Prediction of Hungarian mortality rates using Lee-Carter method*

S. Baran†, J. Gáll, M. Ispány and G. Pap
Faculty of Informatics, University of Debrecen
P. O. Box 12, H-4010 Debrecen, Hungary

Abstract

A modified of version the popular Lee-Carter method (Lee and Carter, 1992) is applied for prediction of the mortality rates in Hungary for the period 2004 – 2040 on the basis of mortality data of men and women of years 1949 – 2003 and another case is also considered based on a restricted data set corresponding to the period 1989 – 2003. The model fitted on data of period 1949 – 2003 predicts increasing mortality rates for men of ages 45–55 which shows that the Lee-Carter method is hardly applicable for the countries where the mortality rates are so changing as in Hungary. However, the models fitted to the data of the last 15 years both for men and women forecast decreasing trends similarly to case of countries where the method was successfully applied. Hence one gets a better fit in this way, however, further concerns suggests that the Lee-Carter model does not necessarily give sufficiently good prediction.

1. Introduction

In the last century the life expectancy in Hungary increased from 37.3 years (1900) to 71.3 years (2000) and this trend seems to continue. The prediction of life expectancy and mortality became a key problem in actuarial- and social sciences as well. In order to forecast mortality it must be accurately modeled and in the past twenty years several models has been developed (see e.g. Alho (1990), Alho and Spencer (1985, 1990), McNown and Rogers (1989) or Tabeau (2001)).

Naturally, the mortality rates of a country depend on the level of the development of the given country and also on medical, social etc. influences. However, Lee and Carter (1992) introduced a new method which is based on time series analysis and does not incorporate knowledge about the previously mentioned influences. With the help of this method Lee and Carter (1992) modeled and predicted US mortality. Later Lee (2000) gave various extensions of the method and applied them for date from US and Chile. Booth *et al.* (2002) applied the Lee-Carter method to Australian data, while Renshaw and Haberman (2003) compared the forecasts based on Lee-Carter with the forecasts based on generalized linear and bilinear models with Poisson error structures.

*Research is supported by the ING Insurance Co. Ltd. Hungary, and partially supported by the Hungarian Scientific Research Fund under Grants No. OTKA-F046061/2004 and OTKA-T048544/2005.

†Corresponding author. E-mail: barans@inf.unideb.hu

The aim of our research is to estimate the Hungarian mortality rates for the period 2004 – 2040 using the Lee-Carter method on the basis of mortality data between 1949 and 2003 for ages from 0 to 100 years both for men and women. The source of the data is the Demographic Yearbook of the Hungarian Central Statistical Office.

2. The Lee-Carter model

Let $m_{x,t}$ denote the central death rate for age x in year t, $x = 1, \dots, N,\ t = 1, \dots, T$. The model for the mortality is

$$\ln(m_{x,t}) = a_x + b_x \cdot k_t + \varepsilon_{x,t}, \tag{2.1}$$

where a_x and b_x are parameters depending only on age, k_t is a stochastic process depending only on the year of observation and $\varepsilon_{x,t}$ are independent error terms with mean 0 and variance σ_ε^2. The value k_t can be interpreted as the index of mortality at year t, while $\exp(a_x)$ is the general pattern of mortality by age. Since if we forget about the error term $\varepsilon_{x,t}$ we have

$$\frac{\mathrm{d}\ln(m_{x,t})}{\mathrm{d}t} = b_x \frac{\mathrm{d}k_t}{\mathrm{d}t},$$

b_x can be considered as the sensitivity of the mortality rate to the change of the mortality index.

In order to use model (2.1) for prediction, first we have to fit it, i.e. to estimate its parameters $\{a_x : x = 1, \dots, N\}$, $\{b_x : x = 1, \dots, N\}$ and $\{k_t : t = 1, \dots, T\}$. The least squares method, which means the minimization of

$$\sum_{x=1}^{N} \sum_{t=1}^{T} (\ln(m_{x,t}) - a_x - b_x \cdot k_t)^2$$

yields the following system of equations

$$T a_x + b_x \sum_{t=1}^{T} k_t - \sum_{t=1}^{T} \ln(m_{x,t}) = 0, \qquad x = 1, \dots, N, \tag{2.2}$$

$$a_x \sum_{t=1}^{T} k_t + b_x \sum_{t=1}^{T} k_t^2 - \sum_{t=1}^{T} k_t \ln(m_{x,t}) = 0, \qquad x = 1, \dots, N, \tag{2.3}$$

$$\sum_{x=1}^{N} a_x b_x + k_t \sum_{x=1}^{N} b_x^2 - \sum_{x=1}^{N} b_x \ln(m_{x,t}) = 0, \qquad t = 1, \dots, T. \tag{2.4}$$

