



Mathematical Statistics
Stockholm University

**Fitting and Forecasting Mortality for
Sweden: Applying the Lee-Carter Model**

Jenny Zheng Wang

Examensarbete 2007:1

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

<http://www.math.su.se/matstat>



Mathematical Statistics
Stockholm University
Examensarbete 2007:1,
<http://www.math.su.se/matstat>

Fitting and Forecasting Mortality for Sweden: Applying the Lee-Carter Model

Jenny Zheng Wang*

March 2007

Abstract

The Lee-Carter Model is one of the most popular methodologies for forecasting mortality rates. The model is widely known to be simple and has been used very successfully in U.S. and several countries. In the present paper, the Lee-Carter model is applied to data from Sweden in the long-term perspective. The Singular Value Decomposition (SVD) was used to estimate the model's parameters. Identification of a common trend of mortality change has been attempted by fitting a standard Lee-Carter model to different time series (1860-2004, 1900-2004, 1950-2004 and 1980-2204). We concluded by forecasting the mortality rates for 1901-2004 and 1951-2004 based on a total of seven different estimation periods and comparing them with the results obtained by application of the extended Lee-Carter model with a constant \hat{b}_x . The results indicate that the selection of an appropriate estimation period is important for forecasting mortality.

*Postal address: Dept. of Mathematical Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: wzheng99@hotmail.com. Supervisor: Anders Martin-Löf.

Acknowledgement

I would like to give my sincere thanks to my supervisor Erik Alm, General Manager and Chief Actuary Hannover Life Re Sweden, for his support, patience and guidance through all this work. Erik has given me many inspiring ideas and a lot of productive discussions during my study. His help and encouragement are essential to this work. I wish to thank the staff of Hannover Life Re Sweden for providing me a friendly environment to work during the summer.

I would also like to thank my supervisor at Mathematical Statistical Institution of Stockholm University, Professor Anders Martin-Löf, for his valuable thoughts and comments. I am also grateful to the faculties of Mathematical Statistical Institution at Stockholm University for guiding me into the fascinating field of the mathematical and statistical sciences.

Special thanks are extended to Gunnar Brännstam, Chief Actuary Handelsbanken Liv/ SPP, for leading me into the world of the actuary. A very special thank to Brigitte Koerfer for her kind encouragement and support.

Last, but not the least, I would like to express my great appreciation to my parents, my sister and my husband for always being there and supporting me.

Contents

- 1. Introduction 4
- 2. The data and method 5
 - 2.1 The data 5
 - 2.2 The Lee-Carter Model 8
- 3. Fitting and Applying the Model 9
 - 3.1 The SVD method 10
 - 3.2 Applying the Lee-Carter Model 12
 - 3.2.1. The period of 1860-2004 12
 - 3.2.2. Periods of 1900-2004 / 1950-2004 / 1980-2004 18
 - 3.2.3. Graphical Presentation of Residual Term 20
- 4. Forecasting 27
 - 4.1. Forecast the mortality index 27
 - 4.2 Performance and result 30
 - 4.3 Discussion: Forecasting with a constant \hat{b}_x 41
- 5. Conclusion 46
- Appendix: 48
- References 50

1. Introduction

The attempt to find an appropriate mortality curve has a long history in demography and actuarial sciences. Traditionally, a parametric curve was fitted to annual mortality rates. The most famous researchers in the history are deMoivre (1725), Gompertz (1825), Makeham (1860), Sang (1868) and Weibull (1939). Over the past ten years, a number of new approaches have been developed for forecasting mortality using stochastic models, such as Alho (1990, 1992), Alho and Spencer (1985, 1990), McNown and Rogers (1989, 1992), Bell and MOnsell (1991), and Lee and Carter (1992).

Recently, the Lee-Carter model became more and more popular and was applied for long-run forecasts of age specific mortality rates from many countries and time periods. This model is computationally simple to apply and it has given successful results for various countries, for instance, U.S. (Lee and Carter 1992), Canada (Lee and Nault 1993), Chile (Lee and Rofman 1994), Japan (Wilmoth 1996), the seven most economically developed nations (G7) (Tuljapurkar et al. 2000) and Belgium (Brouhns et al. 2002). Interestingly, the model did not succeed in Australian data (Booth et al. 2002) and the U.K. (Renshaw and Haberman 2003).

As is well known, Sweden has a long tradition of applying the Makeham method to adjust mortality rates. It could therefore be useful to study other methods for Swedish data as well. At present, several researchers have applied Lee-Carter to Swedish data — Hans Lundström and Jan Qvist (2002) examined how the Lee-Carter model operates with Swedish data for the period 1901-2001. Peter Wohlfart (2006) compared the differences in mortality rates and expected lifetime between Sweden and Denmark by using the Lee-Carter model, whereas Estrella Zarate (2006) applied the Lee-Carter model to the insured individuals' data for a pension fund, KP Pension.

In this paper, we will focus on a long-term study of mortality rates for Swedish population data, and give an overview of the Lee-Carter model through describing the basic method; discussing applications and extensions and evaluating the performance of the method. In addition, we will study how a forecaster could get a better performance of the model by selection of an optimal time period to fit the model.

The paper is divided into several sections. In Section 2, we introduce some historical background on Swedish data, discuss how to deal with data and describe the basic model of the Lee-Carter model. Section 3 we present the method for deriving the model's components in details and discuss the model and its properties by applying the method to different time series. Residual terms are discussed in this part as well. In section 4, we forecast the mortality rate and present the results from the evaluation. Finally, Section 5 gives a summary of our study.

2. The data and method

2.1 The data

The data source used for the studies made in this paper is “Human Mortality Database” (www.mortality.org).

The Human Mortality Database (HMD) began in the year 2000 and was launched in May 2002 after its first phase of development. It received financial and logistical support from its two sponsoring institutions — the Department of Demography at the University of California, Berkeley in US and at the Max Planck Institute for Demographic Research in Rostock, Germany. It also has financial support from the National Institute on Aging, USA and received technical advice and assistance from many other international collaborators.

The database provides detailed mortality and population data to researchers, students, journalists, policy analysts, and others interested in the history of human longevity and this database is free of charge. Currently, it contains detailed data for a collection of 28 countries.

The information in the HMD is standardized and includes the following types of data:

- Live birth counts,
- Death counts,
- Population size on January 1st,
- Population exposed to risk of death (period & cohort: period data are indexed by year of death, whereas cohort data are indexed by year of birth),

- Death rates (period & cohort), and
- Life tables (period & cohort).

All HMD data files are organized by sex, age and time. Population size is given for one-year and five-year age groups. One-year age groups means 1, 2, ..., 109, 110+; and five-year age groups means 0, 1-4, 5-9, 10-14, ..., 105-109, 110+. Age groups are defined in terms of actual age, for instance, “10-14” extends from exact age 10 to right before the 15th birthday. In this paper, the data we will use is the death rate *period* and the population size is *five-year* age groups.

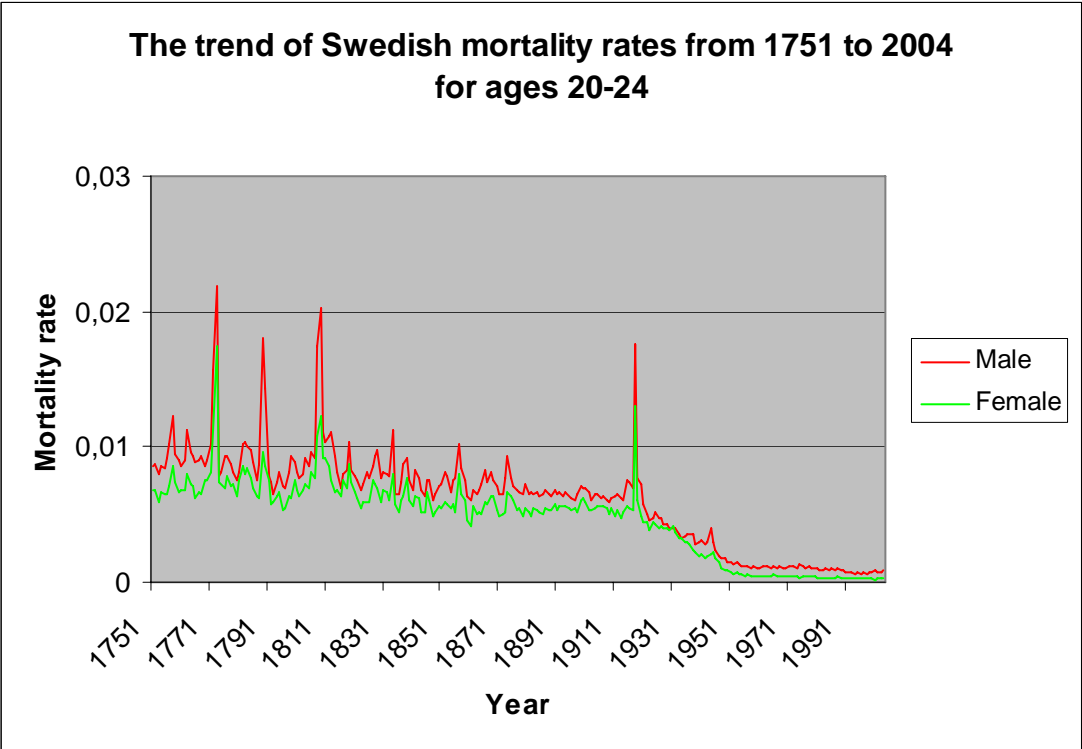
In the HMD, the mortality series for Sweden begins in 1751, which is the earliest record in the database. In fact, Sweden already began to keep a complete and continuously updated register of its population more than 300 years ago. The registration was implemented based on local initiatives. In 1686, these different initiatives were followed by a unitary decree for all of Sweden. From that year, it was mandatory for each parish to keep registers on baptisms, burials, marriages, divorces, and migration, as well as a population register. The ecclesiastical decree of 1686, which prescribed that everyone in Sweden was to profess allegiance to the Lutheran church, was the most important event in making the Swedish population registration efficient, even at an early stage.

The first attempts to collect population statistics from the church registers began in 1721. It followed by a plan for systematic collection of population statistics, the so-called *Tabellverket*. Data collection started in 1749, the clergy was asked to complete forms using data from the parish registers in every parish. Summaries were then made for the rectorial districts, the rural deaneries, and the diocese. Finally, the forms for the diocese were sent to the Chancery Committee.

Problems of missing forms and other errors in each stage of the process were resolved over time. The statistics can be considered more reliable after 1802. In 1858, with the founding of Statistics Sweden [Statistiska Centralbyrån], the compilation of population statistics was reorganized. Since 1860, the compilation has been based on copies of all parish registers sent to Statistics Sweden. Thus, for the first time, it is feasible to check all information very carefully. Population data from 1860 can be considered to be of very high quality.

In general, Swedish mortality rates have a downward trend from 1750 to 2004. However the declining mortality was interrupted by periods of high death rates due to epidemics. In 1771-1772, a harvest failure led to famine and epidemics, resulting in increased deaths during 1772-1773. Another increase in mortality during the first decade of the 19th century coincided with the Finnish War of 1808-09. The Spanish influenza epidemic of 1918-19 also resulted in increased death rates, especially among those aged 15 to 40. Because Sweden remained neutral during both world wars, it was minimally impacted by the war relative to other European countries. In Figure 1, we take the age group 20-24 as an example to present the trend of Swedish death rates since 1750. The periods of high death rates due to above-mentioned epidemics are shown clearly.

Figure1. : The trend of Swedish mortality rates, 1751-2004, ages 20-24.



2.2 The Lee-Carter Model

In 1992, based on a combination of statistical time series methods, Lee and Carter have developed a new model that could be used for the extrapolation of trends and age patterns in mortality.

The Lee-Carter methodology for forecasting mortality rates is a simple bilinear model in the variables x (age) and t (calendar year). The model is defined as:

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

Where

$m_{x,t}$: observed central death rate at age x in year t

a_x : average age-specific pattern of mortality

b_x : pattern of deviations from the age of profile as the k_t varies

k_t : a time-trend index of general mortality level

$\varepsilon_{x,t}$: the residual term at age x and time t .

The time component k_t captures the main time trend on the logarithmic scale in mortality rates at all ages. The model includes no assumption about the nature of the trend in k_t . The age component b_x modifies the main time trend according to whether change at a particular age is faster or slower than the main trend. In principle, not all the b_x need have the same sign, in which case movement in opposite directions could occur. In practice, all the b_x do have the same sign, at least when the model is fit over fairly long periods. The model assumes that b_x is invariant over time.

In order to obtain a unique solution for the system of equations of the model, a_x is set equal to the averages over time of the $\ln m_{x,t}$, the square values of b_x sum to unity, and k_t values sum to zero, these are of the forms:

$$a_x = \frac{1}{T} \sum_t \ln m_{x,t} \quad , \quad \sum_x b_x^2 = 1 \quad , \quad \sum_t k_t = 0 \quad (2)$$

3. Fitting and Applying the Model

To ensure good performance, an appropriate method to estimate the model's parameters has to be decided carefully at the very beginning.

In their original paper, Lee and Carter (1992) applied a two-stage estimation procedure. In the first stage, singular value decomposition (SVD) is applied to the matrix of $\{\log(m_{x,t}) - a_x\}$ to obtain estimates of b_x and k_t . Then in the second stage, the time series of k_t is re-estimated by the method of so called "second stage estimation". Lee and Carter noticed that once b_x and k_t have been estimated, the observed total number of deaths $D_t \equiv \sum_x D_{xt}$ is not guaranteed to be equal to the fitted number of deaths. Therefore, they made a second stage estimation of k_t by finding a value that makes the observed number of deaths equal to the predicted number of deaths. That is, they searched for k_t such that

$$D_t = \sum_x \{\exp(a_x + b_x k_t) N_{x,t}\}, \quad (3)$$

where

D_t is the total number of deaths in year t ;

$N_{x,t}$ is the population (exposure to risk) of age x in year t .

There are several advantages to make a second stage estimate of the k_t , which were described in details by Lee and Carter (1992). These can be useful in the life table presentation of the data and especially in cases where only the total, rather than age-specific, death rates are known in certain years.

However, different criteria have been proposed for this method. Wilmoth (1993) believe the reason for the differences in the observed number of deaths and the fitted number is that the estimates of k_t are computed by minimizing the least square error over log-mortality rather than mortality itself. As a result, age groups with small numbers of deaths had the same weight as age groups with large numbers of death, even though they contributed very little to the overall mortality. Therefore, Wilmoth developed two alternative one-stage estimation strategies — a weighted least square (WLS) and a maximum likelihood (MLE) technique.

Currently, the Singular Value Decomposition (SVD), the Weighted Least Square method (WLS) and the Maximum Likelihood Estimate (MLE) have become the three most used methods for estimating the model's parameters. A study by Marie Claire Koisst, Arnold Shapiro, and Göran Högnäs has compared these three methods. Their results showed that SVD is the best alternative for the mortality index k , especially for the data from Nordic countries. Another advantage of SVD is that it can be easily facilitated by using *Biplot** (see *Appendix2*) in Microsoft Excel.

In this study, we skip the re-estimation stage and apply the first stage SVD method directly.

3.1 The SVD method

From equation (2) in section 2.2, the parameter vector a_x can be easily computed as the average over time of the logarithm of the central death rate. Then we apply the Singular Value Decomposition to matrix $Z_{x,t} = \ln(m_{x,t}) - \hat{a}_x$, producing the matrices $ULV' = SVD(Z_{x,t}) = L_1 U_{x1} V_{t1} + \dots + L_X U_{xX} V_{tX}$. Approximation to the first term gives the estimates $\hat{b}_x = U_{x1}$ and $\hat{k}_t = L_1 V_{t1}$. The whole application of the SVD method is relative simple and follows six steps:

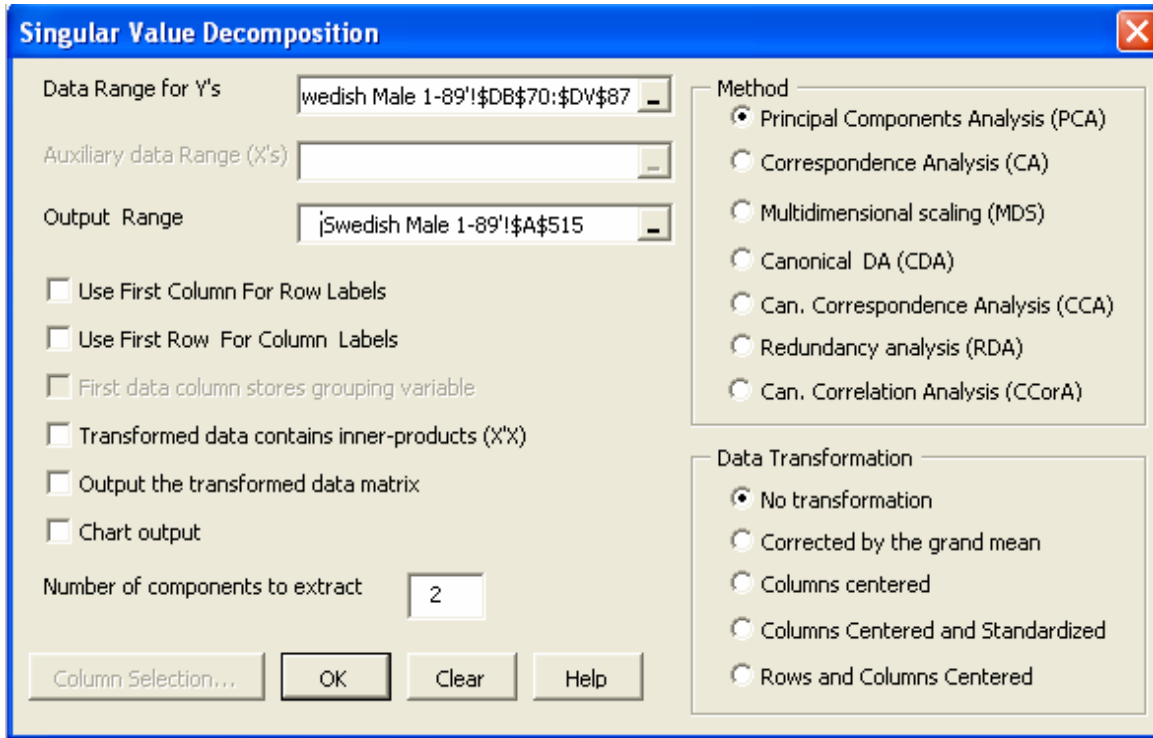
Step 1.
$$\hat{a}_x = \frac{1}{T} \sum_{t=t_1}^{t_n} \ln(m_{x,t})$$

Step 2. Create a matrix $Z_{x,t}$ for estimating b_x and k_t , where $Z_{x,t} = \ln(m_{x,t}) - \hat{a}_x = b_x k_t$

Step 3. Apply the Singular Value Decomposition to matrix $Z_{x,t}$, which decomposes the matrix of $Z_{x,t}$ into the product of three matrices : $ULV' = SVD(Z_{x,t}) = L_1 U_{x1} V_{t1} + \dots + L_X U_{xX} V_{tX}$, where U representing the age component, L is the singular values and V representing the time component.

