate per thousand women for females of age $x$ in year $t$ and $M_t(x)$ be the female resident population of age $x$ at 30 June in year $t$. Then, for fertility data, $\sigma_t^2(x) \approx [1000 - p_t(x)]M_t^{-1}(x)p_t^{-1}(x)$.

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 & Carter (1992); they proposed a methodology for modelling and extrapolating long-term trends in mortality rates. The methodology has since become very widely used and there have been various extensions and modifications proposed (e.g., Booth, Maindonald & Smith, 2002; and Renshaw & Haberman, 2003). The methodology has also been applied to fertility (Lee, 1993). 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 & Carter (1992) in that it also involves a principal component decomposition of the mortality or fertility rates. However, we differ in several important respects. First, we use more than one principal component. Second, we use the functional data paradigm (Ramsay & Silverman, 1997) 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. Third, we propose a robust version of principal components to avoid difficulties with outlying years (which often occur in mortality data due to wars and epidemics).

### Modelling

We smooth the data for each $t$ using a nonparametric smoothing method to estimate $f_t(x)$ for $x \in (x_1, x_p)$ from $\{x_i, y_t(x_i)\}$, $i = 1, 2, \dots, p$. We use constrained and weighted penalized regression splines for mortality and fertility data. The weights are $\sigma_t^{-2}(x)$ and we apply either a monotonic constraint (for mortality rates at upper ages) or a concavity constraint (for fertility rates).

Next we 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 $e_t(x) \sim \mathrm{N}(0, v(x))$. We use the $L_1$-median (Hössjer & Croux, 1995) to estimate $\mu(x)$. The basis functions are obtained using a robust method for functional principal components based on ideas from Ramsay & Dalzell (1991) and Hubert, Rousseeuw & Verboven (2001).

### Forecasting

We need to forecast $\beta_{t,k}$ for $k = 1, \dots, K$ and $t = n+1, \dots, n+h$. 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 $k \neq \ell$. There may still be cross-correlations at non-zero lags, but these are likely to be small given the zero contemporaneous correlations. Therefore, univariate methods will be adequate for forecasting each series $\{\hat\beta_{t,k}\}$.

Now combining (1) with (2) we obtain

$$y_t(x_i) = \mu(x_i) + \sum_{k=1}^{K} \beta_{t,k}\,\phi_k(x_i) + e_t(x_i) + \sigma_t(x_i)\varepsilon_{t,i}\,. \tag{3}$$

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

$$\hat y_{n,h}(x) = \mathrm{E}[y_{n+h}(x) \mid \mathcal{I}, \mathbf{\Phi}] = \hat\mu(x) + \sum_{k=1}^{K} \tilde\beta_{n,k,h}\,\hat\phi_k(x)\,, \tag{4}$$

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 (3):

$$\zeta_{n,h}(x) = \text{Var}[y_{n+h}(x) \mid \mathcal{I}, \mathbf{\Phi}] = \hat{\sigma}^2_\mu(x) + \sum_{k=1}^{K} u_{n+h,k}\, \hat{\phi}^2_k(x) + v(x) + \sigma^2_t(x) \tag{5}$$

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 model error variance $v(x)$ is estimated by averaging $\hat{e}^2_t(x)$ for each $x$.

Because of the way the model has been constructed, each component is orthogonal to the other components and so the forecast variance is a simple sum of component variances. A prediction interval for $y_{n+h}(x)$ is constructed as $\hat{y}_{n,h}(x) \pm 1.96\sqrt{\zeta_{n,h}(x)}$ assuming the various sources of error are all normally distributed.

Let $e_{n,h}(x) = y_{n+h}(x) - \hat{y}_{n,h}(x)$ denote the forecast error and define the Integrated Squared Forecast Error as $\text{ISFE}_n(h) = \int_x e^2_{n,h}(x)\, dx$. We choose the order $K$ of the model, by minimizing $\displaystyle\sum_{t=N}^{n-h}\sum_{h=1}^{m} \text{ISFE}_t(h)$ where $N$ is the minimum number of years used to fit the model.

## Application: Australian fertility forecasting

The Australian fertility data shown in Figure 1 reveal 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 method with $N = 20$ gives a model with $K = 3$ basis functions. The forecast methodology used in these computations was the single source of error state space model (Hyndman, et al., 2002) which underlies the damped Holt's method. This was selected as it extrapolates the local trends seen in the coefficient series while damping them to avoid nonsensical long-term forecasts.