As for all $c \in \mathbb{R}$

$$a_x + b_x \cdot k_t = (a_x - c b_x) + b_x \cdot (k_t + c)$$

and for all $0 \neq c \in \mathbb{R}$

$$a_x + b_x \cdot k_t = a_x + (c b_x) \cdot (k_t / c),$$

the solution of system (2.2)–(2.4) is not unique and one can find a solution of this system satisfying boundary conditions

$$\sum_{x=1}^{N} b_x = 1 \qquad \text{and} \qquad \sum_{t=1}^{T} k_t = 0. \tag{2.5}$$

Equations (2.2) and (2.5) imply that the estimators of the parameters $\{a_x : x = 1, \dots, N\}$ are

$$\widehat{a}_x = \frac{1}{T}\sum_{t=1}^{T} \ln(m_{x,t}), \qquad x = 1, \dots, N.$$

Now, instead of solving the system (2.2)–(2.4) the least squares estimators of the remaining parameters can be obtained from the singular value decomposition (SVD) of the $N \times T$ matrix M with entries

$$M_{x,t} := \ln(m_{x,t}) - \widehat{a}_x, \qquad x = 1, \dots, N, \quad t = 1, \dots, T. \tag{2.6}$$

If the SVD of M is $U \cdot D \cdot V^\top$ then

$$\widehat{b}_x = \frac{1}{c} U_{x,1} \qquad \text{and} \qquad \widehat{k}_t = c D_{1,1} V_{1,t}$$

is a solution, where $D_{1,1}$ is the largest singular value of M, $U_{x,1}$ is the value at entry $(x, 1)$ of U, $V_{1,t}$ is the value at entry $(1, t)$ of V, and $c := \sum_{x=1}^{N} U_{x,1}$ which ensures

$$\sum_{x=1}^{N} \widehat{b}_x = 1.$$

The estimates $\{\widehat{k}_t : t = 1, \dots, T\}$ can be considered as a sample from a time series, and the classical Box-Jenkins method can be used to fit an appropriate time series model. For instance, Lee and Carter (1992) used an ARIMA(0, 1,0) model (random walk with a drift) to describe k_t, i.e.,

$$k_t = k_{t-1} + c + \varepsilon_t, \quad t = 1, \dots, T. \tag{2.7}$$

Here the least squares estimator $\widehat{c}$ of c equals

$$\widehat{c} = \frac{1}{T-1}(k_T - k_1).$$

Using the time series model fitted on $\{\widehat{k}_t : t = 1, \dots, T\}$ one can calculate the predictions $\{\widehat{k}_{T+t} : t = 1, \dots T'\}$ of k_{T+t} for the time points $T+1, \dots, T+T'$. For instance, in case of model (2.7) we have $\widehat{k}_{T+t} := \widehat{k}_{T+t-1} + \widehat{c}$, hence $\widehat{k}_{T+t} := \widehat{k}_T + \widehat{c}t$, $t = 1, \dots T'$. The forecast of the mortality rate is

$$\widehat{m}_{x,T+t} = \exp\left(\widehat{a}_x + \widehat{b}_x \cdot \widehat{k}_{T+t}\right), \qquad t = 1, \dots, T'.$$

As a generalization of (2.1) Booth *et al.* (2002) used a higher order approximation, namely

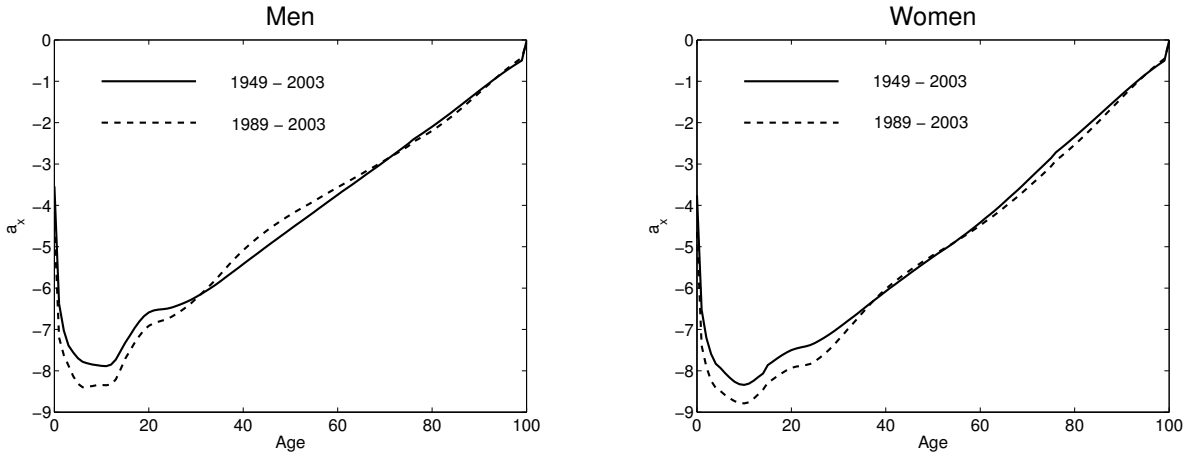
$$\ln(m_{x,t}) = a_x + b_x^{(1)} \cdot k_t^{(1)} + \cdots + b_x^{(\ell)} \cdot k_t^{(\ell)} + \varepsilon_{x,t}, \qquad x = 1, \dots N, \quad t = 1, \dots, T, \tag{2.8}$$