Step 4. Select *Singular Value Decomposition Dialog* (see the graph) from Biplot in Microsoft Excel, by running the program. \hat{k}_t is derived from the first vector of the time-

component matrix and the first singular value ($\hat{k}_t = L_1 V_{t1}$), and \hat{b}_x is derived from the first vector of the age-component matrix ($\hat{b}_x = U_{x1}$).



Step 5. (Lee-Carter) Approximate a new matrix $\hat{Z}_{x,t}$ by the product of the estimated parameters \hat{b}_x and \hat{k}_t and get $\hat{Z}_{x,t_1} = \hat{b}_x \hat{k}_t$.

$$\hat{Z}_{x,t} = \begin{bmatrix} \hat{Z}_{x_1 t_1} & \hat{Z}_{x_1 t_2} & \dots & \hat{Z}_{x_1 t_n} \\ \hat{Z}_{x_2 t_1} & \hat{Z}_{x_2 t_2} & \dots & \hat{Z}_{x_2 t_n} \\ \dots & \dots & \dots & \dots \\ \hat{Z}_{x_A t_1} & \hat{Z}_{x_A t_2} & \dots & \hat{Z}_{x_A t_n} \end{bmatrix} \hat{b}_x \hat{k}_t$$

Step 6. Estimate the logarithm of the central death rate, $\hat{\ln}(m_{x,t}) = \hat{a}_x + \hat{Z}_{x,t} = \hat{a}_x + \hat{b}_x \hat{k}_t$

3.2 Applying the Lee-Carter Model

3.2.1. The period of 1860-2004

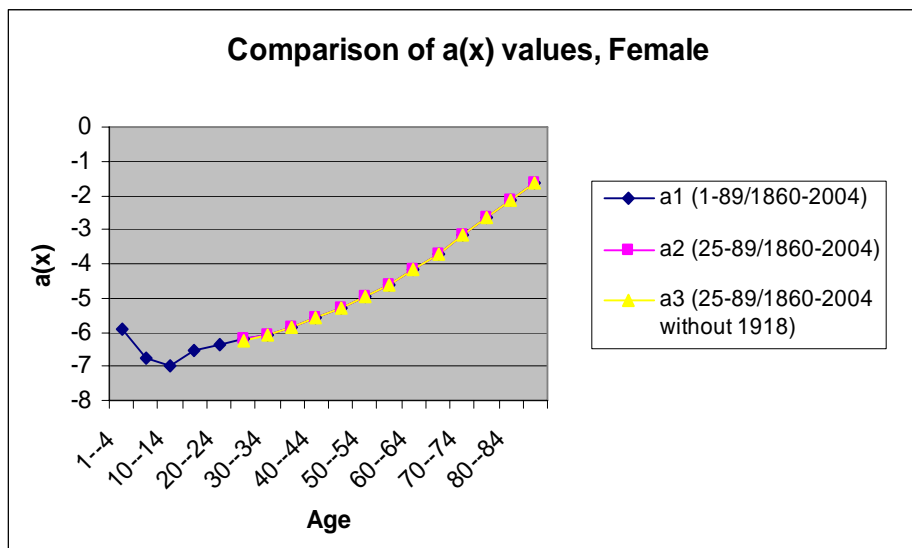
We start by applying the Lee-Carter model for estimating the whole period 1860-2004 for all ages 1-89. By using the SVD (*Singular Value Decomposition*) method, we obtain the values for model components \hat{a}_x , \hat{b}_x , and \hat{k}_t . Before going to the next step of estimation, we would like to check our model by checking two more things: sums of squares of residuals per age (SResi.A.) and sums of squares of residuals per year (SResi.Y) (residual terms $\varepsilon = \ln_{M_{x,t}} - \hat{a}_x - \hat{b}_x \hat{k}_t$). As the Table 1 shows, the SResiA we obtain for ages 1-24 is quite high when we consider the whole period and all ages 1-89. Therefore, we decide to remove those youngsters and obtain the new vectors of \hat{a}_x , \hat{b}_x and SResi, the resulted values of SResi.A became lower in general. We thereafter take the SResi.Y (Sums of squares of residuals per year) into consideration, and observe that the value of SResi.Y for year 1918 is extremely high. This is probably due to the effect of Spanish influenza epidemic. In order to get a better performance, we decide to remove the year 1918 from the data and calculate the parameters again. In Figure 2, we take the female as an example and compare the values of \hat{a}_x , \hat{b}_x , \hat{k}_t , sums of squares of residuals per age and sums of squares of residuals per year for three different conditions: whole period 1860-2004 & all ages 1-89; whole period 1860-2004 & ages 25-89 and period 1860-2004 (without 1918) & ages 25-89. The estimation data from 1860-2004 (without 1918) & ages 25-89 give the best result. Therefore, all the data we used in the rest of our study only cover the age from 25 to 89 and skip the data from year 1918.

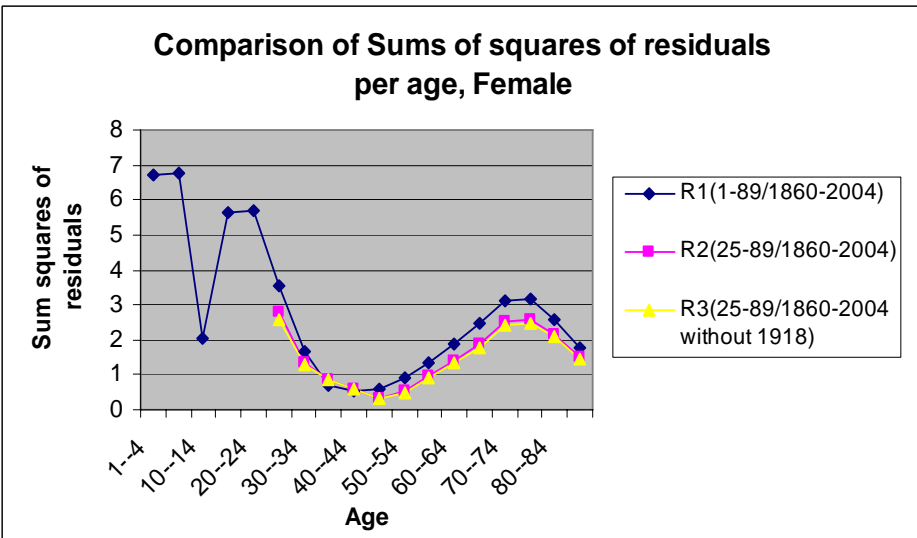
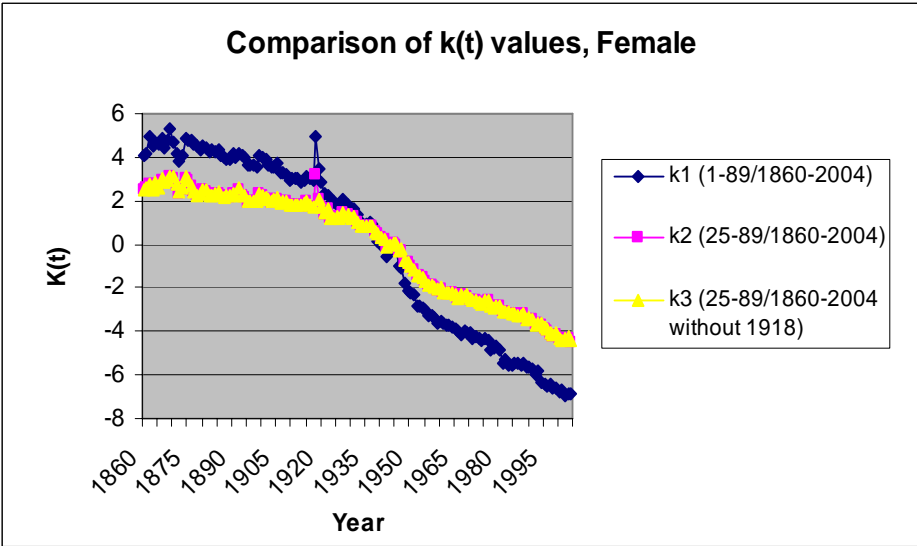
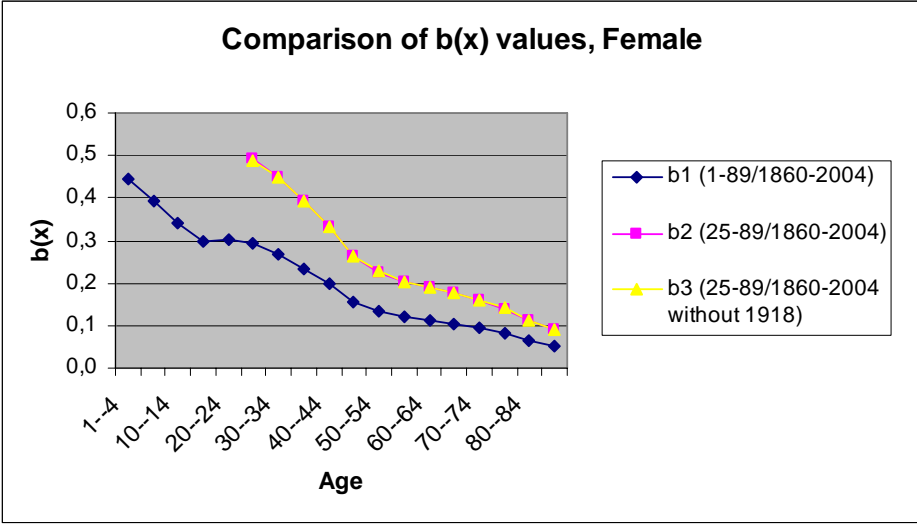
Table 1: \hat{a}_x , \hat{b}_x and Sums of squares of residuals per age for different periods and ages.

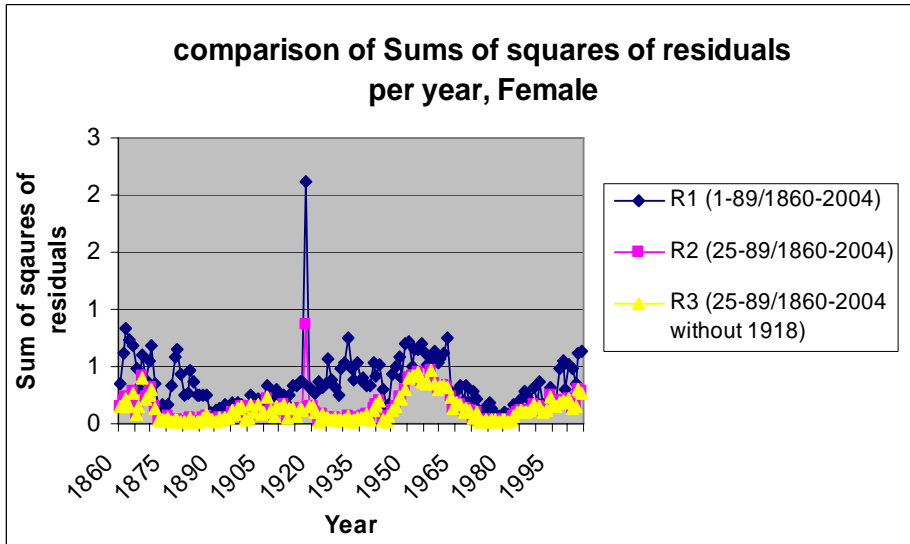
| Age / Period Female | Whole period 1860-2004 & all ages 1-89 | | | Whole period 1860-2004 & ages 25-89 | | | 1860-2004 (without 1918) & ages 25-89 | | |
|---------------------------|---|--------|-----------|--|--------|-----------|--|--------|-----------|
| | a1 | b1 | SResiA*.1 | a2 | b2 | SResiA*.2 | a3 | b3 | SResiA*.3 |
| 1--4 | -5,8916 | 0,4436 | 6,6950 | | | | | | |
| 5--9 | -6,7557 | 0,3932 | 6,7757 | | | | | | |
| 10--14 | -6,9633 | 0,3413 | 2,0563 | | | | | | |
| 15--19 | -6,5354 | 0,2982 | 5,6407 | | | | | | |
| 20--24 | -6,3501 | 0,3008 | 5,7052 | | | | | | |
| 25--29 | -6,2206 | 0,2919 | 3,5493 | -6,2206 | 0,4911 | 2,7756 | -6,2347 | 0,4893 | 2,588332 |
| 30--34 | -6,0644 | 0,2673 | 1,6389 | -6,0644 | 0,4491 | 1,3614 | -6,0766 | 0,4479 | 1,27936 |
| 35--39 | -5,8414 | 0,2348 | 0,6824 | -5,8414 | 0,3940 | 0,8853 | -5,8508 | 0,3936 | 0,883463 |
| 40--44 | -5,5866 | 0,1971 | 0,5496 | -5,5866 | 0,3308 | 0,6014 | -5,5933 | 0,3312 | 0,591558 |
| 45--49 | -5,3029 | 0,1559 | 0,5954 | -5,3029 | 0,2624 | 0,3191 | -5,3080 | 0,2628 | 0,304492 |
| 50--54 | -4,9706 | 0,1340 | 0,9309 | -4,9706 | 0,2259 | 0,5294 | -4,9744 | 0,2267 | 0,492283 |
| 55--59 | -4,6080 | 0,1192 | 1,3584 | -4,6080 | 0,2014 | 0,9475 | -4,6110 | 0,2023 | 0,892954 |
| 60--64 | -4,1785 | 0,1111 | 1,9035 | -4,1785 | 0,1881 | 1,4149 | -4,1809 | 0,1892 | 1,337243 |
| 65--69 | -3,6990 | 0,1040 | 2,4667 | -3,6990 | 0,1765 | 1,8894 | -3,7010 | 0,1776 | 1,796457 |
| 70--74 | -3,1800 | 0,0938 | 3,1294 | -3,1800 | 0,1598 | 2,5046 | -3,1816 | 0,1609 | 2,412575 |
| 75--79 | -2,6455 | 0,0815 | 3,1454 | -2,6455 | 0,1392 | 2,6036 | -2,6463 | 0,1405 | 2,4879 |
| 80--84 | -2,1294 | 0,0656 | 2,5610 | -2,1294 | 0,1123 | 2,1473 | -2,1304 | 0,1132 | 2,090969 |
| 85--89 | -1,6420 | 0,0520 | 1,7713 | -1,6420 | 0,0892 | 1,4883 | -1,6424 | 0,0901 | 1,427563 |

* SResi.A = Sums of squares of residuals per age

Figure 2. Estimate performances with three different conditions

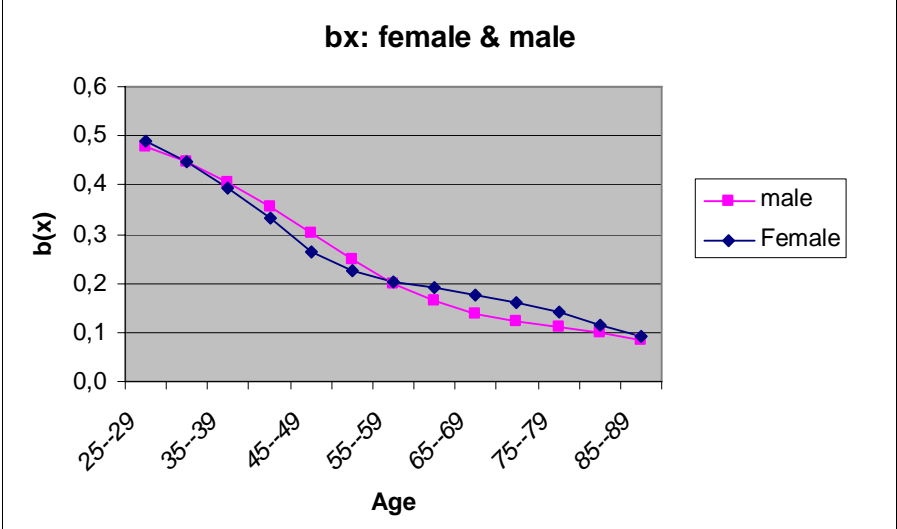
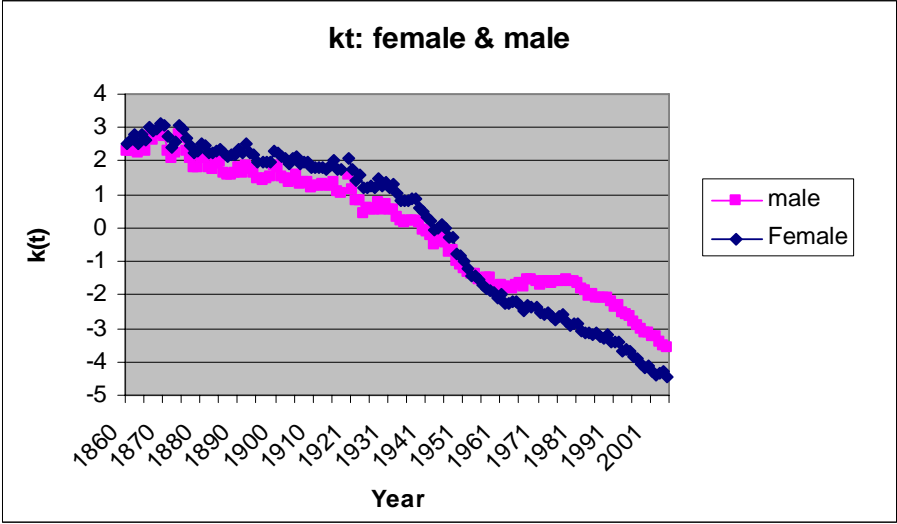
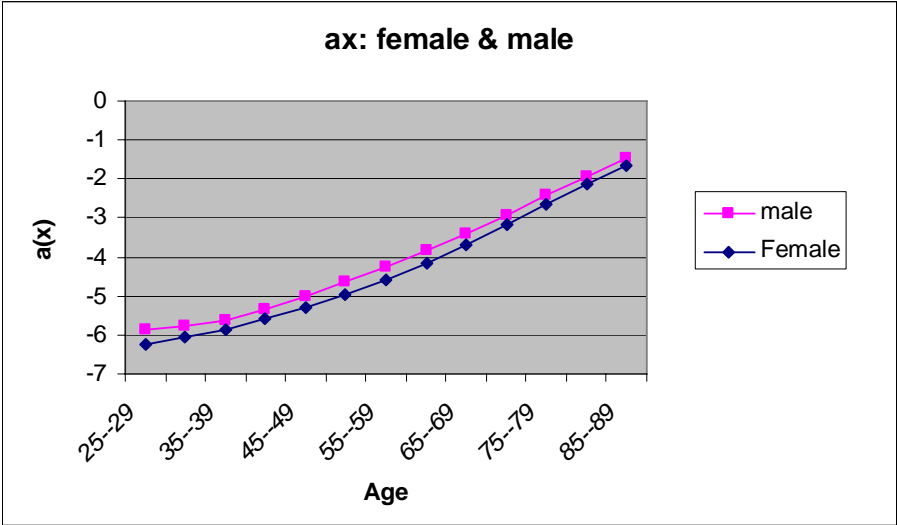