The fitted bases $\hat{\phi}_k(x)$ and associated coefficients $\hat{\beta}_{t,k}$ are shown in Figure 2 (left). The basis functions explain 86.3%, 10.3% and 2.6% of the variation respectively, leaving only 0.8% unexplained. From Figure 2, it is apparent that the basis functions are modelling the fertility rates of females 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 above. 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 shift to later ages for giving birth.

Twenty-year forecasts of the coefficients are shown in Figure 2 (left). The grey shaded regions are 80% prediction intervals computed from the underlying state space model.

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 2 (right) along with 80% prediction intervals computed using the variance given by (5). 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.

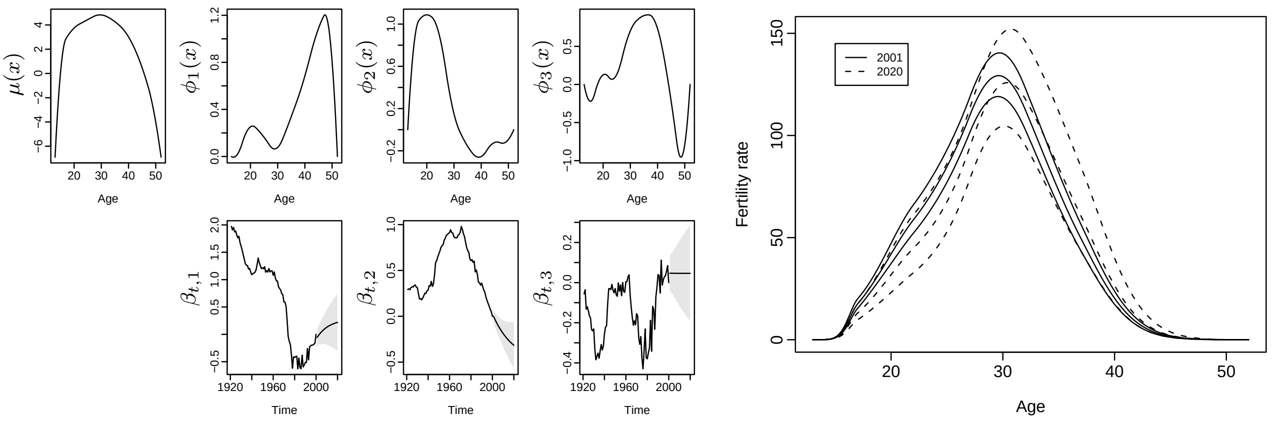

**Figure 2:** *Left: basis functions and associated coefficients for the data shown in Figure 1. A model of order $K = 3$ has been used. Forecasts of the coefficients are shown with 80% prediction intervals. Right: forecasts of fertility rates for 2001 and 2020, along with 80% prediction intervals.*

## REFERENCES

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

Hössjer, O., and Croux, C. (1995) Generalized univariate signed rank statistics for testing and estimating a multivariate location parameter. *Nonparametric Statistics*, **4**, 293–308.

Hubert, M., Rousseeuw, P.J., and Verboven, S. (2002), A fast robust method for principal components with applications to chemometrics, *Chemometrics and Intelligent Laboratory Systems*, **60**, 101–111.

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

Lee, R. (1993) Modeling and forecasting the time series of US fertility: age patterns, range, and ultimate Level. *International Journal of Forecasting*, **9**, 187–202.

Lee, R., and Carter, L. (1992) Modelling and forecasting the time series of US mortality. *Journal of the American Statistical Association*, **87**, 659–671.

Ramsay, J.O., and Dalzell, C.J. (1991) Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society*, Series B, **53**(3), 539–572.

Ramsay, J.O., and Silverman, B.W. (1997) *Functional data analysis.* Springer-Verlag: New York

Renshaw, A., and 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.

## RÉSUMÉ

*We propose a new method for forecasting age-specific mortality and fertility rates observed over time. We combine ideas from functional data analysis, nonparametric smoothing and robust statistics to form a methodology that is widely applicable to any functional time series data, and age-specific mortality and fertility in particular. Our approach provides a modelling framework that is easily adapted to allow for constraints and other information. The model used can be considered a generalization of the Lee-Carter model commonly used in mortality and fertility forecasting. The methodology is applied to Australian fertility data.*