where ℓ is the rank of the approximation which can vary between 1 and T. To estimate the parameters one can again consider the SVD of the matrix M defined by (2.6). The estimates are

$$\widehat{b}_x^{(i)} = \frac{1}{c\,i} U_{x,i}, \qquad \text{and} \qquad \widehat{k}_t^{(i)} = c_i D_{i,i} V_{i,t},$$

where $D_{i,i}$ is the ith singular value, $i = 1, \dots, T$, while $U_{x,i}$ and $V_{i,t}$ are appropriate entries of matrices U and V, respectively. Again, for each value of i, the norming constant c_i should equal $\sum_{x=1}^{N} U_{x,i}$ to ensure

$$\sum_{x=1}^{N} b_x^{(i)} = 1,$$



a. b.

Figure 1: General pattern of mortality.

and $\widehat{k}_t^{(i)},\ t = 1, \ldots, T$, can be considered as a sample from a time series.

The rank of the approximation should be chosen to keep the ratio

$$\psi_\ell = \big(D_{1,1} + \cdots + D_{\ell,\ell}\big) / \big(D_{1,1} + \cdots + D_{T,T}\big)$$

high, and at the same time each $\{k_t^{(i)} : t = 1, \ldots, T\},\ i > \ell$, should be close to a white noise.

3. The fitted models

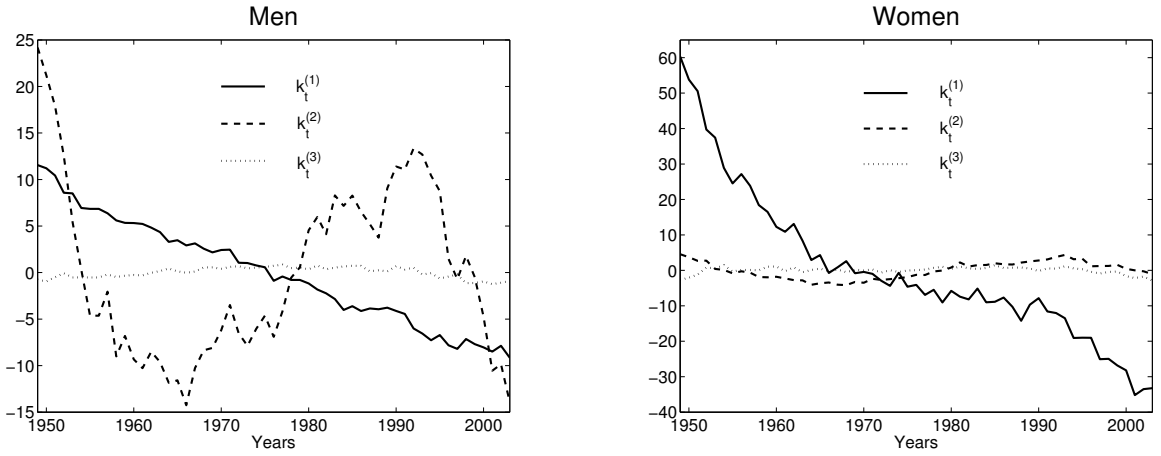
We investigated two time periods for fitting the models. The first is the whole observed period (1949 – 2003, long data set), while the second contains only the last 15 years (1989 – 2003, short data set). The idea behind the examination of a shorter time period is that the changes in the mortality rates between 1949 and 2003 due to historical reasons are extremely high, while the rates after 1989 for all ages show a constant improvement. We remark that in period 1949 – 1998 mortality rates at age 100 given by the Hungarian Central Statistical Office are considered to be 1.