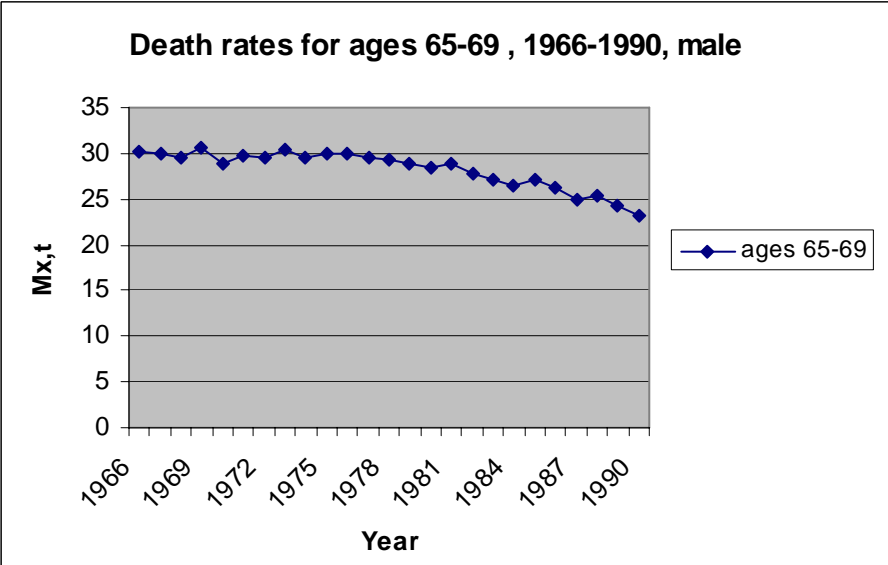
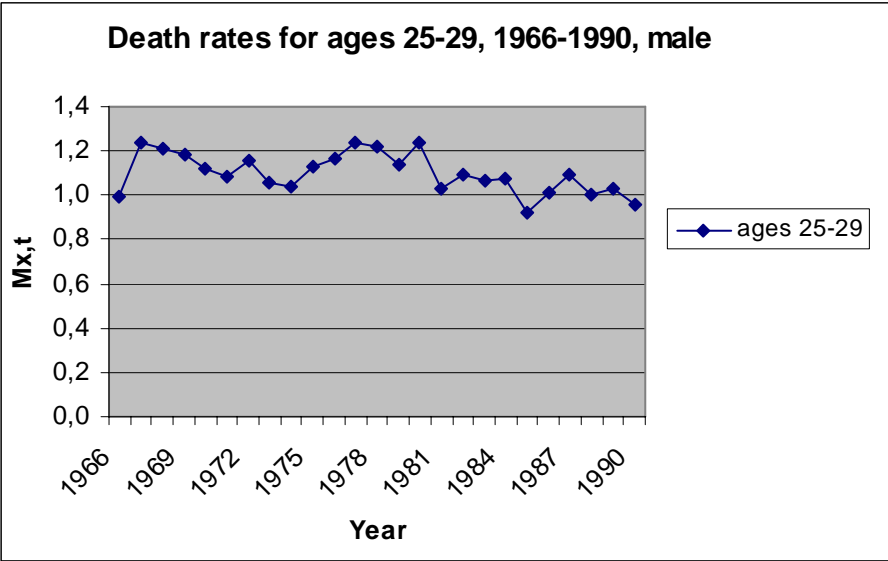
In order to explore the properties of the three components of the model: \hat{a}_x , \hat{b}_x and \hat{k}_t , we plot them for females and males in Figure 3. Parameter \hat{a}_x represents the general age shape of mortality. From this figure, we find that both females and males have upward trend of mortality in general, whereas the younger ages have a lower mortality and the older ages have a higher mortality. The mortality index, \hat{k}_t , captures the main time trend on the logarithmic scale in death rates at all ages. From Figure 3, we find that men have a significant change of mortality rates after 1965, which was probably caused by a change of smoking habit at that time. Parameter \hat{b}_x describes the tendency of mortality at age x to change as the general level of mortality (\hat{k}_t) changes. This indicates that when \hat{b}_x is large for some x , the death rate at age x varies a lot than the general level of mortality change and when \hat{b}_x is small, then the death rate at that age varies a little.

Figure 3: Comparison of the components between female and male, period 1860-2004.



These characteristics can be observed in Figure 4, where we compare the mortality between men aged 25-29 and aged 65-69 for the period of 1966-1990. This period was chosen since we have demonstrated that was a change in mortality pattern for men after 1965. The results showed that men aged 25-29 with larger value of \hat{b}_x (se Figure 2) have a much more fluctuant mortality than men aged 65-69.

Figure 4: Performance of death rates for selected ages and periods, male.
 ($M_{x,t}$ Display in thousands)



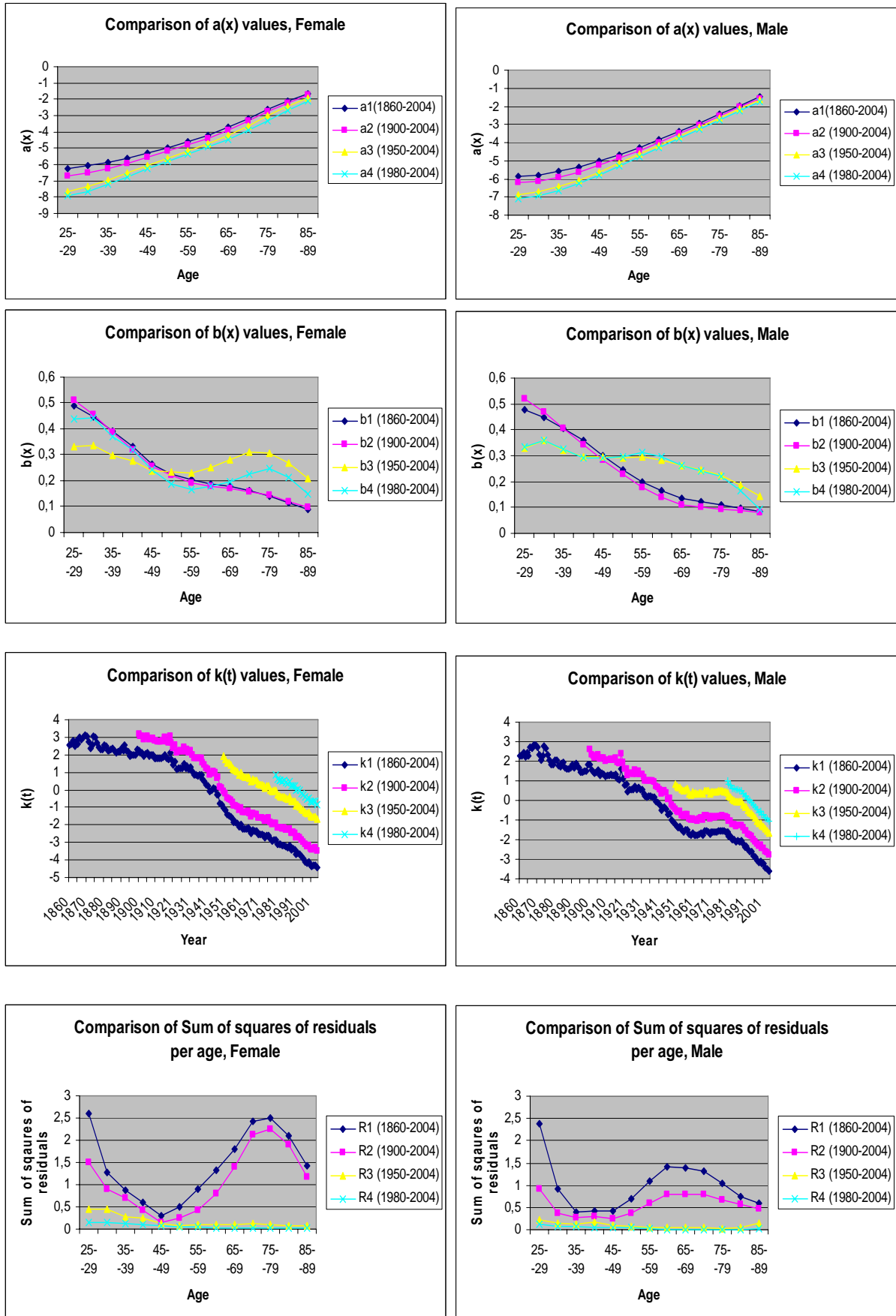
3.2.2. Periods of 1900-2004 / 1950-2004 / 1980-2004

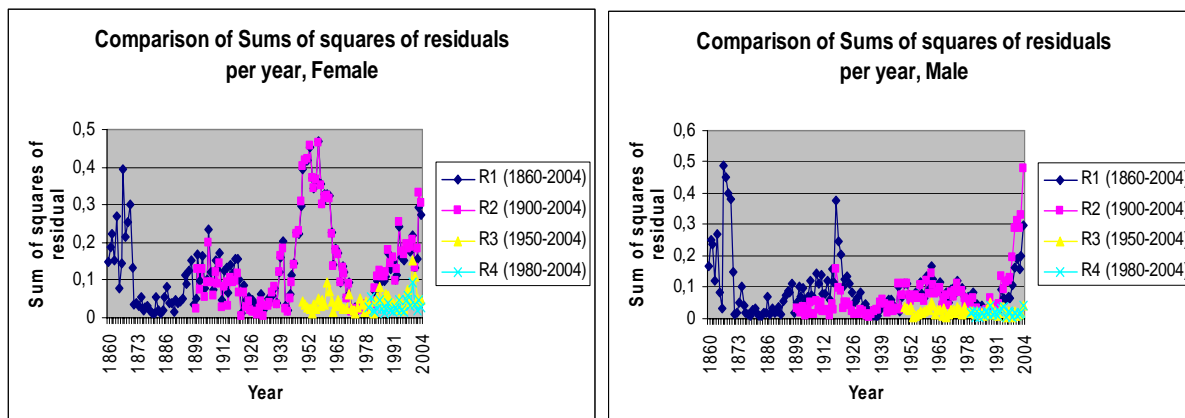
We have applied the Lee-Carter model to mortality rates in time series from 1860 to 2004. We now apply this method to three sub-samples of 1900-2004, 1950-2004 and 1980-2004. In the same way as before, we estimate the \hat{a}_x and \hat{b}_x for each set of data and then estimate \hat{k}_t for these years based on these \hat{a}_x and \hat{b}_x estimates.

The comparison among the entire time series 1860-2004 and the three sub-samples is presented in Figure 5. In general, mortality has declined continuously over the course of the 20th century. We also observe that the pattern in 1860-2004 is similar to the period of 1900-2004; while 1950-2004 and 1980-2004 give similar patterns. The main difference between these two “blocks” is resulting from growing importance of medical care. Throughout the later 19th century and first half of the 20th century, infectious diseases were the leading cause of death. Pneumonia, tuberculosis and influenza were the biggest killers. Early in the 20th century, mortality began to decline, thanks to public health and economic measures that improved peoples’ ability to withstand diseases. Difference in age group is also observed — mortality reductions are concentrated at younger ages. As shown in Figure 5, mortality for younger ages declines a lot when we compare age patterns of mortality of 1900-2004 and 1950-2004. This mortality decline is very likely resulting from the more significant impact of nutrition and public health on the young people than the old.

By the mid 20th century, infectious diseases continued to decline, which is probably due to medical factors. Antibiotics, including penicillin and sulfa drugs, became important contributors to mortality reduction in this era. Antibiotics help the elderly as well as the young, so mortality reductions became more widespread across the age distribution, which could explain mortality declines almost at same rate for young and old ages in 1950-2004 and 1980-2004.

Figure 5: Comparison of the components of the entire time series 1860-2004 and three selected sub-samples 1900-2004, 1950-2004 and 1980-2004, male and female.





3.2.3. Graphical Presentation of Residual Term

Figures 7, 8, 9 and 10 are the graphical presentations of the residual term of the mortality rate on a logarithmic scale. In order to highlight the relevant features, the unit term is standardized in the corresponding periods. Figure 7 shows the residual term of the entire time series of 1860-2004. Most parts of the graph appear random or less systematic, but a systematic pattern is still noticeable, such as during period 1945-1975 for both females and males. We think tuberculosis could probably be a contributing factor for the systematic pattern. As we mentioned previously, tuberculosis has been the leading cause of high death rates, especially for women. Figure 6 demonstrates this unusual period more clearly by plotting the sum of squares of residuals per year. The same effect can be easily recognized in Figure 8 as well.

However, in Figures 9 and 10, we recognize a (very weak) cohort effect as a diagonal line for males. We may conclude that the males born in 1943 have slightly higher death rates. An example is given in Figure 9. We could follow a man who is 27 years old at year 1970 and thus 37 years old at 1980 and 57 years old at 2000. We find that it is almost a diagonal line in blue color from year 1970 to 2000 corresponding to the age groups from 25-29 to 55-59. Furthermore, we express the residual term as residual death rates instead of a logarithmic value in Figure 11. It shows this effect more clearly, even though the effect is not very significant. Finally, most of the high value of residuals occurs at younger ages, which reflecting both the greater irregularity in death rates at these ages and the smaller weights at ages with smaller numbers of deaths. We also noticed that young women have especially larger value of residuals than men during the recent 20 years. It is probably caused by much

smaller numbers of deaths in females than males at the young age; young men have a higher chance to be killed in accidents, such as motorcycle accidents.

Figure 6: Sum of squares of residuals per year, Female & Male, 1860-2004

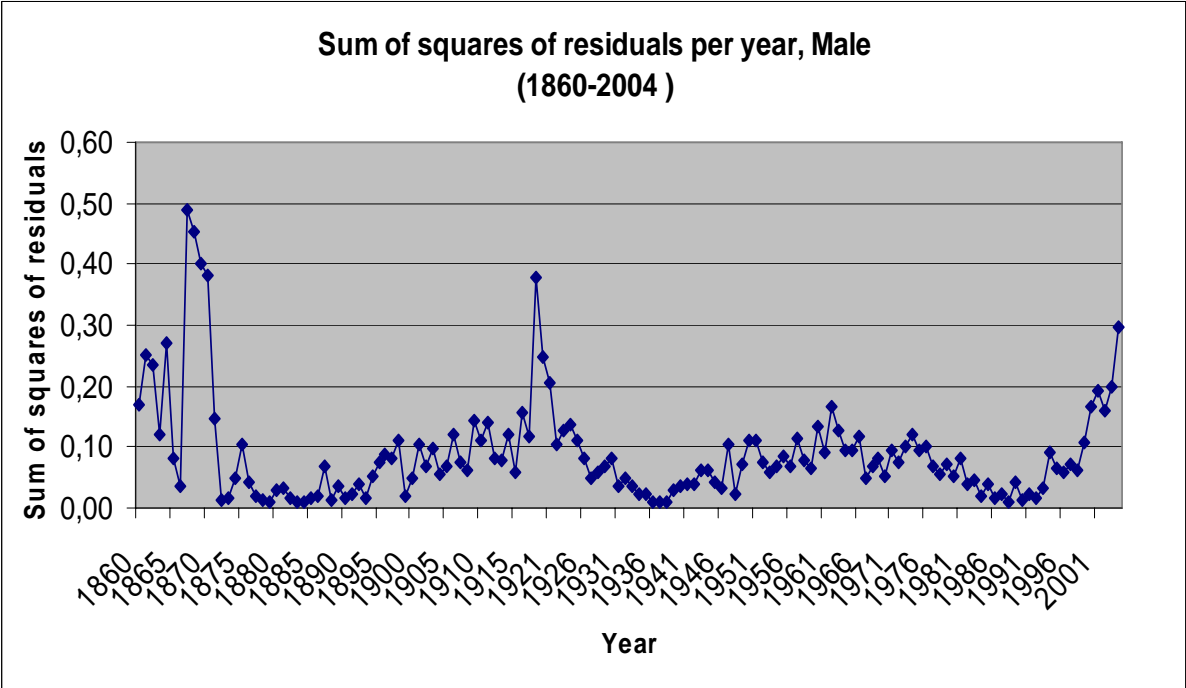
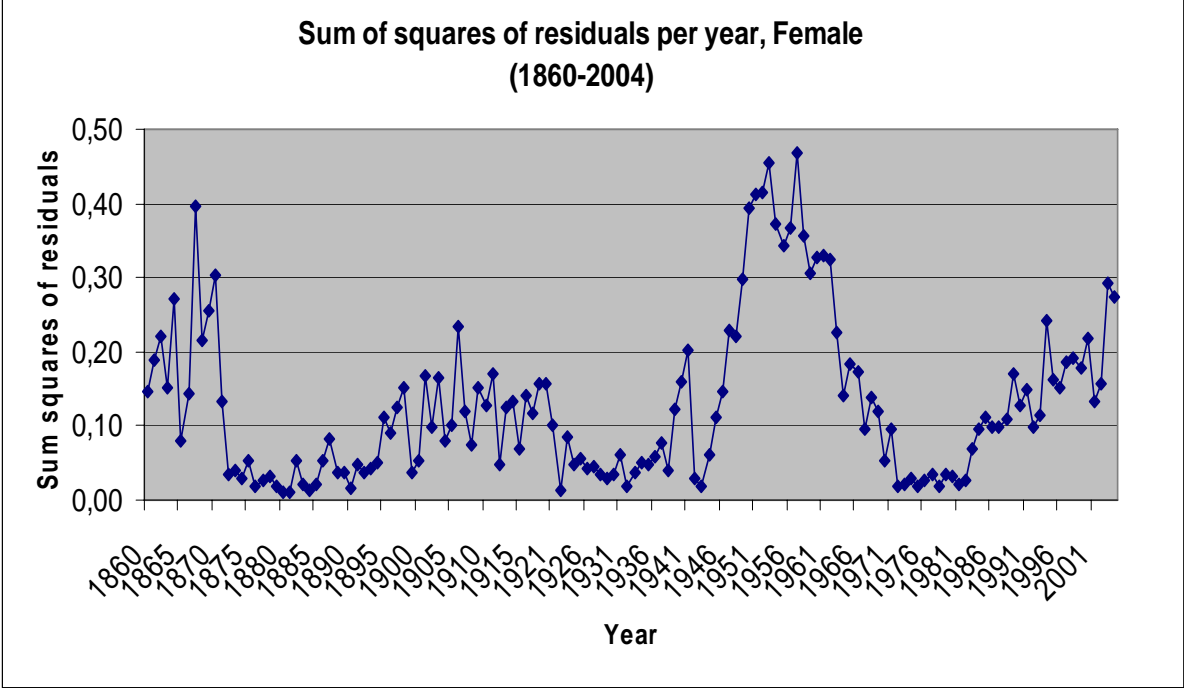
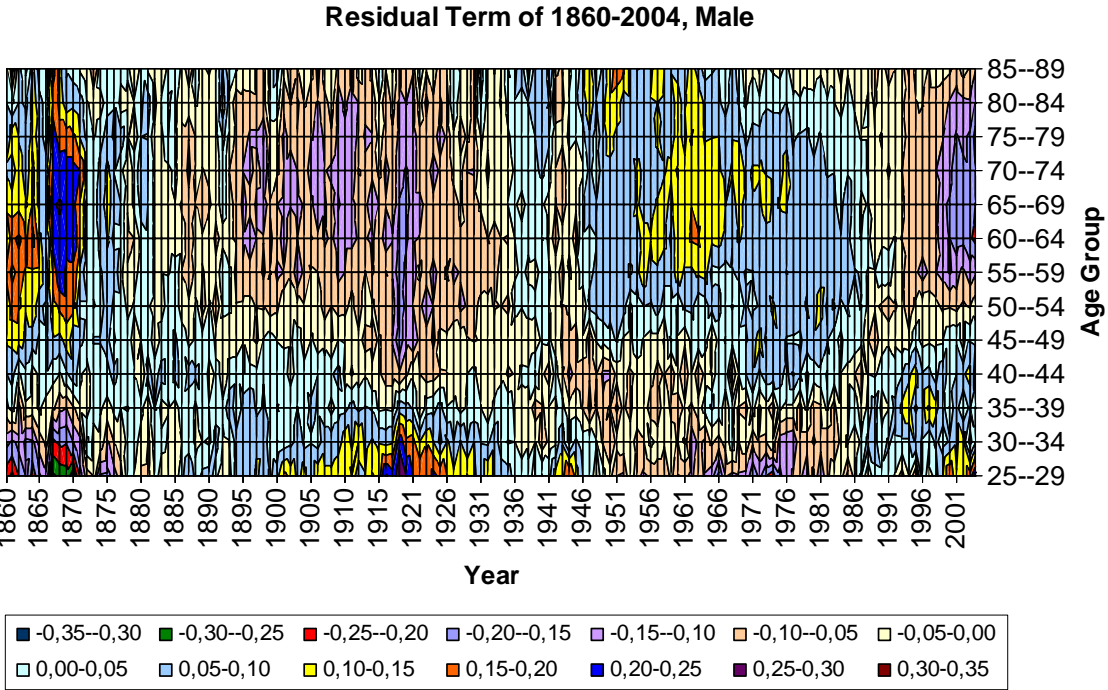
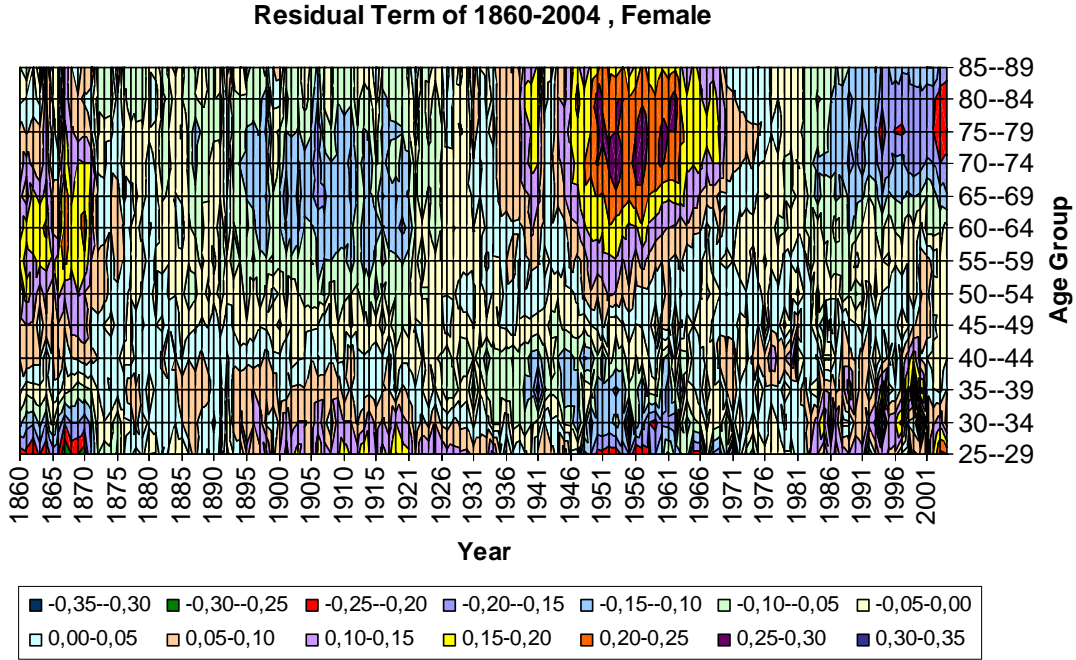
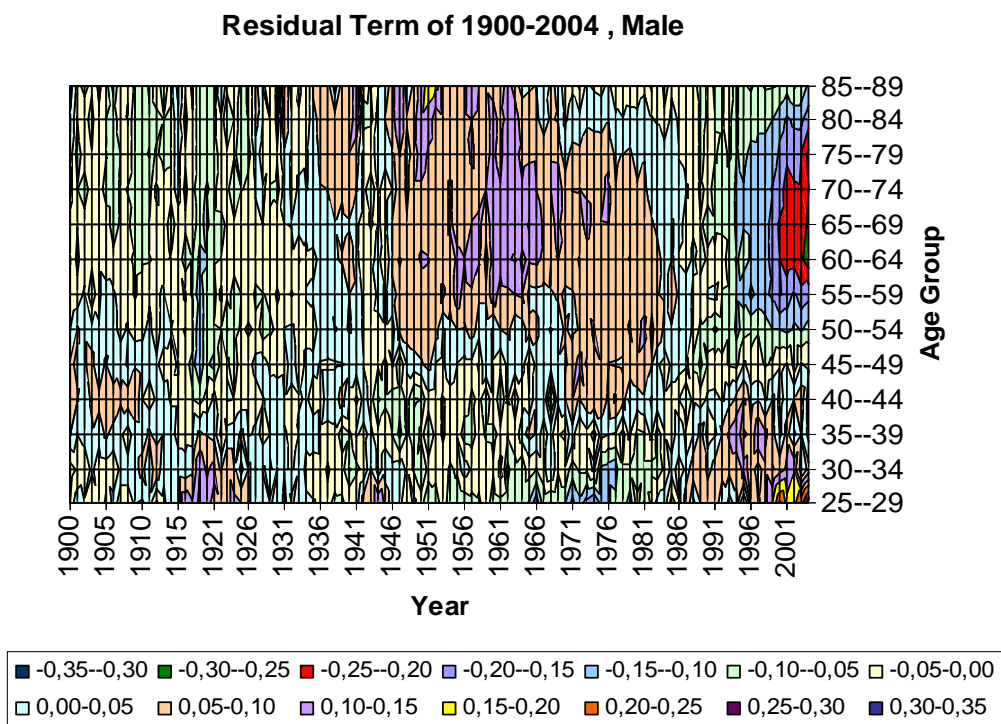
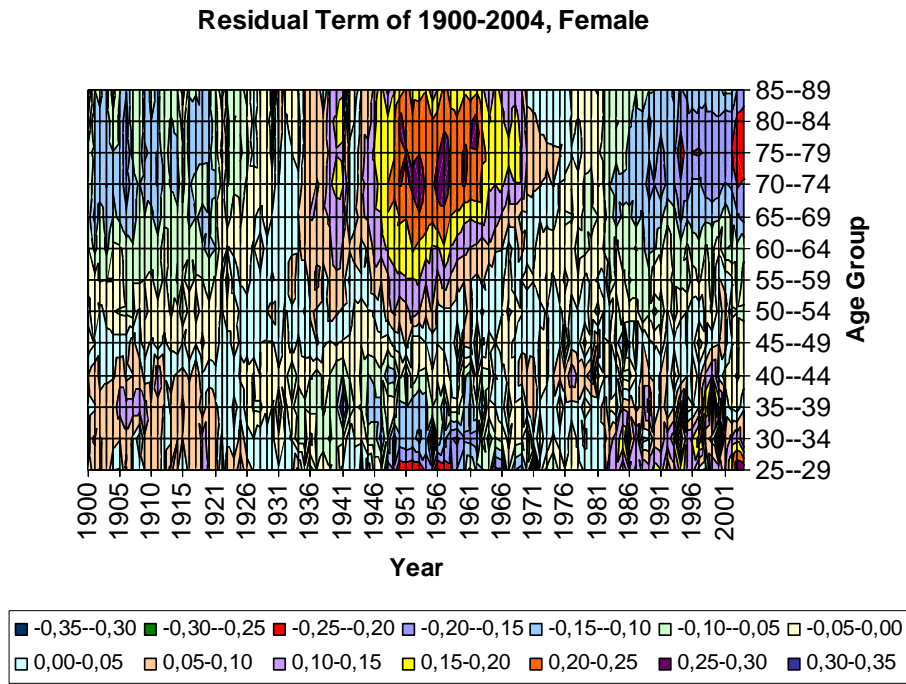


Figure 7: Lee-Carter Residual Term – Female & Male, 1860-2004



Note: Residuals are on a logarithmic scale.

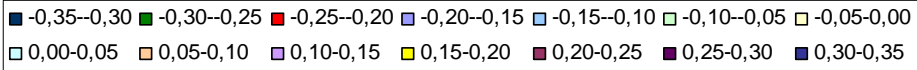
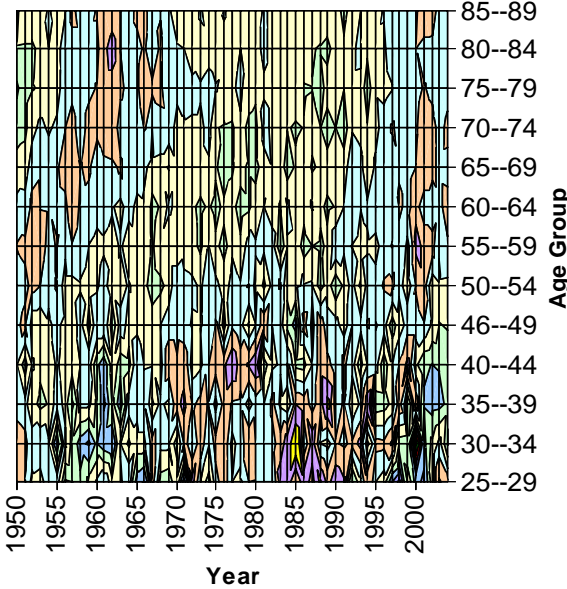
Figure 8: Lee-Carter Residual Term – Female & Male, 1900-2004



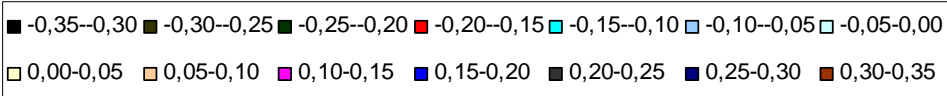
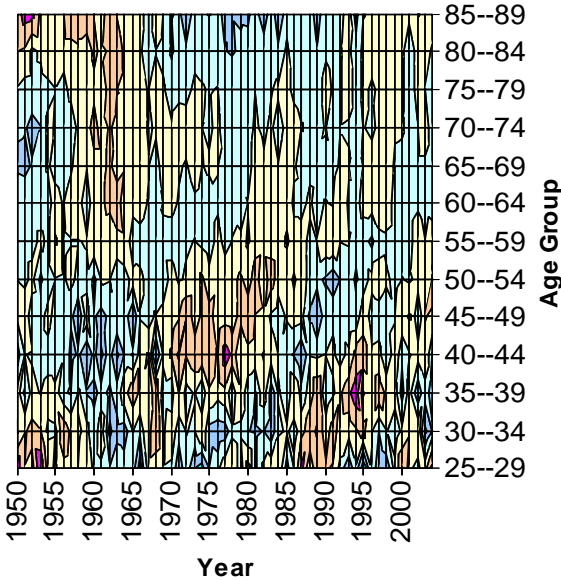
Note: Residuals are on a logarithmic scale.

Figure 9: Lee-Carter Residual Term – Female & Male, 1950-2004

Residual Term of 1950-2004, Female

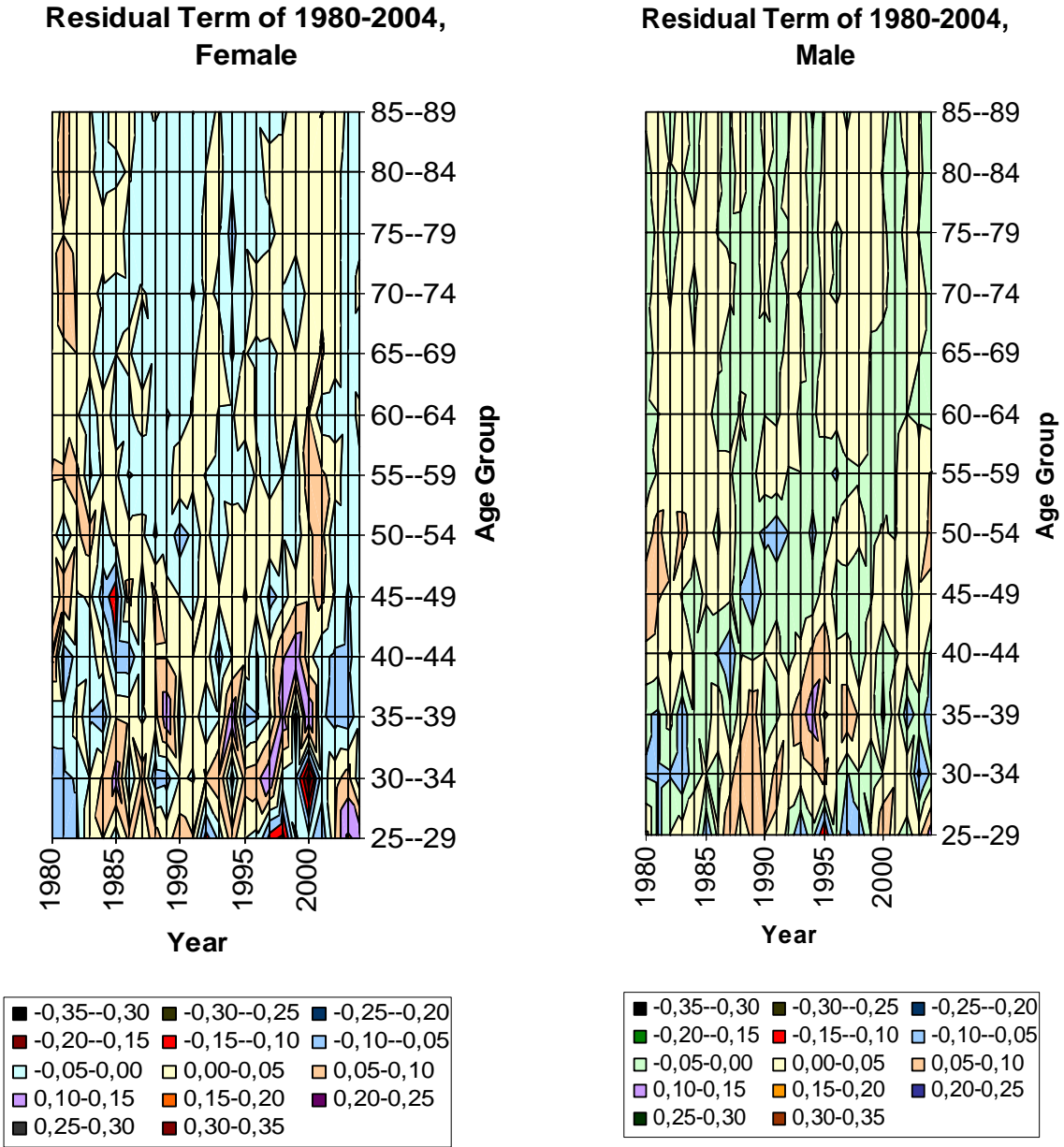


Residual Term of 1950-2004, Male



Note: Residuals are on a logarithmic scale.

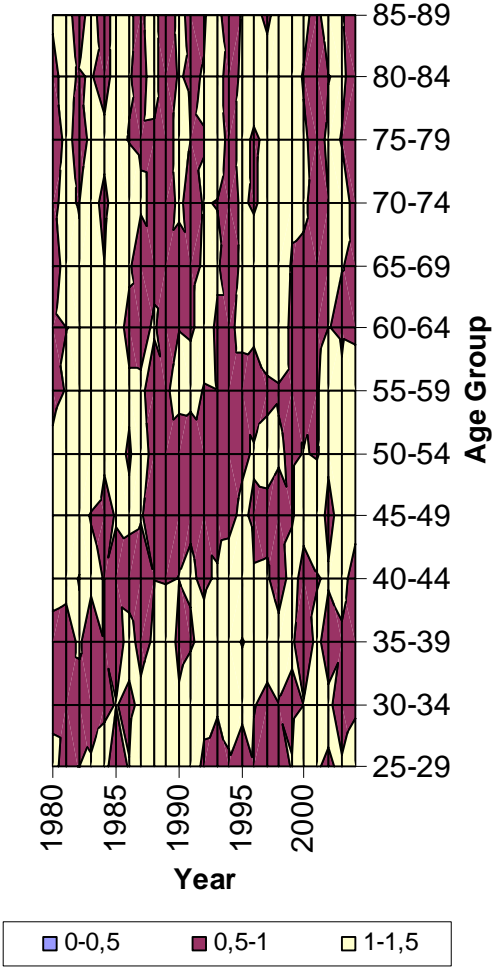
Figure 10: Lee-Carter Residual Term – Female & Male, 1980-2004



Note: Residuals are on a logarithmic scale.

Figure 11: Lee-Carter Residual death rates.

Residual death rates of 1980-2004, male



4. Forecasting

4.1. Forecast the mortality index

One advantage of the LC (Lee and Carter) approach is that once the data are fitted to the model and the values of the vectors \hat{a}_x , \hat{b}_x , and \hat{k}_t are found; only the mortality index \hat{k}_t needs to be predicted. Lee and Carter predicted the mortality index \hat{k}_t in their original paper by a standard univariate time series model ARIMA (0,1,0). They demonstrated that other ARIMA models might be preferable for different data sets, but in practice the random walk with drift model (RWD) for k_t has been used almost exclusively. The model is as follows:

$$\hat{k}_t = \hat{k}_{t-1} + \theta + \varepsilon_t \quad (4)$$

where θ is known as the *drift parameter* and

$$\hat{\theta} = \frac{\hat{k}_T - \hat{k}_1}{T - 1} \quad (5)$$

which means $\hat{\theta}$ only depends on the first and last of the k_t estimates; while ε_t is the *error term*. Then to forecast two periods ahead, we just substitute for the definition of \hat{k}_{t-1} moved back in time one period:

$$\begin{aligned} \hat{k}_t &= \hat{k}_{t-1} + \hat{\theta} + \varepsilon_t \\ &= (\hat{k}_{t-2} + \hat{\theta} + \varepsilon_{t-1}) + \hat{\theta} + \varepsilon_t \\ &= \hat{k}_{t-2} + 2\hat{\theta} + (\varepsilon_{t-1} + \varepsilon_t) \end{aligned} \quad (6)$$

To forecast \hat{k}_t at time $T + (\Delta t)$ with data available up to period T , we follow the same procedure and iterate (Δt) times and obtain:

$$\hat{k}_{T+(\Delta t)} = \hat{k}_T + (\Delta t)\hat{\theta} + \sum_n^{(\Delta t)} \varepsilon_{T+n-1}$$

$$= \hat{k}_T + (\Delta t)\hat{\theta} + \sqrt{(\Delta t)}\varepsilon_t \quad (7)$$

If ignore the error term, we can obtain forecast point estimates, which follow a straight line as a function of (Δt) , with slope $\hat{\theta}$:

$$\begin{aligned} \hat{k}_{T+(\Delta t)} &= \hat{k}_T + (\Delta t)\hat{\theta} \\ &= \hat{k}_T + (\Delta t)\frac{\hat{k}_T - \hat{k}_1}{T-1} \end{aligned} \quad (8)$$

The forecasting of \hat{k}_t is thus very simple: Extrapolate from a straight line drawn through the first \hat{k}_1 and last \hat{k}_T points, and all other \hat{k}_t points are ignored.