Figures 1.a and 1.b show the general patterns of mortality both for men and women for the two observed time periods. We remark that these figures are rather similar to the ones obtained by Lee (2000) for US and Chile and Booth *et al.* (2002) for Australia.

Consider first the long data set. For men we have $\psi_3 = 0.6076$ while for women this ratio equals 0.5861. Hence in both cases a third order model

$$\ln(m_{x,t}) = a_x + b_x^{(1)} \cdot k_t^{(1)} + b_x^{(2)} \cdot k_t^{(2)} + b_x^{(3)} \cdot k_t^{(3)} + \varepsilon_{x,t} \tag{3.9}$$

is fitted to the data. Figures 2.a and 2.b show the samples from time series $k_t^{(i)},\ i = 1, 2, 3$, calculated from the SVD of the 101×55 matrix M using the mortality rates of men and women, respectively.



a. b.

Figure 2: The estimated time series, 1949 – 2003.

In case of men the fitted ARIMA time series models are

$$\begin{aligned} k_t^{(1)} &= -0.3834 + k_{t-1}^{(1)} + \delta_t^{(1)}, & \sigma_\delta^{(1)} &= 0.592066, \\ k_t^{(2)} &= k_{t-1}^{(2)} + \delta_t^{(2)}, & \sigma_\delta^{(2)} &= 3.114578, \\ k_t^{(3)} &= 0.8825 \cdot k_{t-1}^{(3)} + \delta_t^{(3)}, & \sigma_\delta^{(3)} &= 0.295489, \end{aligned} \tag{3.10}$$

where the innovations $\delta_t^{(i)},\ i = 1, 2, 3$, are uncorrelated random variables with mean 0 and standard deviation $\sigma_\delta^{(i)}$.

In case of women the fitted time series models are of the same type, they differ only in the values of the constants, that is

$$\begin{aligned} k_t^{(1)} &= -1.7224 + k_{t-1}^{(1)} + \delta_t^{(1)}, & \sigma_\delta^{(1)} &= 3.260100, \\ k_t^{(2)} &= k_{t-1}^{(2)} + \delta_t^{(2)}, & \sigma_\delta^{(2)} &= 0.758579, \\ k_t^{(3)} &= 0.7918 \cdot k_{t-1}^{(3)} + \delta_t^{(3)}, & \sigma_\delta^{(3)} &= 0.709897. \end{aligned} \tag{3.11}$$

In case of the short data set for men we have $\psi_3 = 0.5700$ while for women it equals 0.5283. Figures 3.a and 3.b show the samples from $k_t^{(i)},\ i = 1, 2, 3$, estimated from the model (3.9) using the mortality data of men and women, respectively. The investigations of the autocovariance structures of these time series show that in case of men $k_t^{(3)}$, while in case of women both $k_t^{(2)}$ and $k_t^{(3)}$ might be considered as white noises. Hence, for men the appropriate model for the mortality rates is

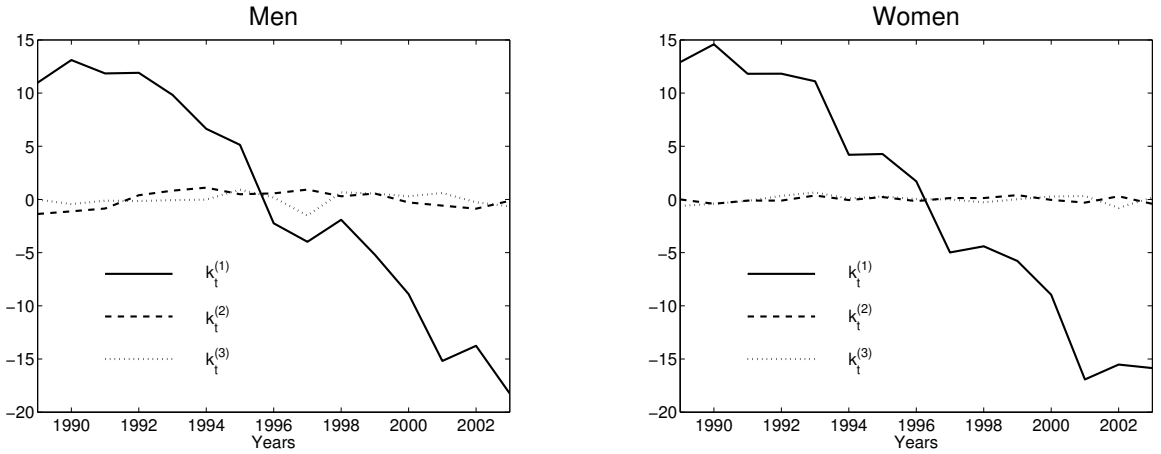
$$\ln(m_{x,t}) = a_x + b_x^{(1)} \cdot k_t^{(1)} + b_x^{(2)} \cdot k_t^{(2)} + \varepsilon_{x,t}, \tag{3.12}$$