In this paper, instead of extrapolating from a straight line drawn through the first \hat{k}_1 and last \hat{k}_T points, which would make the slope $\hat{\theta}$ depending on only the first and last of the k_t estimates, we decide to extrapolate from a straight line drawn through a new point $\hat{k}_{T'}$ which is the average value of the last five points up to period T, and we also make a different value for the *drift parameter* $\hat{\theta}$. The new model is defined as:

$$\hat{k}_{T+(\Delta t)} = \hat{k}_{T'} + (\Delta t)\hat{\theta}_{T'} \quad (9)$$

where

$$\hat{k}_{T'} = \frac{\hat{k}_T + \hat{k}_{T-1} + \hat{k}_{T-2} + \hat{k}_{T-3} + \hat{k}_{T-4}}{5} \quad (10)$$

We then apply the *least squares estimation* to find the slope $\hat{\theta}_{T'}$.

First we assume the point of $(\hat{k}_{T'}, T')$ is an origin (0, 0) and simplified the above equation as:

$$\hat{k}_{T+(\Delta t)} = (\Delta t)\hat{\theta}_{T'} \quad (11)$$

Then we obtain

$$\begin{aligned} \text{Min}_{\hat{\theta}_{T'}} \sum_{\Delta t} (\hat{k}_{T+(\Delta t)} - (\Delta t)\hat{\theta}_{T'})^2 &= Q \Rightarrow \\ \frac{\partial Q}{\partial \hat{\theta}_{T'}} &= 2 \sum_{\Delta t} (\hat{k}_{T+(\Delta t)} - (\Delta t)\hat{\theta}_{T'})\Delta t = 0 \Rightarrow \end{aligned}$$

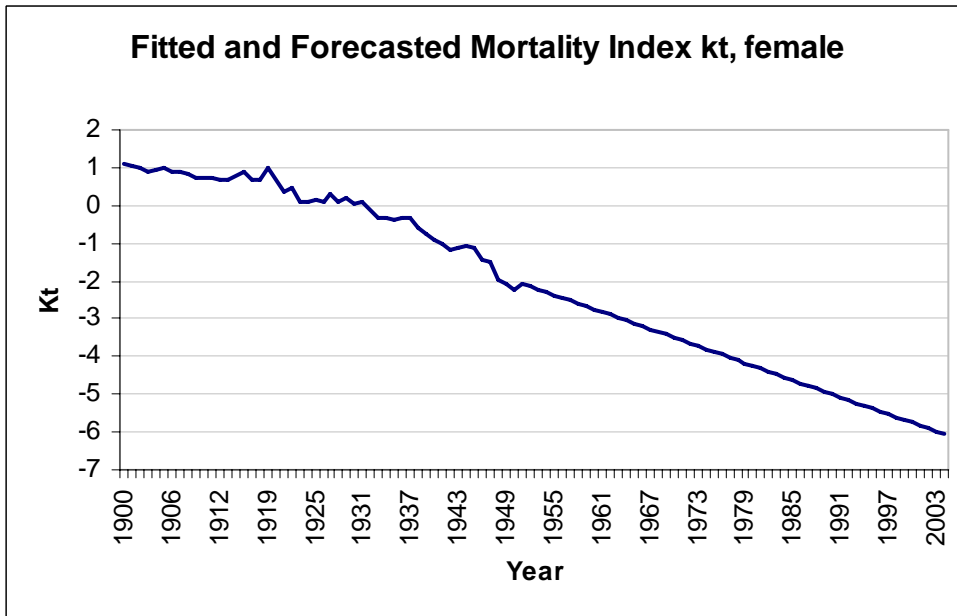
$$\sum_{\Delta t} \hat{k}_{T+(\Delta t)} \Delta t - \hat{\theta}_{T'} \sum_{\Delta t} (\Delta t)^2 = 0 \Rightarrow$$

$$\hat{\theta}_{T'} = \frac{\sum_{\Delta t} \hat{k}_{T+(\Delta t)} \Delta t}{\sum_{\Delta t} (\Delta t)^2} \quad (12)$$

We plug this expression into the equation (9) to make a point forecast for k_t .

Figure 12 illustrates the fitted and forecasted mortality index, obtained with our transformed Lee-Carter approach.

Figure 12: Fitted and Forecasted Mortality Index k_t .



Since the mortality index k_t is the only parameter that needs to be predicted, we could then easily get forecasts for the future mortality rates with the estimation period ends at calendar year T:

$$\ln(m_{x,T+\Delta t}) \approx \hat{a}_x + \hat{b}_x \hat{k}_{T+\Delta t} \quad (13)$$

4.2 Performance and result

We have constructed two experiments using the LC model to generate forecasts. The purpose is to study how the performance of the predictions would have changed if we had changed the length of the estimation period.

In *the first experiment*, the period 1901-2004 is assumed as “future” and its mortality rates will be predicted. In this case, we always using the year 1900 as the end year for our estimation period, while we allow the first year for the estimation period vary as the following: 1875-1900, 1850-1900 and 1800-1900. It means that we are using 25-, 50- and 100- years time interval for estimation. We will then predicting the “future” mortality rates for the next 100 years basing on those different estimation periods. In *the second experiment* we study another four different estimation periods: 1940-1950, 1925-1950, 1900-1950 and 1850-1950, which have the year 1950 as the end year. The prediction period in this case is 1951-2004. According to the procedure described previously, we have a total of seven different estimation periods. For each of these period, we will get a corresponding matrix of mortality rates by estimating and forecasting LC parameters \hat{a}_x , \hat{b}_x , and \hat{k}_t .

In the following part, we present the results obtained when we have different durations of the estimation periods for different spans of the prediction periods. From these results we will try to determine the optimal length of the estimation periods. A limitation for the graphical presentations is that it would be too many graphs if we presented the situation for all the ages. To resolve this dilemma, we present the results for both males and females aged 25-29, 45-49, 65-69 and 85-89 as examples. An extensive overview is given in Figure 13 to 20, and the performance of the predictions is illustrated by comparing the predicted mortality to the observed mortality. The evaluation depends on how the predicted mortality rates resemble the actual mortality rates observed. If the predicted mortality rates correspond to the observed rates, the model is considered to have a good performance. By using this approach, we made additionally two evaluation tables — Table 2 and Table 3, so that the results can be observed more intuitively. Moreover, we format the y-axis scale as standard units for appropriate age in order to facilitate the comparison of the performances.

Forecast 1901-2004

Figure 13: Difference in observed and fitted and forecasted mortality for estimates based on the time series end up with 1900, for ages 25-29, female and male.

($M_{x,t}$ Display in thousands)

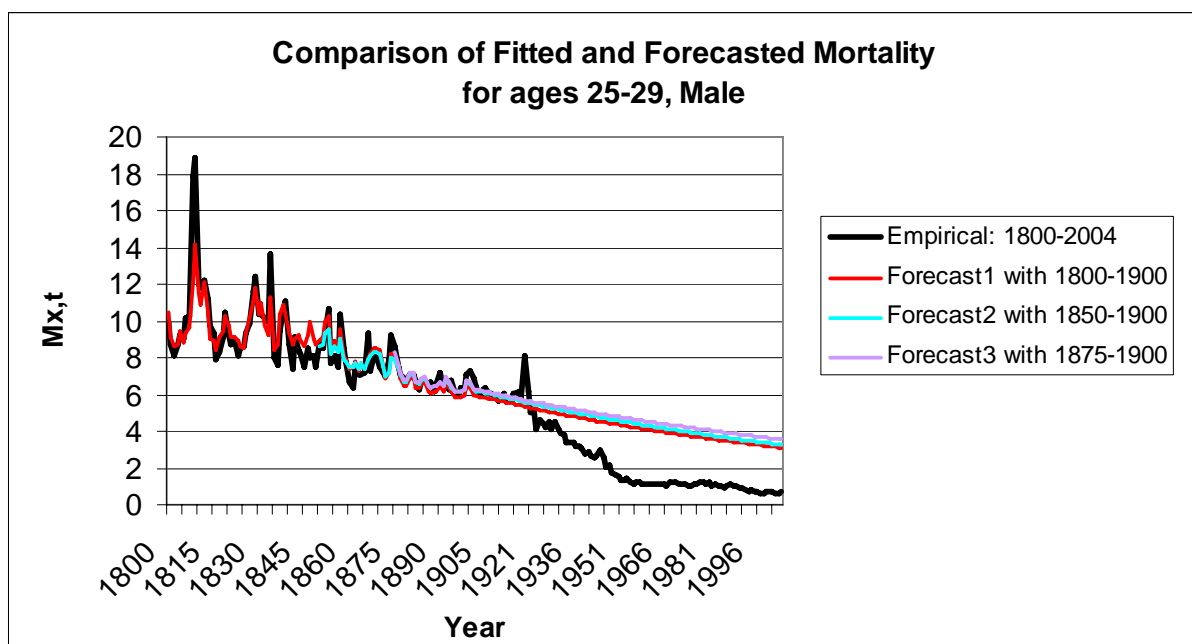
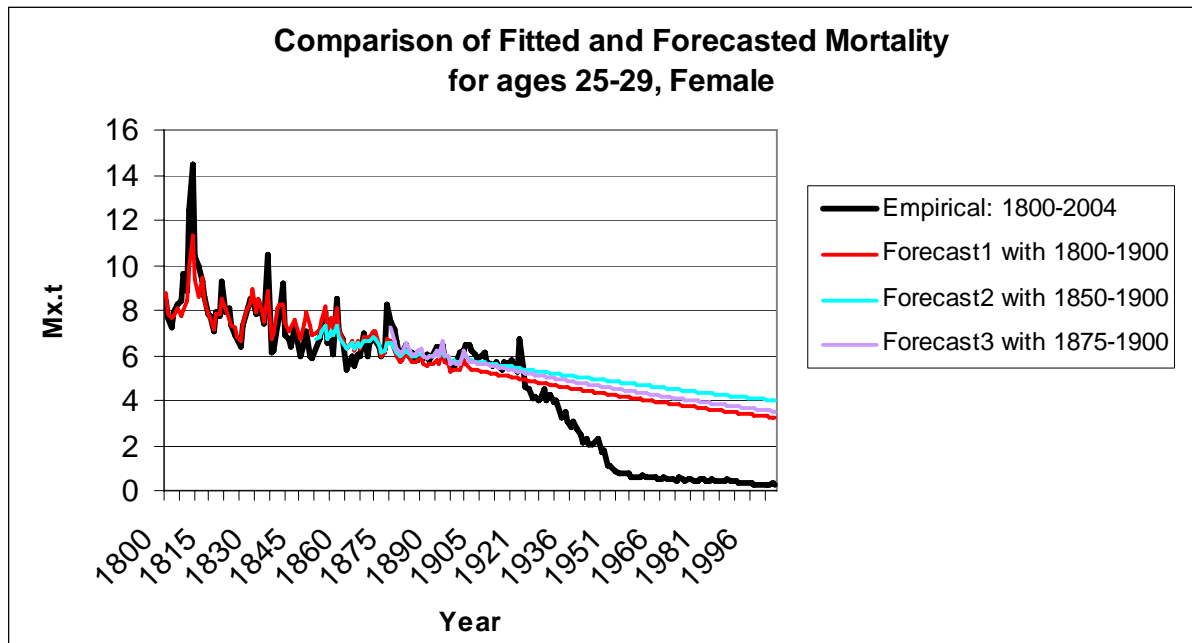


Figure 14: Difference in observed and fitted and forecasted mortality for estimates based on the time series end up with 1900, for ages 45-49, female and male.

($M_{x,t}$ Display in thousands)

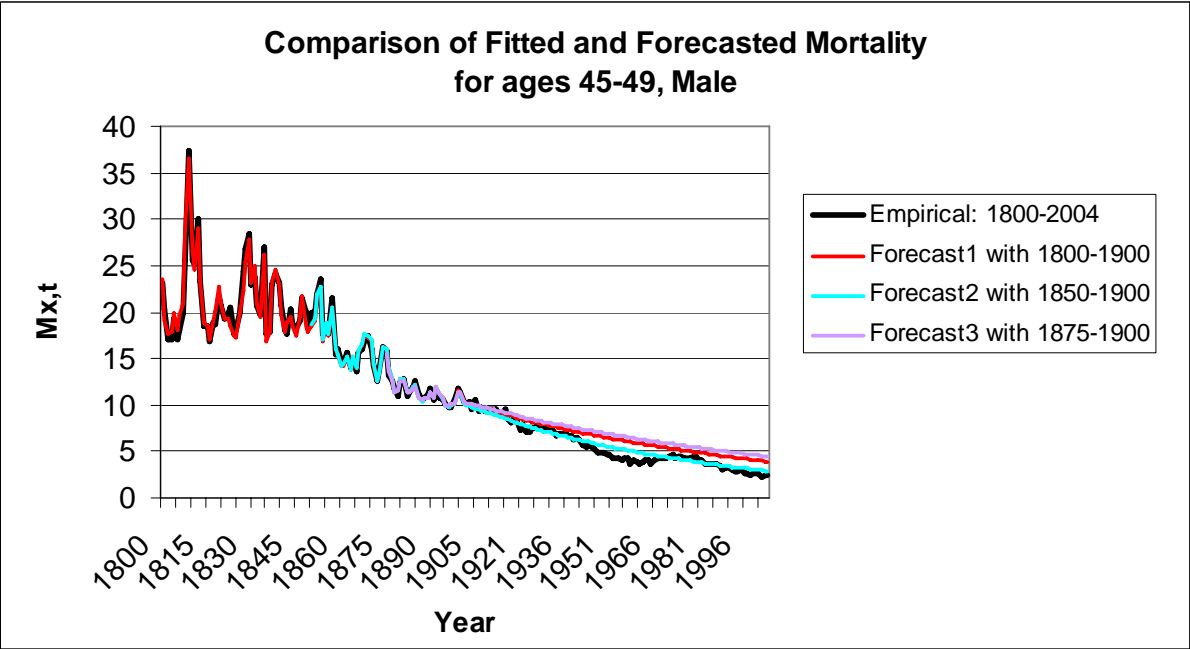
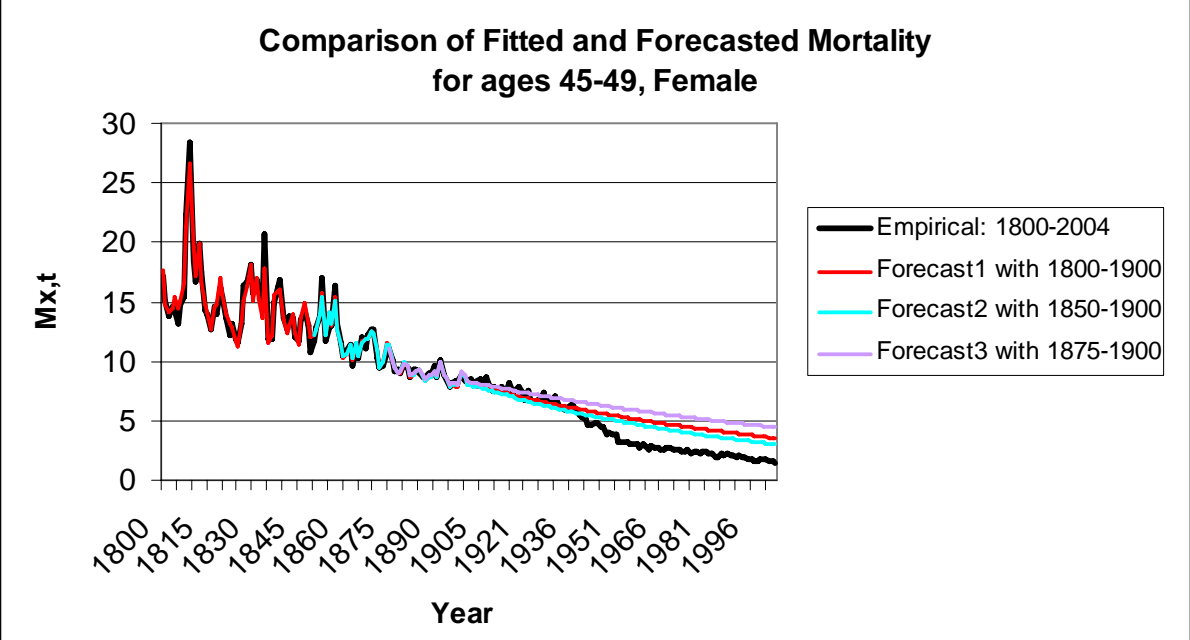


Figure 15: Difference in observed and fitted and forecasted mortality for estimates based on the time series end up with 1900, for ages 65-69, female and male.