where

$$\begin{aligned} k_t^{(1)} &= -2.0882 + k_{t-1}^{(1)} + \delta_t^{(1)}, & \sigma_\delta^{(1)} &= 2.906170, \\ k_t^{(2)} &= 0.7526 \cdot k_{t-1}^{(2)} + \delta_t^{(2)}, & \sigma_\delta^{(2)} &= 0.547201. \end{aligned} \tag{3.13}$$

For women the classical first order Lee-Carter model (2.1) is the best one, i.e.

$$\ln(m_{x,t}) = a_x + b_x^{(1)} \cdot k_t^{(1)} + \varepsilon_{x,t}, \tag{3.14}$$



a. b.

Figure 3: The estimated time series, 1989 – 2003.

where similarly to Lee and Carter (1992)

$$k_t^{(1)} = -2.0544 + k_{t-1}^{(1)} + \delta_t^{(1)}, \qquad \sigma_\delta^{(1)} = 3.148897. \tag{3.15}$$

4. Predictions

On the basis of the mortality rates of period 1949 – 2003 with the help of the fitted time series models (3.10) for men and (3.11) for women, three different predictions for the time interval 2004 – 2040 were calculated using the first (1. model), second (2. model) and third (3. model) order Lee-Carter models (3.14), (3.12) and (3.9), respectively. Figures 4.a, 4.c and 4.e show both the observed and the predicted mortality rates for men of ages 25, 50 and 75, while Figures 4.b, 4.d and 4.f show the same for women of the same ages, respectively. Naturally, the higher the order of the model, the better the prediction is. On Figure 4.c one can see that for 50 years old men the mortality rates are increasing which is rather surprising and fairly unrealistic as Hungary now is a well developing country. However, as the data show, the decrease of mortality rates of the last ten years can not compensate the huge increase between 1963 and 1993. This fact justifies the use of predictions based on the data of the last 15 years. For this we also note that the Hungarian demographic history of the last five decades involves strange trends and (regime) changes (compared to similarly developed countries), which are still not fully understood and hence subject to further scientific research for experts. Our choice of the length of the short period (15 years) is based on the above considerations. However, we must emphasize that such a short period does not give sufficiently good estimates (e.g. small standard errors of the estimates) for a long term prediction for actuarial use. Our predictions are made in a way that they are all comparable and that is the reason why we always give a relatively long term prediction compared to the size of the data. The continual re-estimation of the model parameters in the future might definitely give better results and therefore it is fairly suggested for actuarial applications.

In case of the short data set the predictions for men were calculated with the help of the fitted time series models (3.13) using the first and second order Lee-Carter models (3.14) and (3.12), respectively. On Figures 5.a, 5.c and 5.e both the observed and the predicted

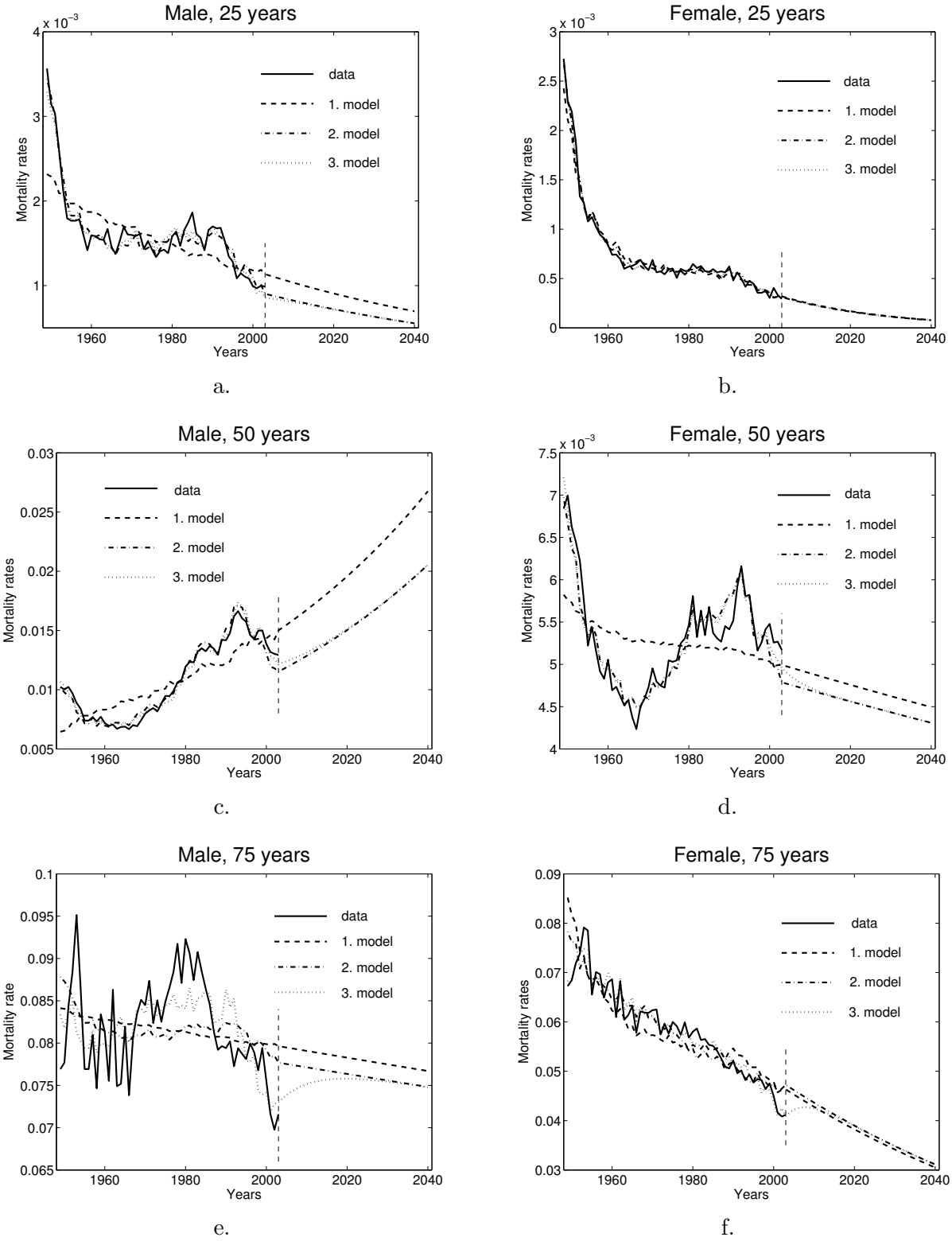


Figure 4: Predicted mortality rates for different ages. Data from period 1949 – 2003.

mortality rates are plotted for men of ages 25, 50 and 75, respectively. On Figures 4.c and 5.c one can clearly see the difference between the predictions using the long and the short dataset.

For women we could only predict using the first order Lee-Carter model (3.14) for the