($M_{x,t}$ Display in thousands)

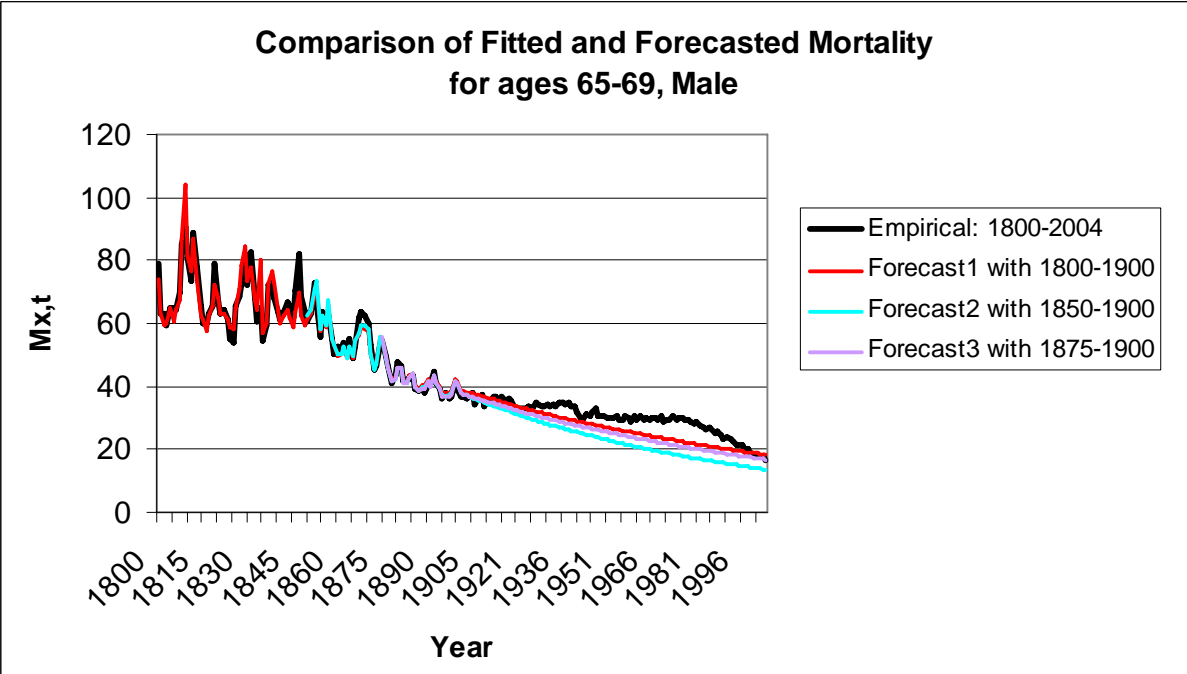
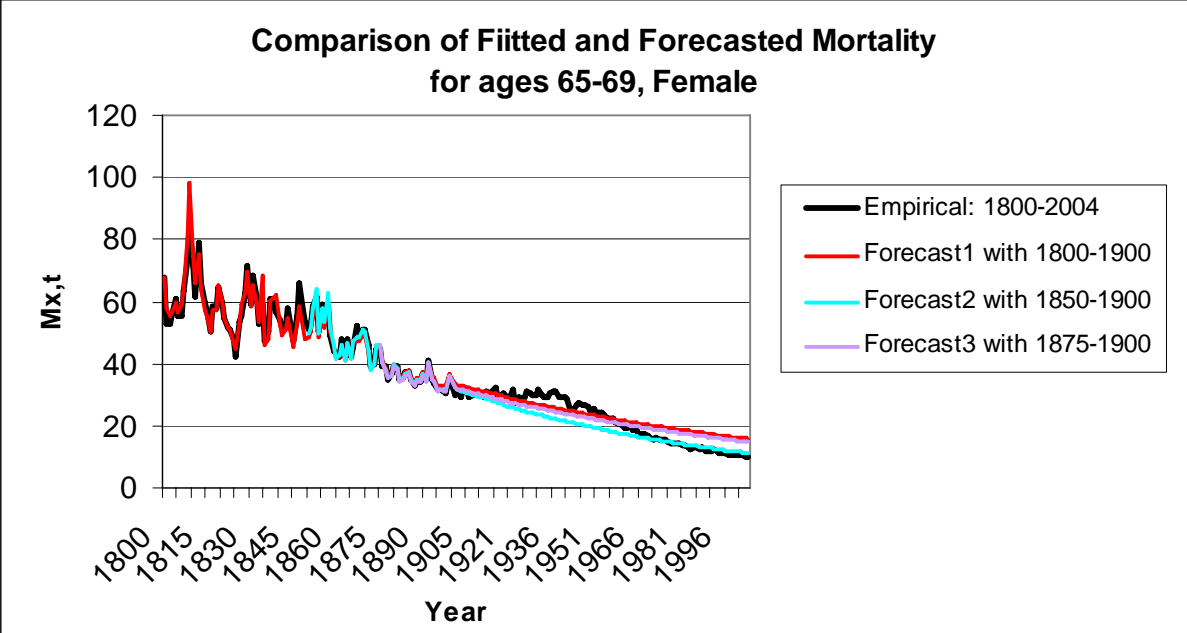
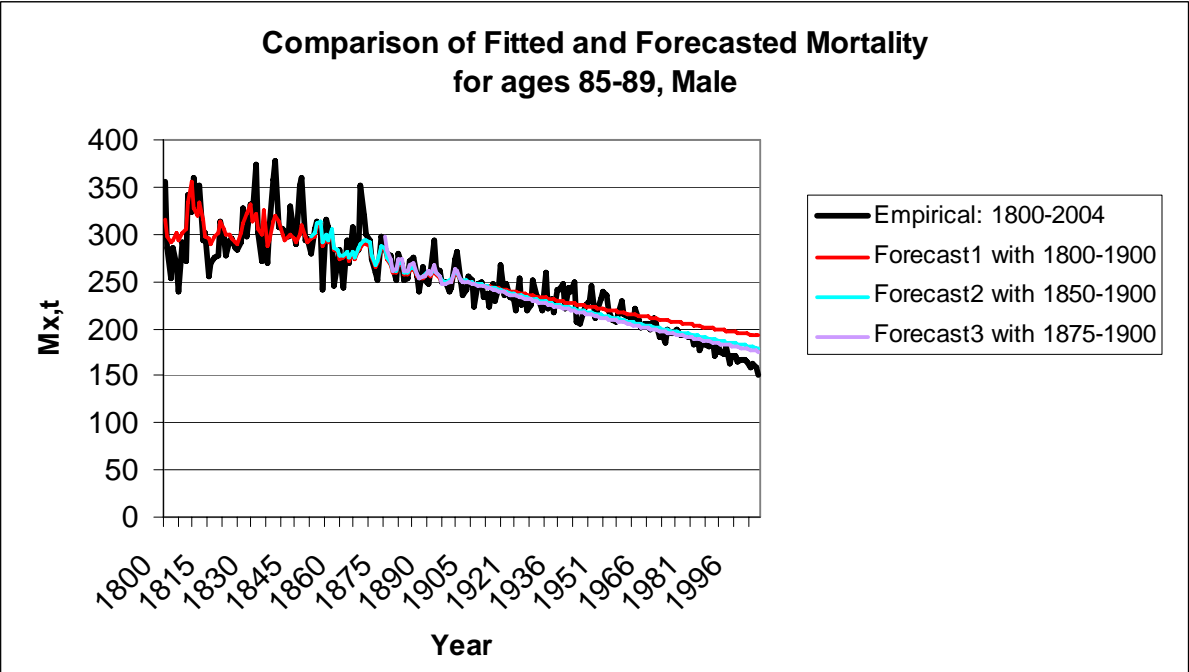
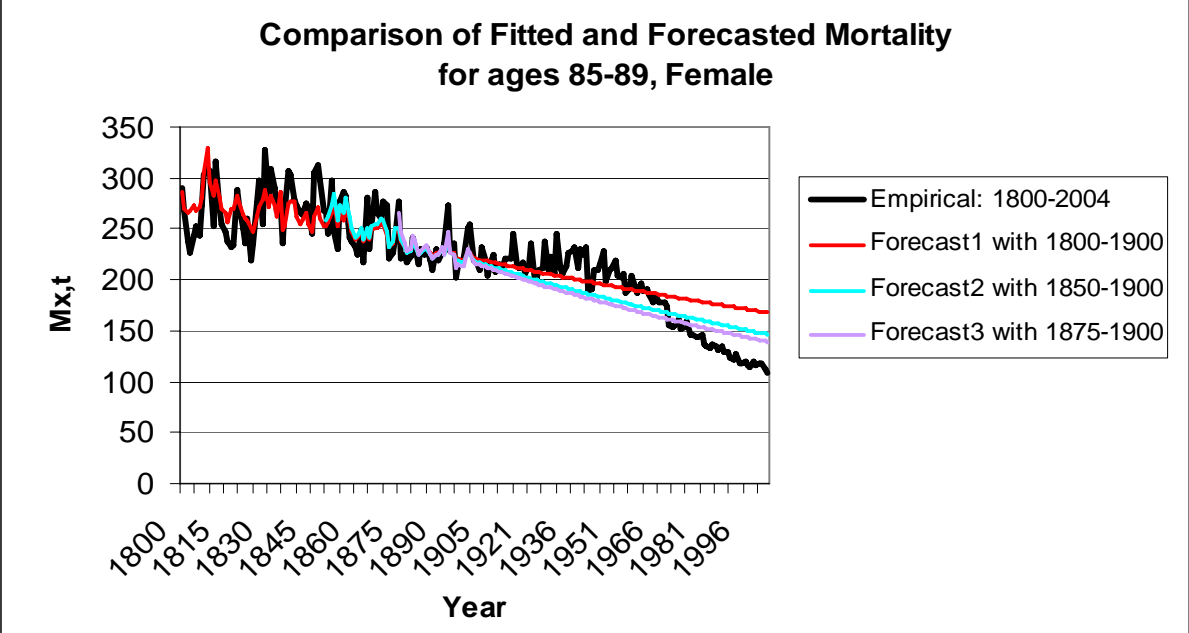


Figure 16: Difference in observed and fitted and forecasted mortality for estimates based on the time series end up with 1900, for ages 85-89, female and male.

($M_{x,t}$ Display in thousands)



Forecast 1951-2004

Figure 17: Difference in observed and fitted and forecasted mortality for estimates based on the time series end up with 1950, for ages 25-29, female and male.

($M_{x,t}$ Display in thousands)

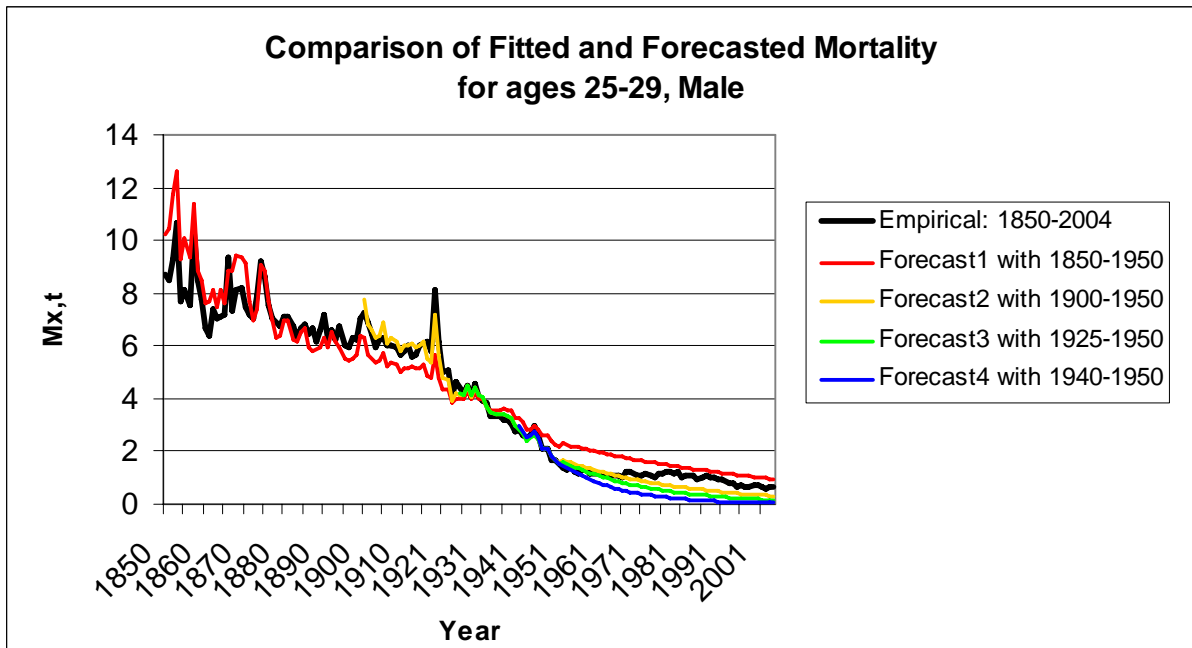
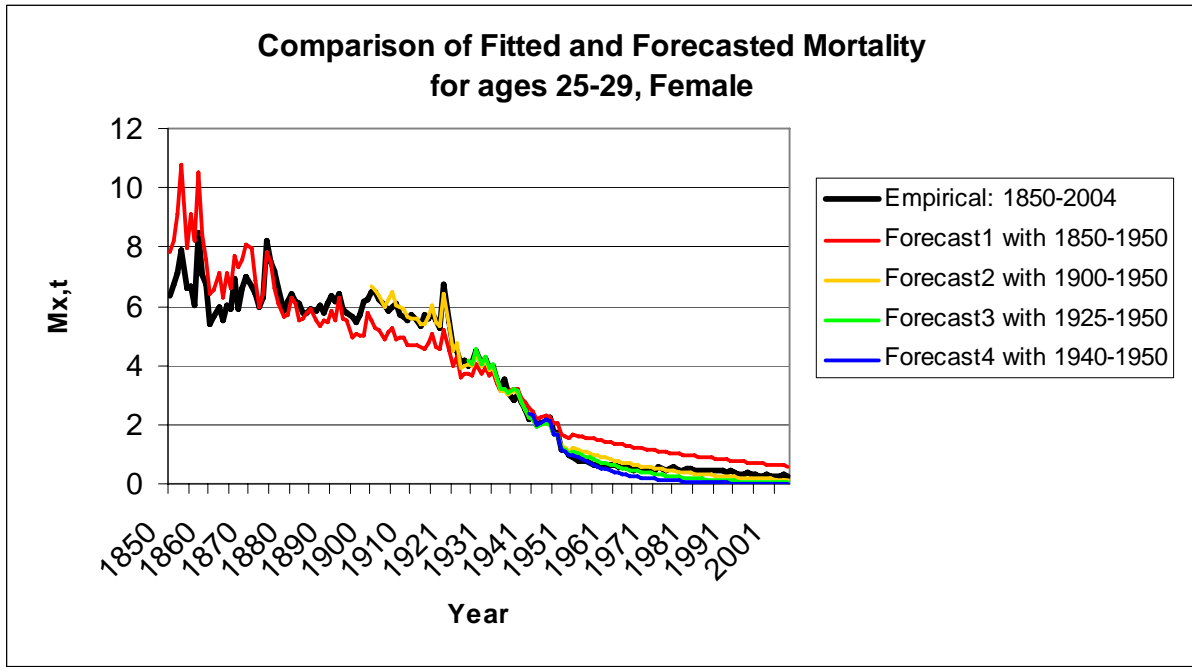


Figure 18: Difference in observed and fitted and forecasted mortality for estimates based on the time series end up with 1950, for ages 45-49, female and male.

($M_{x,t}$ Display in thousands)

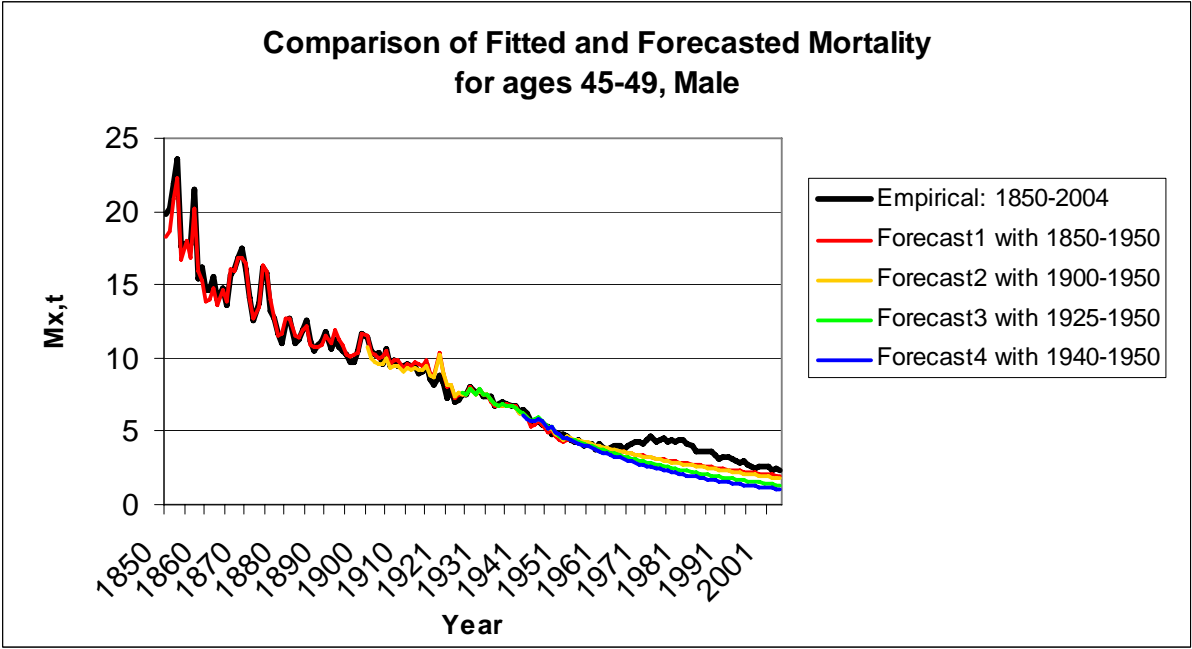
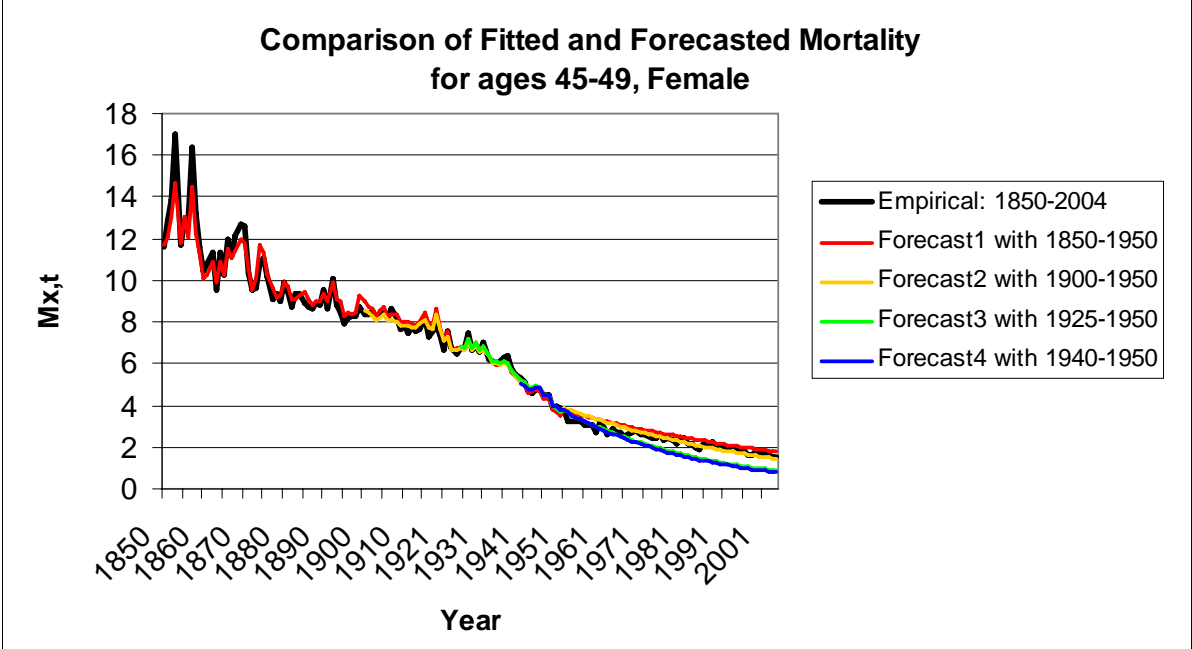


Figure 19: Difference in observed and fitted and forecasted mortality for estimates based on the time series end up with 1950, for ages 65-69, female and male.