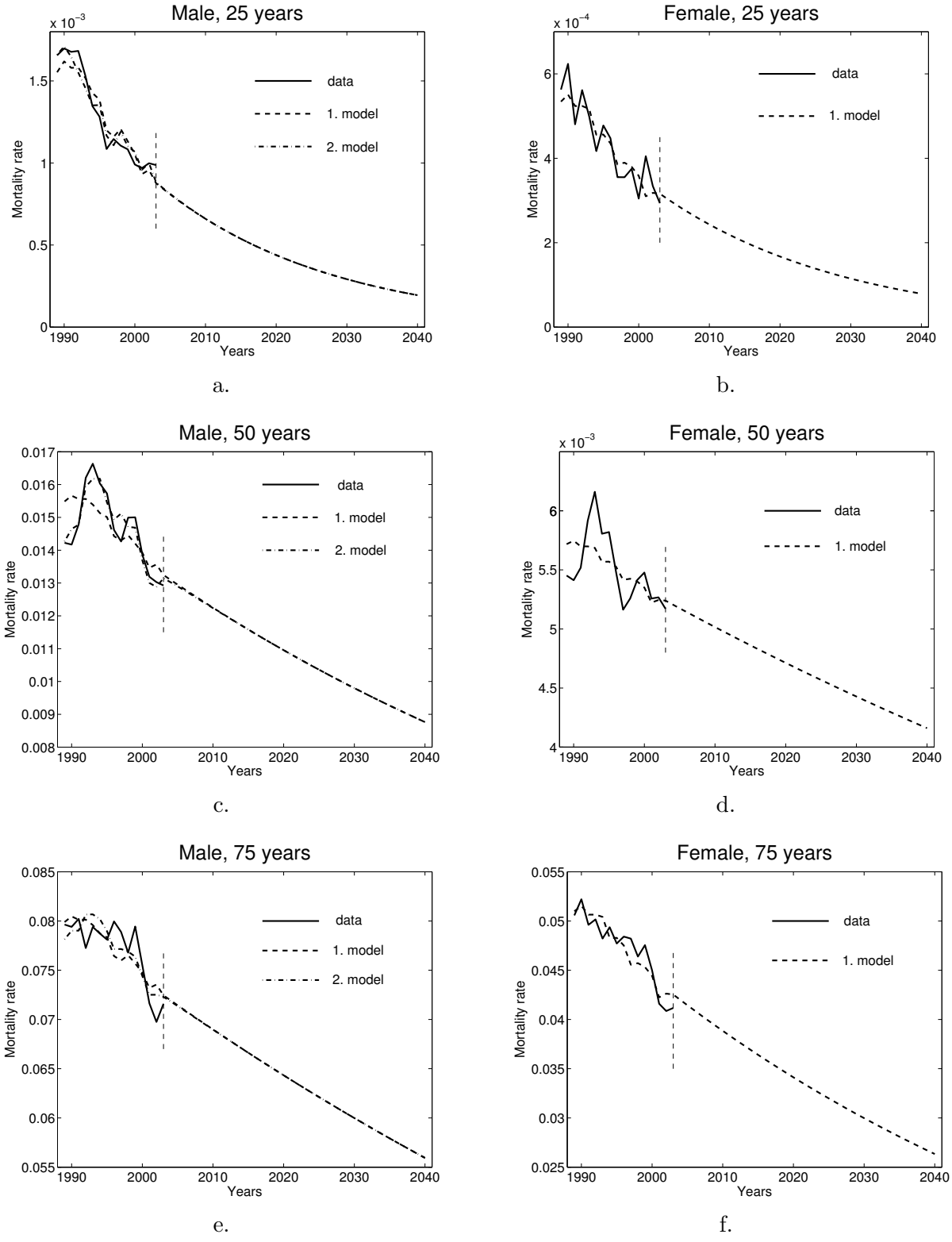
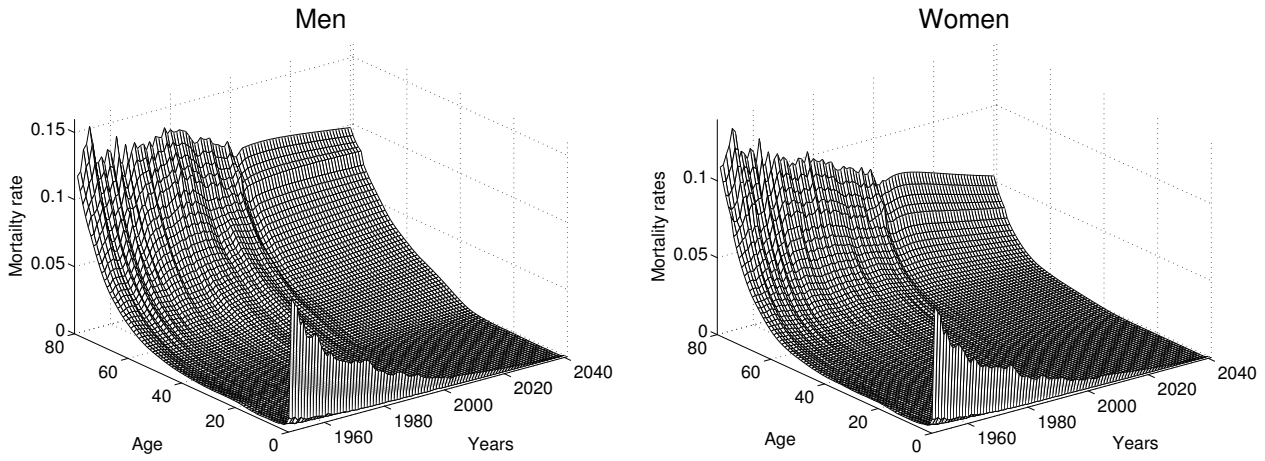


Figure 5: Predicted mortality rates for different ages. Data from period 1989 – 2003.

log mortality and the time series model (3.15). For the three ages considered before the mortality data and the predictions for the period 2004 – 2040 are indicated on Figures 5.b, 5.d and 5.f.

At the end, Figures 6.a and 6.b show the mortality rates of period 1949 – 2004 for ages



a. b.

Figure 6: Data from period 1949 – 2003 and predictions using the best model.

from 0 to 80 together with predictions for 2004 – 2040 based on the best fitted models (third order model for both sexes) for men and women, respectively.

Similarly, on Figures 7.a and 7.b the mortality rates of period 1989 – 2004 are plotted for ages from 0 to 80 together with predictions for 2004 – 2040 based on the best fitted model (second order for men, first order for women) for men and women, respectively. One can conclude, that in this case the mortality rates for all ages both for men and woman show a decreasing trend which is rather similar to the data from the U.S. (Lee and Carter, 1992) and Australia (Booth *et al.*, 2002) where the Lee-Carter method was successfully applied.

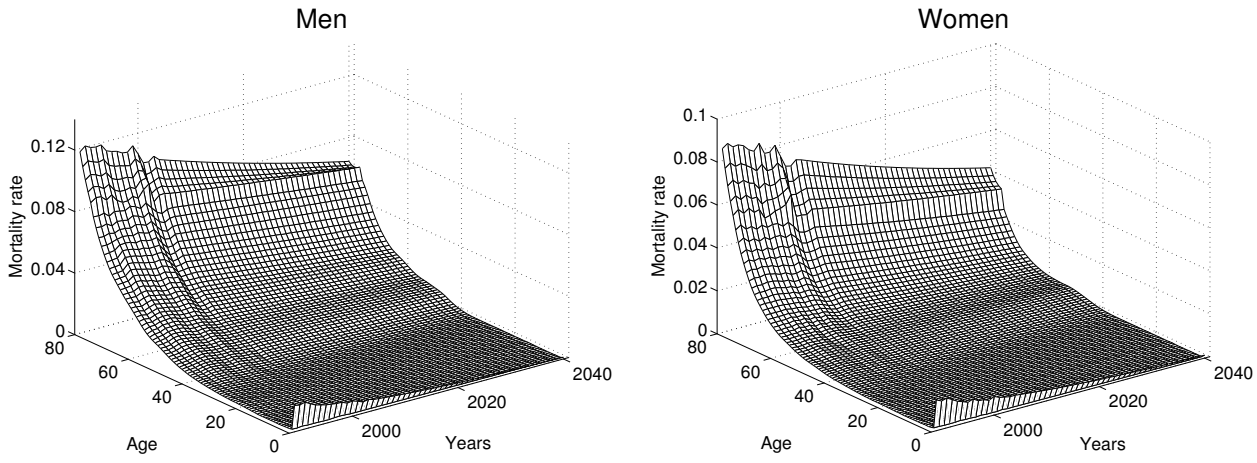
We note, however, that the decreasing trend of mortality is certainly not the only property that should be satisfied by the prediction of a good model. The Lee-Carter method is based on time series analysis and does not take into consideration further variables (covariates) to explain the mortality in a better way. Hence, though our results concerning the short data sets look similar to that of other countries, we do not suggest that Lee-Carter method gives a sufficiently good prediction e.g. for actuarial applications. Having a look at the standard errors of some predictions and also at the size of the confidence intervals derived these concerns are confirmed. The estimation of standard errors, calculation of confidence intervals and, by the aid of them, the analysis of optimistic and pessimistic scenarios for the evolution of mortality were all subjects of our research. However, these results are beyond the scope of the present paper.

References

Alho, J. M. (1990). Stochastic methods in population forecasting. *International Journal of Forecasting* **6**, no. 4, 521–530.

Alho, J. M. and Spencer, B. D. (1985). Uncertain population forecasting. *Journal of the American Statistical Association* **80**, no. 390, 306–314.

Alho, J. M. and Spencer, B. D. (1990). Error models for official mortality forecasts. *Journal of the American Statistical Association* **85**, no. 410, 609–616.



a. b.

Figure 7: Data from period 1989 – 2003 and predictions using the best model.

Booth, H., Maindonald, J. and Smith, L. (2002). Age-time interactions in mortality projection: applying Lee-Carter to Australia. *Working Paper*, ANU.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting the time series of U.S. mortality. *Journal of the American Statistical Association* **87**, no. 419, 659–671.

Lee, R. D. (2000). The Lee-Carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal* **4**, no. 1, 80–91.

McNown, R. and Rogers, A. (1989). Forecasting mortality: a parametrized time series approach. *Demography* **26**, no. 3, 433–458.

Renshaw, A. E. and Haberman, S. (2003). Lee-Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics* **33**, no. 2, 255–272.

Tabeau, E. (2001). A review of demographic forecasting models for mortality. In Tabeau, E., van den Berg Jets, A. and Heathcote, C. (eds) *Forecasting Mortality in Developed Countries: Insights from a Statistical, Demographic and Epidemiological Perspective.* Kluwer, Dordrect.