($M_{x,t}$ Display in thousands)

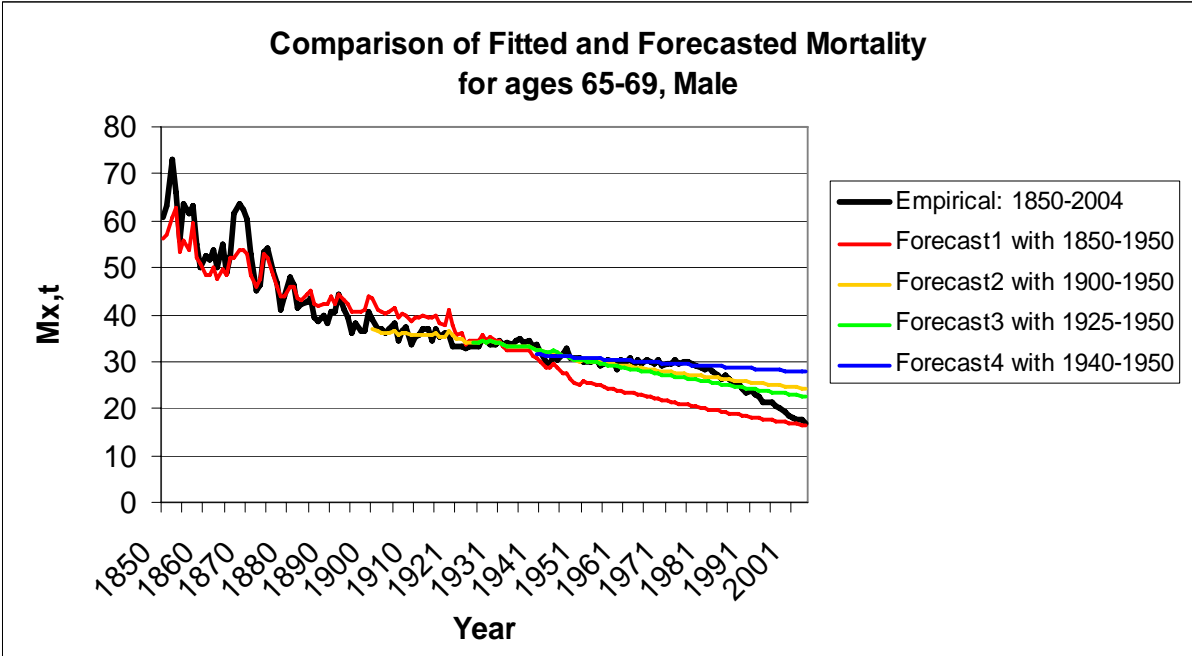
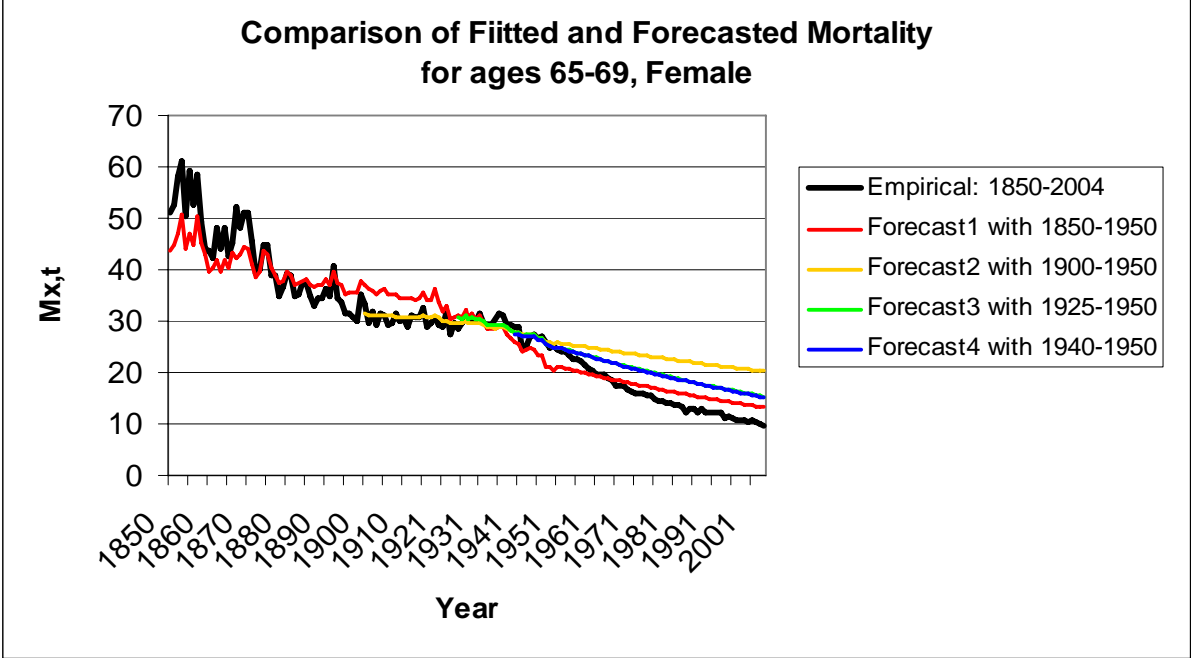


Figure 20: Difference in observed and fitted and forecasted mortality for estimates based on the time series end up with 1950, for ages 85-89, female and male.

($M_{x,t}$ Display in thousands)

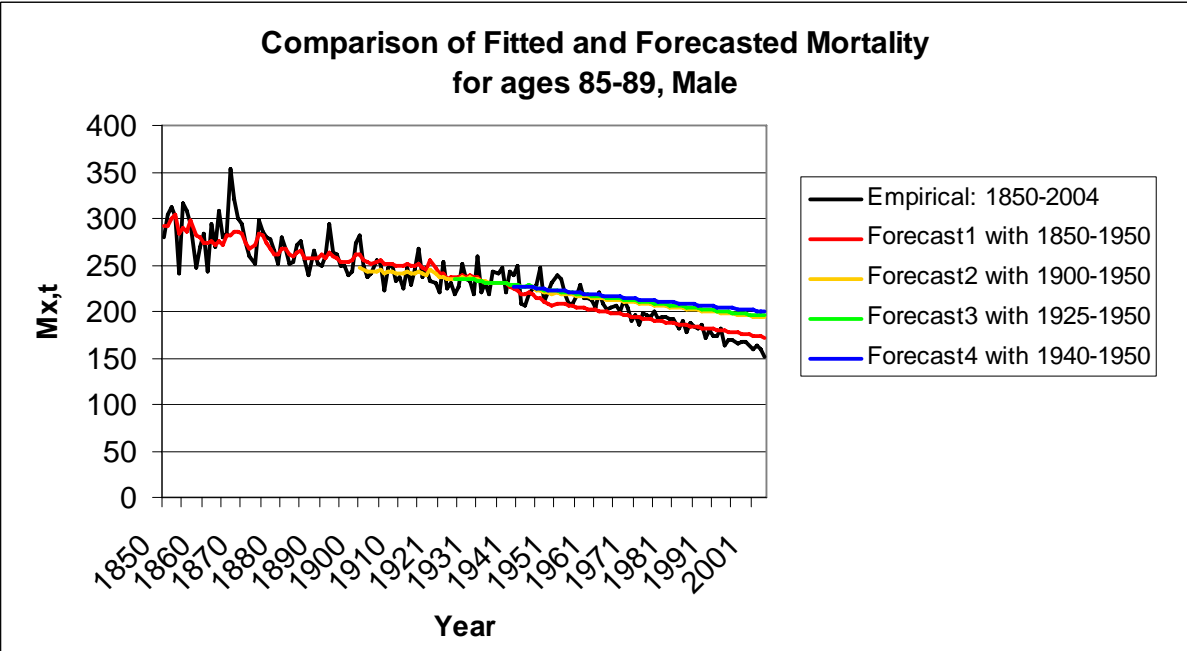
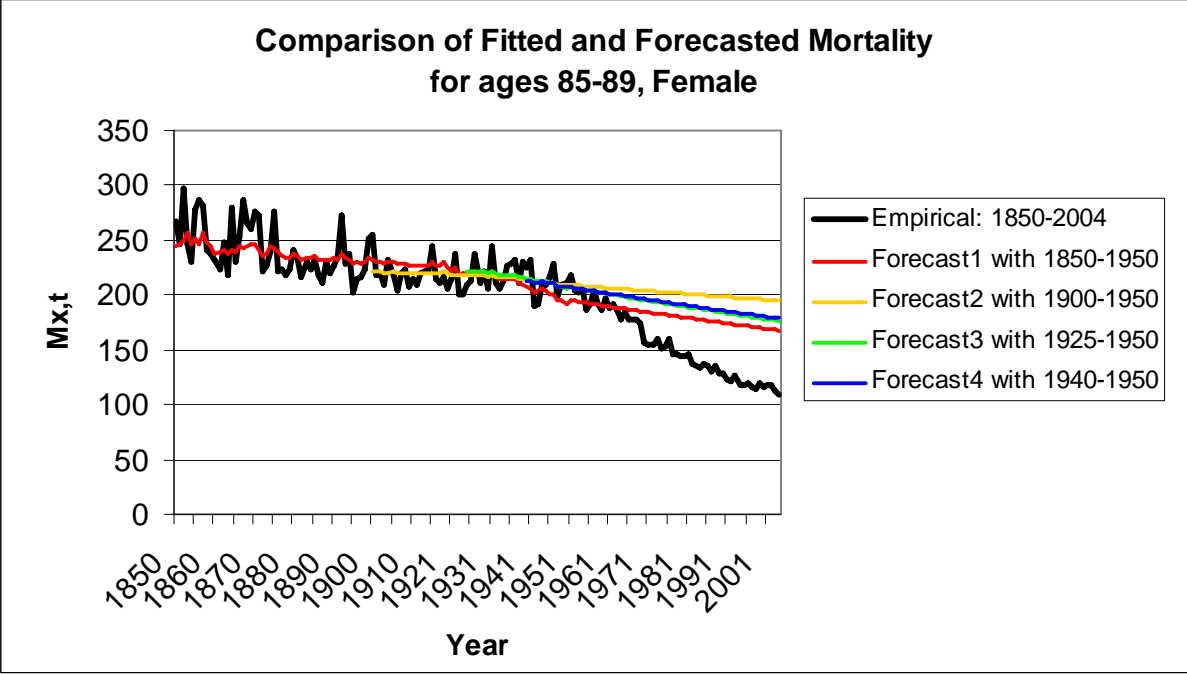


Table 2: Evaluation of the predicted period: 1901-2004.

| Estimation Periods: | Prediction Periods: 1901-2004 | | | | | | | |
|---------------------|-------------------------------|------|-------------|------------------|------------------|-------------|------------|------------------|
| | ages 25-29 | | ages 45-49 | | ages 65-69 | | ages 85-89 | |
| | Female | Male | Female | Male | Female | Male | Female | Male |
| Performance: | Very bad | Bad | Good | Very good | Very good | Good | OK | Very good |
| 1800-1900 | *** | *** | ** | ** | *** | *** | * | ** |
| 1850-1900 | * | ** | *** | *** | ** | * | *** | *** |
| 1875-1900 | ** | * | * | * | *** | ** | ** | *** |

- Notes: 1.) Performance certificate: very bad → bad → ok → good → very good
 2.) "*" is the rating star, for example: *** is the best rating for both "good" and "bad" among three different estimation periods by given fixed age and sex.
 3.) The best performance in the same column has been marked as red.

Table 3: Evaluation of the predicted period: 1951-2004.

| Estimation Periods: | Prediction Periods: 1951-2004 | | | | | | | |
|---------------------|-------------------------------|-------------|------------------|-------------|------------|------|------------|------------------|
| | ages 25-29 | | ages 45-49 | | ages 65-69 | | ages 85-89 | |
| | Female | Male | Female | Male | Female | Male | Female | Male |
| Performance: | Very good | Good | Very good | Good | Ok | Ok | Bad | Very good |
| 1850-1950 | * | ** | **** | **** | **** | ** | **** | **** |
| 1900-1950 | **** | **** | **** | **** | ** | **** | * | *** |
| 1925-1950 | *** | *** | *** | *** | *** | **** | *** | *** |
| 1940-1950 | ** | ** | *** | *** | *** | *** | ** | *** |

- Notes: 1.) Performance certificate: very bad → bad → ok → good → very good
 2.) "*" is the rating star, for example: **** is the best rating of both "good" and "bad" among four different estimation periods by given fixed age and sex.
 3.) The best performance in the same column has been marked as red.

From Figure 13 to Figure 16, we present the comparison of fitted and forecasted mortality among three predictions which based on different spans of estimation periods and the observed mortality. On some occasions, the results of the three predictions vary a lot, while sometimes they are very similar. This characteristic has been shown even more obviously in Figure 17 - 20, in which we compared all four different predictions together with the observed mortality.

In Table 2 and 3, we evaluated the forecast results obtained for the two different prediction periods of 1901-2004 and 1951-2004. Five grades were used to measure the performances: very bad, bad, ok, good and very good. At the same time, rating stars are used to compare the performances within the same category. For example, the evaluation of the forecasted performance for female aged 25-29, is “very bad” using each of the three estimation periods as shown in Table 2. However, a relatively better performance was observed while using the estimation period of 1800-1900, therefore we give it a “3-star” rating.

Table 2 shows that predictions for ages 25-29 are bad for all the estimation periods for both men and women. However, it worked quite well for men and women aged 45-49 and aged 65-69 and men aged 85-89. In general, Table 3 presents better performance than Table 2. The short prediction period is likely to contribute to the lower uncertainty of predictions.

Contrary to Table 2, prediction results in Table 3 are very good for young people but are not so good for old females. In Table 3, we found that estimation period of 1850-1950 and 1900-1950 gave better performances than the periods of 1925-1950 and 1940-1950.

In both table 2 and table 3, we could see those estimation periods of 50 years and longer give better results than shorter estimation periods. However, the 100 years estimation period does not give better results than the 50 years estimation period.

In the most of cases, the performance for males and females is similar for the same time period.

4.3 Discussion: Forecasting with a constant \hat{b}_x

As we have shown in the previous sections, the Lee-Carter model is a useful and appropriate approach to extrapolate historical trends in the level and age distribution of mortality. However, there have been a number of criticisms and discussions about the problems and limitations of the method. One of the problems is that the model assumes a certain pattern of change in the age distribution of mortality, such that the rates of the decline at different ages given by $\hat{b}_x (dk_t/dt)$ always maintain the same ratio to one another over time. However, in reality, the relative speed of decline at different ages may vary. As shown in Figure 21(a) and Figure 22(a), the fitted and forecasted mortality of the recent years become more and more concave. This phenomenon indicates that the value of \hat{b}_x at the concave point is much higher than others. As we have predicted, the mortality index \hat{k}_x is a straight downward line, a large \hat{b}_x value will yield smaller mortality.

We would like to study whether the model is more predictive when we assume that \hat{b}_x is constant. The simplest way is to take a mean value for \hat{b}_x as the constant. In Figure 21(b) and Figure 22(b), two examples are chosen to present how the shape of fitted and forecasted mortality will change when we use a constant value for \hat{b}_x . The first example is the prediction of 1901-2004 using the estimation period of 1800-1900 for females; while the second example is the prediction of 1901-2004 but using the estimation period of 1850-1900 for males. The idea behind this is that the first example has been performed very poorly for female aged 25-29 and the second example has been performed very well for male aged 85-89 (see Table 2 in Section 4.2). We are interested to see how the results change by changing \hat{b}_x to a constant. Figure 23 and Figure 24 display the comparison of the predictions by using a constant and variable \hat{b}_x . The performance with a constant \hat{b}_x in Figure 23 seems slightly

better than the result using a variable \hat{b}_x , presented in section 4.2. On the contrary, the performance in Figure 24 using a constant \hat{b}_x is much worse than before.

In order to compare the performances between the variable \hat{b}_x and constant \hat{b}_x , we also make two tables (Table 4 and 5) for constant \hat{b}_x following the same approach as in Table 2 and 3 in section 4.2. Similarly, the evaluations in Table 4 and 5 also show that the estimation periods of 1850-1900 and 1900-1950 give the best performance than the other periods. If we take age into account, the performance from constant \hat{b}_x for ages 85-89 is very bad for all estimation periods and predicted periods. Comparing with Table 3, Table 5 shows a poor performance result for ages 25-29 for both females and males. However, the result is quite good for ages 45-49 and ages 65-69 (except the males). One could see a tendency that the performances for shorter estimation periods are better compared to the longer estimation periods, when a variable \hat{b}_x is used. Generally, there is absolutely no guarantee that the extend method with a constant \hat{b}_x would perform better than a variable \hat{b}_x .

Figure 21: Fitted and Forecasted mortalities with different \hat{b}_x with estimates based on time series 1800-1900 for females all ages, on a logarithmic scale.

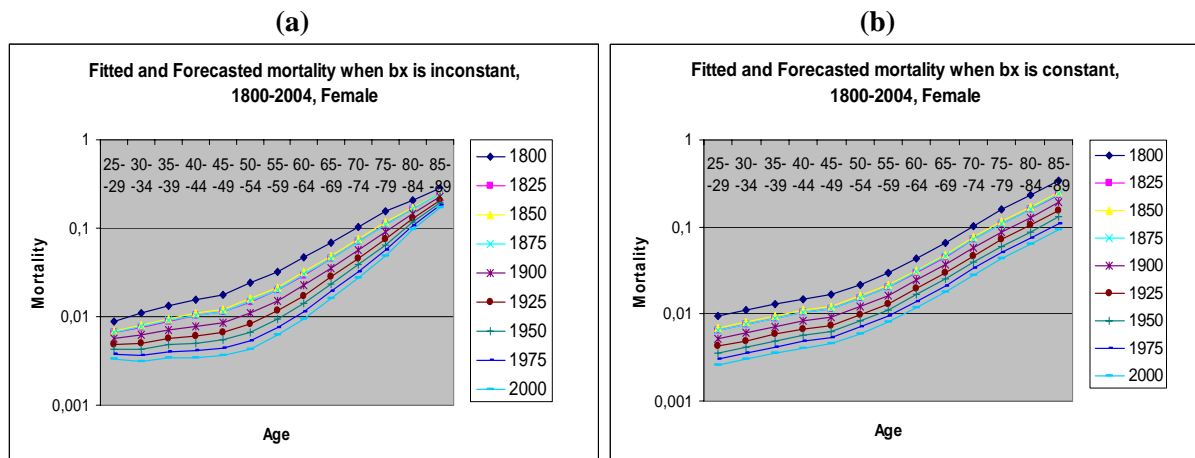


Figure 22: Fitted and Forecasted mortalities with different \hat{b}_x , with estimates based on time series 1850-1900 for males all ages, on a logarithmic scale.

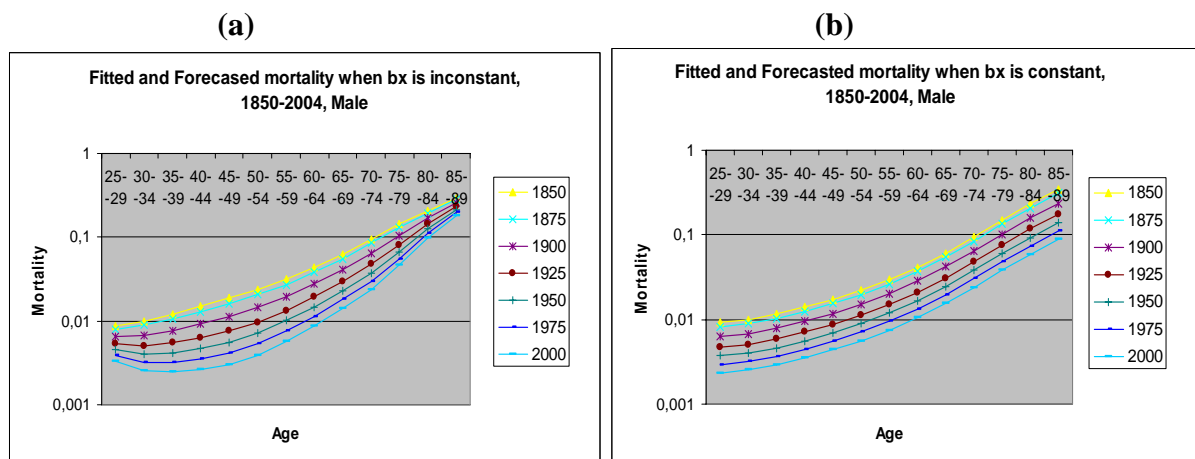


Figure 23: Compare the prediction results with constant and inconstant \hat{b}_x for ages 25-29 female and estimates based on the time series 1800-1900.

($M_{x,t}$ Display in thousands)

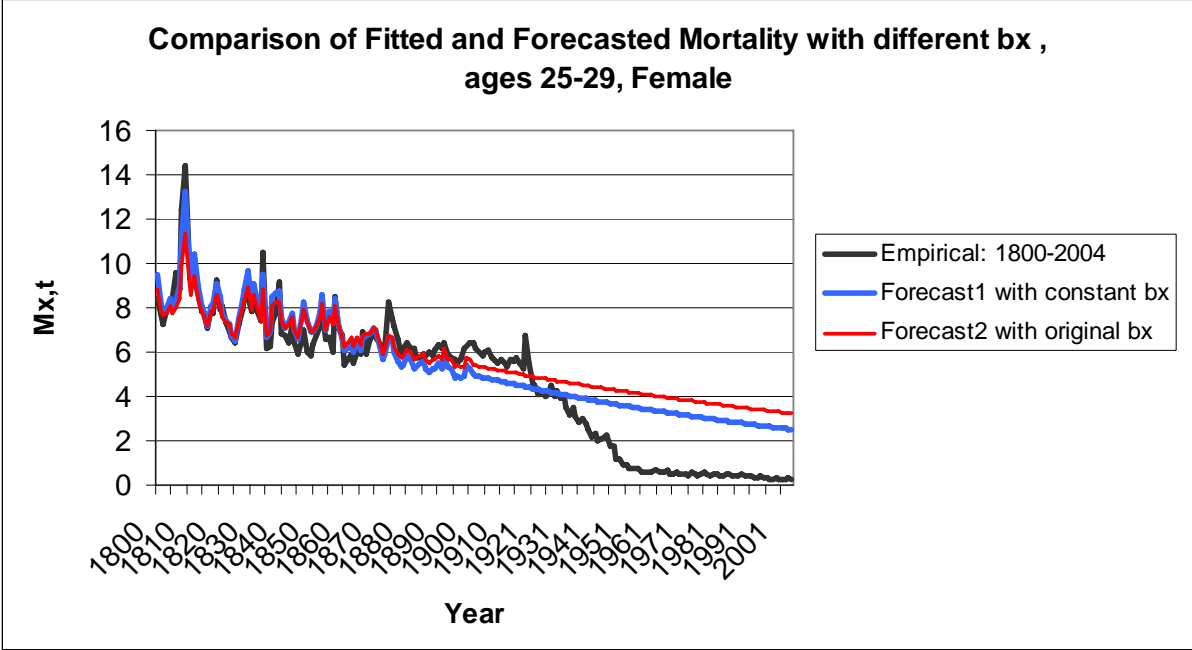


Figure 24: Compare the prediction results with constant and variable \hat{b}_x for males ages 85-89 and estimates based on the time series 1850-1900.

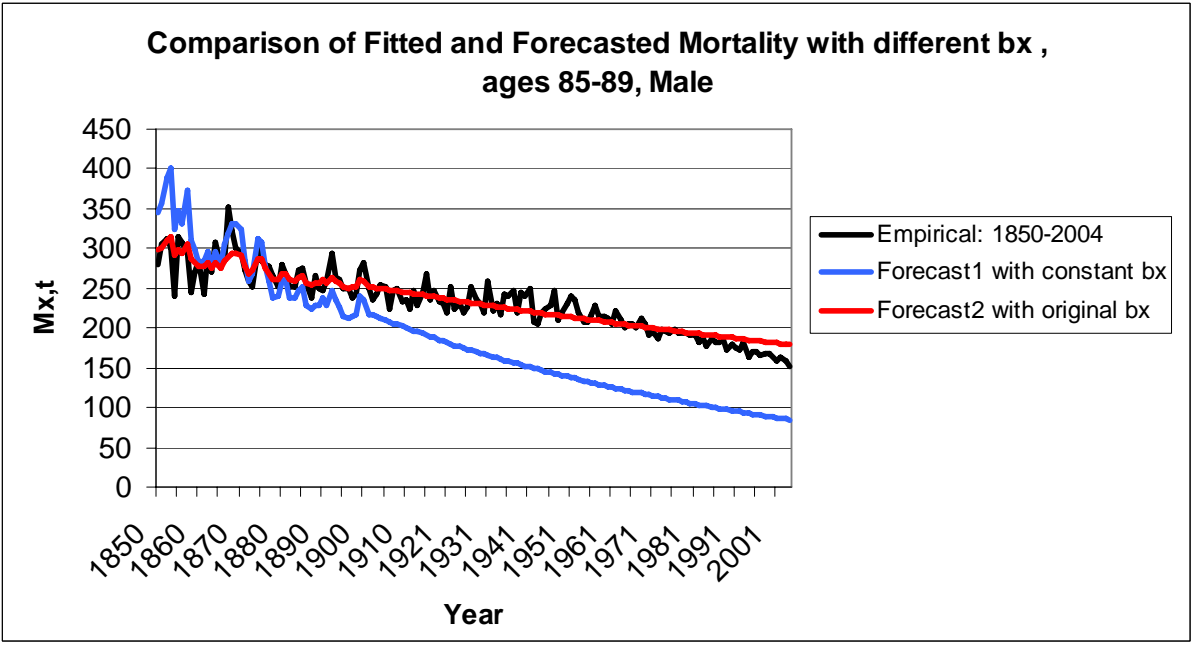


Table 4: Evaluation of the predicted period: 1901-2004 with constant \hat{b}_x .

| Estimation Periods: | Prediction Periods: 1901-2004 | | | | | | | |
|---------------------|-------------------------------|------|------------|------|------------|------|------------|----------|
| | ages 25-29 | | ages 45-49 | | ages 65-69 | | ages 85-89 | |
| | Female | Male | Female | Male | Female | Male | Female | Male |
| Performance: | Bad | Bad | Good | Good | Good | Good | Very bad | Very Bad |
| 1800-1900 | *** | ** | ** | ** | ** | *** | ** | ** |
| 1850-1900 | *** | *** | *** | *** | *** | ** | ** | ** |
| 1875-1900 | ** | * | ** | ** | ** | *** | *** | *** |

- Notes: 1.) Performance certificate: very bad → bad → ok → good → very good
 2.) "*" is the rating star, for example: *** is the best rating for both "good" and "bad" among three different estimation periods by given fixed age and sex.
 3.) The best performance in the same column has been marked as blue.

Table 5: Evaluation of the predicted period: 1951-2004 with constant b_x .

| Estimation Periods: | Prediction Periods: 1951-2004 | | | | | | | |
|---------------------|-------------------------------|------|------------|-----------|------------|------|------------|----------|
| | ages 25-29 | | ages 45-49 | | ages 65-69 | | ages 85-89 | |
| | Female | Male | Female | Male | Female | Male | Female | Male |
| Performance: | Very bad | OK | Very good | Very good | Very good | bad | Very bad | Very bad |
| 1850-1950 | * | * | ** | *** | **** | ** | * | ** |
| 1900-1950 | ** | ** | **** | **** | *** | ** | **** | **** |
| 1925-1950 | *** | *** | **** | ** | ** | *** | *** | **** |
| 1940-1950 | **** | **** | ** | * | ** | **** | ** | *** |

- Notes: 1.) Performance certificate: very bad → bad → ok → good → very good
 2.) "*" is the rating star, for example: *** is the best rating for both "good" and "bad" among three different estimation periods by given fixed age and sex.
 3.) The best performance in the same column has been marked as blue.

5. Conclusion

We have attempted to identify the common trend of mortality change by fitting a standard Lee-Carter model to Swedish historical population data. The model's parameters were estimated with the Singular Value Decomposition (SVD). The computations and comparisons were completed using the data from time series of 1860-2004 and three selected sub-samples of 1900-2004, 1950-2004 and 1980-2004. The estimates for the components from data in 1860-2004 and 1900-2004 are most alike whereas 1950-2004 and 1980-2004 are similar to each other. This is due to the growing importance of medical care in mortality decline in the later half of the 20th century.

To study the efficiency of the Lee-Carter model, the residual term on a logarithmic scale has been examined as well. The cohort effect does not appear very significant, and most of the residual terms show a similar lack of systematic pattern. However, the residuals still cannot be described as random as the marked clusters do occur. The occurrence is probably a result of the infectious diseases, which were the leading cause of death during certain periods. Insufficient weight for young people at ages with smaller numbers of deaths also contributed to this observation.

Whether any empirical pattern will continue in the future is of course the subject of almost every forecasting work. In this paper, we simulated two prediction periods: 1901-2004 and 1951-2004 as the "future" periods. Three different estimation periods of 1800-1900, 1850-1900 and 1875-1900 were used to predict the mortality rates for 1901-2004. Four estimation periods of 1850-1950, 1900-1950, 1925-1950 and 1940-1950 were used to predict the mortality rates for 1951-2004. Our purpose is to study how the performance of the predictions would change if we change the span of the estimation period. The forecasting of the mortality index \hat{k}_t is very simple: extrapolate from a straight line drawn through the point \hat{k}_T , which is the average value of the last five points up to period T, while all other \hat{k}_t points are ignored. The results showed that the estimation periods of 1850-1900 and 1900-1950 yield the best forecasting performances for prediction series of 1901-2004 and 1951-2004, respectively. Not surprisingly, the prediction with short estimation period like 1875-1900 and 1940-1950 did not work well.

In addition, we have also discussed whether the model would perform better when we assume \hat{b}_x as a constant in section 4.3. The value of \hat{b}_x is fixed using the mean value. The results showed that the model does perform better than the forecasting using a variable \hat{b}_x in certain estimation periods and age groups. However, poorer performance was also observed using other estimation periods and age groups. The improvement of the performance by using a constant \hat{b}_x is not reliable, as there is absolutely no guarantee that the extend method with a constant \hat{b}_x would perform better than a variable \hat{b}_x . Nevertheless, similar to the previous forecasts, the prediction with a constant \hat{b}_x also demonstrated that the estimation period of a 50-year time interval, e.g. 1850-1900 and 1900-1950 are the optimal span for the estimation period. This indicates that the selection of an appropriate estimation period is important for forecasting mortality. Moreover, males and females almost have similar performances for the same time period.

Since Lee and Carter published this model for long-run forecasts of the level and age pattern of mortality in 1992, there have been a number of extensions of the method, including the development of coherent forecasts by sex and by race, and forecasts for regions comprising a national system and so on. This paper is only an initial investigation into the attributes of an original Lee-Carter model used for estimation and forecasting mortality. Hence there is still a room left for an after work thinking, for instance, \hat{b}_x may assumed to be a function of age factor x for receiving a better performance.

Moreover, by studying Swedish historical data, we found that infectious diseases and medical factors are always very important contributors to the trend of mortality. Therefore, in practice, the model could be combined with the addition of expert opinions to estimate future trends, such as the opinions about medical developments, environments and new diseases.

Finally, we experienced the user-friendly application of the model as we have completed all the calculations using functions in Microsoft Excel.

Appendix:

I. Proof Lee-Carter Approximation

Given a matrix $\ln_{M_{x,t}} = a_x + b_x k_t$ with two conditions of $\sum_t k_t = 0$ and $\sum_x b_x^2 = 1$. Parameters a_x , b_x and k_t are determined by minimizing the Q (a, b, k), where

$$Q(a, b, k) = \sum_{x,t} (\ln_{M_{x,t}} - a_x - b_x k_t)^2$$

Let $\frac{\partial Q}{\partial a_x} = \frac{\partial Q}{\partial b_x} = \frac{\partial Q}{\partial k_t} = 0$, and we can get

$$\frac{\partial Q}{\partial a_x} = 2 \sum_t (\ln_{M_{x,t}} - a_x - b_x k_t) = 0 \quad (1)$$

$$\frac{\partial Q}{\partial b_x} = 2 \sum_t (\ln_{M_{x,t}} - a_x - b_x k_t) k_t = 0 \quad (2)$$

$$\frac{\partial Q}{\partial k_t} = 2 \sum_x (\ln_{M_{x,t}} - a_x - b_x k_t) b_x = 0 \quad (3)$$

With the help of the equation (1) and the condition 2, we can get a_x directly,

$$\sum_t a_x = \sum_t \ln_{M_{x,t}} - \sum_t b_x k_t = \sum_t \log_{M_{x,t}} \Rightarrow$$

$$a_x = \frac{1}{T} \sum_t \ln_{M_{x,t}}$$

Now let us set a new matrix $Z_{x,t} = \ln_{x,t} - a_x$, then we can get the following equations for b_x and k_t :

$$\sum_t Z_{x,t} k_t = \sum_t (\ln_{x,t} - a_x) k_t = \sum_t (b_x k_t) k_t = b_x \sum_t k_t^2$$

$$\sum_x Z_{x,t} b_x = \sum_x (\ln_{x,t} - a_x) b_x = \sum_x (b_x k_t) b_x = k_t \sum_x b_x^2 = k_t, \text{ (since } \sum_x b_x^2 = 1)$$

By considering of a vector form with $\sum_t k_t^2 = \beta$, we rewrite two above equations as the following:

$Zk = \beta b$ and $Z'b = k$ (where $Z' = Z^T$), which gives:

$$(ZZ') \mathbf{b} = Z\mathbf{k} = \beta \mathbf{b}$$

Therefore, \mathbf{b} is the eigenvector of matrix (ZZ') with the eigenvalue β , under the conditions of $(\mathbf{b}'\mathbf{b}) = \sum_x b_x^2 = 1$ and $(\mathbf{k}'\mathbf{k}) = \sum_t k_t^2 = \beta$.

By using equations: $Z_{x,t} = \ln_{x,t} - a_x$, $\sum_x b_x Z_{x,t} = k_t$, $\sum_x b_x^2 = 1$ and $\sum_t k_t^2 = \beta$, we can simplify $Q(\mathbf{a}, \mathbf{b}, \mathbf{k})$ as the following:

$$\begin{aligned} Q(\mathbf{a}, \mathbf{b}, \mathbf{k}) &= \sum_{x,t} (\ln_{M_{x,t}} - a_x - b_x k_t)^2 \\ &= \sum_{x,t} (Z_{x,t} - b_x k_t)^2 \\ &= \sum_{x,t} Z_{x,t}^2 + \sum_{x,t} b_x^2 k_t^2 - 2 \sum_t k_t \left(\sum_x b_x Z_{x,t} \right) \\ &= \sum_{x,t} Z_{x,t}^2 + 1 \times \beta - 2 \sum_t k_t^2 \quad (\text{since}) \\ &= \sum_{x,t} Z_{x,t}^2 + \beta - 2\beta \\ &= \sum_{x,t} Z_{x,t}^2 - \beta \end{aligned}$$

In order to minimize Q , we have to maximize β . It also gives that \mathbf{b} will be eigenvector of (ZZ') with the maximal eigenvalue β , and $\mathbf{k} = Z'\mathbf{b}$.

II. How to add **Biplot** function in your Ms Excel®?

Download the “Biplot” from

http://filebox.vt.edu/artsci/stats/vining/keying/biplot_final.zip) or from other websites.

To install *biplot01.xla* do the following:

- Save the file “biplot01” on your hard drive (for example: D:\software\biplot01.exl).
- Start MS Excel®
- In Excel go to **Tools|Add-Ins...** and invoke the *Add-ins* dialog
- In the Add-Ins dialog click **Browse** button and select the current *biplot01.xla* from where you have saved on your hard drive (eg: D:\software\biplot01.exl).
- After pressing **OK** button, the menu item "biplot" should appear on your Excel main menu.
- Refer to biplot help for instructions about specific functions.

References

- [1] Booth, Heather; Maindonald, John and Smith, Len. (2002). *Age-Time Interactions in Mortality Projection: Applying Lee-Carter to Australia.*
- [2] Broekhoven, Henk Van. *Market Value of Liabilities Mortality Risk: A Practical Model.*
- [3] Carter, Lawrence R. and Prskawetz, Alexia (2001). *Examining Structural Shifts in Mortality Using the Lee-Carter Method.*
- [4] Cutler, David M. and Meara, Ellen (September 2001). *Changes in the Age Distribution of Mortality over the 20th Century.*
- [5] Gleijer, Dana; Lundström, Hans and Wilmoth, John. (February 2006). *About Mortality Data for Sweden.*
- [6] Human Mortality Database (www.mortality.org).
- [7] King, Gary and Girosi, Federico (March 2005). *A Reassessment of the Lee-Carter Mortality Forecasting Method.*
- [8] Koissi, Marie Claire; Shapiro, Arnold and Högnäs, Göran. *Fitting and Forecasting Mortality Rates for Nordic Countries Using the Lee-Carter method.*
- [9] Kruger, Daniel J. and Nesse, Randolph M. (2004). *Sexual Selection and the Male: Female Mortality Ratio.*
- [10] Leach, Sonia. *Singular Value Decomposition – A Primer*
- [11] Lee, Ronald. *The Lee-Carter Method for Forecasting Mortality, With Various Extensions and Applications.*
- [12] Lee, Ronald and Miller, Timothy (November 2000). *Evaluating the Performance of Lee-Carter Mortality Forecasts.*

- [13] Li, Siu-Hang and Chan, Wai-Sum (January 2005). *The Lee-Carter Model for Forecasting Mortality Revisited*.
- [14] Lipkovich, Ilya and Smith, Eric P. (June 2002). *Biplot and Singular Value Decomposition Macros for Excel*.
- [15] Lundström, Hans and Qvist, Jan (Dec 2002). *Mortality Forecasting and Trend Shifts: an Application of the Lee-Carter Model to Swedish Mortality Data*.
- [16] Maxwell, Andrea. *The Future of the Social Security Pension Provision of Trinidad and Tobago Using Lee-Carter Forecasts*.
- [17] Naoki Sunamoto, FIAJ (September 2005). *Cohort Effect Structure in the Lee-Carter Residual Term*.
- [18] Noymen, Andrew and Garenne, Michel. *The 1918 Influenza Epidemic's Effects on Sex Differentials in Mortality in the United States*.
- [19] Olsén, Jörgen (July 2005) *Modeller och Projektioner för Dödlighetsintensitet – en anpassning till svensk populationsdata 1970-2004*.
- [20] Ozeki, Masakazu (2004). *Applying Mortality Models to Japan*.
- [21] Perls, Thomas T. and Fretts Ruth C. *Why Women Live Longer than Men*.
- [22] Wohlfart, Peter (2006). *Mortality Predictions for Longevity Analysis and Annuity Valuation*.