SFB 649 Discussion Paper 2009-009

Stochastic Population Forecast for Germany and its Consequence for the German Pension System

Wolfgang Härdle*
Alena Mysickova*

*Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".

http://sfb649.wiwi.hu-berlin.de
ISSN 1860-5664

SFB 649, Humboldt-Universität zu Berlin
Spandauer Straße 1, D-10178 Berlin

SFB 649 ECONOMIC RISK BERLIN

Stochastic Population Forecast for Germany and its Consequence for the German Pension System*

Wolfgang Karl Härdle[†] Alena Myšičková[‡]

31st January 2009

Population forecasts are crucial for many social, political and economic decisions. Official population projections rely in general on deterministic models which use different scenarios for future vital rates to indicate uncertainty. However, this technique shows substantial weak points such as assuming absolute correlations between the demographic components. In this paper, we argue that a stochastic projection alternative, with no a priori assumptions provides point forecasts and probabilistic prediction intervals for demographic parameters in addition. Age-sex specific population forecast for Germany is derived through a stochastic population renewal process using forecasts of mortality, fertility and migration. Time series models with demographic restrictions are used to describe immigration, emigration and time varying indices of mortality and fertility rates. These models are then used in the simulation of future vital rates to obtain age-specific population forecast using the cohort-component method. The consequence for the German pension system is discussed. To maintain the actual average pension level the premium rate of the present system rises at least by 50% as the old-age ratio nearly doubles by 2040.

Keywords: Demographic Forecasting; Population Projection; Stochastic Demography
JEL classification: J11; J13; C53; C22

---

*We gratefully acknowledge financial support by the Deutsche Forschungsgemeinschaft and the Sonderforschungsbereich 649 "Ökonomisches Risiko".

[†]Institute of Statistics and Econometrics of Humboldt-Universität zu Berlin and CASE - Center for Applied Statistics and Economics.

[‡]Corresponding author. Research associate at the Institute for Statistics and Econometrics of Humboldt-Universität zu Berlin and CASE-Center for Applied Statistics and Economics, Spandauer Straße 1, 10178 Berlin, Germany. Email: alena.mysickova@wiwi.hu-berlin.de.

1 Introduction

Population forecasts are crucial for many social, political and economic decisions, such as the financing of pension and health systems, labour market development or education planning. Conventional population projections rely on deterministic models which mainly use three different scenarios for future vital rates to indicate uncertainty. The demographic factors (such as population size and its age structure) in each scenario are obtained by an extrapolation of the actual values using the assumed scenario parameters which are pre-specified by the demographers. The forecast interval of the demographic factors is then defined by the combination of the scenarios. However, this technique faces substantial difficulties. Firstly, in the deterministic models one cannot assign any access probability to the various scenarios, so that the resulting population size occurs with the same probability in its whole interval. Secondly, the deterministic models assume a perfect but unrealistic correlation between the demographic components which in consequence leads to wide forecast intervals. Furthermore, the forecasted ranges for age group size, population size and age ratios are not probabilistically consistent with one another.

In the recent past, demographers and statisticians developed alternative methods which allowed for stochastic population forecasts, aimed at calculating confidence intervals for every demographic factor of interest and assigning access probabilities for diverse scenarios. Alho and Spencer (1985), Pflaumer (1988) and Lee and Tuljapurkar (1994) provided a stochastic population projection for the USA, Keilman (2002) showed a probabilistic population forecast by the case of Norway.
A detailed discussion of the deterministic and stochastic methods and a comparison of both methods can be found by Lee (1998) and Babel (2007). A detailed summarization of statistical methods applied on forecasting demographic variables can be found by Girosi and King (2008).

The stochastic approach was firstly applied on German data by Lutz and Scherbov (1998), their method combines assumptions of experts on the demographic parameters with an estimation of the probability distribution using a simulation method. They estimate the Total Fertility Rate (TFR), net migration and life expectancy by a normal distribution. Lipps and Betz (2005) provide a separate stochastic forecast for former East and West Germany using time series models. They apply the Lee-Carter model for mortality, model the TFR by a Random-Walk process and choose a standard autoregressive process to describe the migration level. Babel (2007) compares deterministic and stochastic models for population forecasting applied on East and West German data. In his work, he applies an analogue model to Lipps and Betz (2005) for fertility rates, forecast mortality via panel data procedures and applies a modified Lee-Carter approach on migration.

Our work join these recent developments and perform a stochastic population forecast for up-to-date German data. We apply the classical Lee-Carter method to forecast mortality and modify it for fertility projection. Net migration is modelled as a difference of the immigration and emigration process, we estimate the distribution of migrants age by a nonparametric technique. The forecasts of vital rates are combined in the population renewal process using the cohort-component method to estimate the population size and its structure. Furthermore, probabilistic confidence intervals and the distribution of forecasts are generated by scenario simulation. We compare our forecast with the deterministic population forecast of the Federal Statistical Office. To show the consequence of the shrinking and ageing population for the German pension system, on one hand the minimal premium rate, required to maintain the actual pension level is estimated with its prediction intervals. On the other hand, the future pension level with its confidence intervals is estimated in case the actual premium rate will be fixed.

The paper is structure as follows. In Section 2 we present the modelling of the age-specific mortality where we use the known method developed by Lee and Carter (1992). Section 3 concentrates on the modelling of fertility using a similar method as was used for mortality applied to the age-specific fertility rates. In Section 4 the statistical model for migration is presented. The kernel density estimator is used to estimate the age density of immigrants and emigrants, the level of in-moving and out-moving population is modelled by the appropriate time series processes. In Section 5 we describe the population renewal process using the cohort-component method and present the results of the population size, age structure and the age ratios. In Section 6 the consequence for the German pay-as-you-go pension system is discussed. Section 7 concludes the paper. The computations in this paper were made in Matlab 7.0.0 and R version 2.8.1.

2 Mortality

Due to the medical progress and the improvement of living conditions life expectancy in Germany increased substantially during the second half of the 20th century. Female life expectancy at birth rose from 70.9 years in 1956 to 82.3 years in 2006, male life expectancy at birth increased from 65.9 to 77.2 years in the same period. Beyond the changes in mortality (and the related life expectancy) in the time period, mortality changes in various age groups can also be observed. To be able to forecast mortality one needs to separate the variations in time and over age. In order to do that we use the model published by Lee and Carter (1992), which is described in the Subsection 2.2. The next subsection presents the available mortality data for Germany.

2.1 The Historical Mortality Data for Germany

For the analysis of mortality the annual age-specific periodic central death rates are used, defined as the number of deaths per 1000 living individuals, per one calender year. The annual age-specific death rates for the entire German population are available since German unification for years 1991 to 2006. For the period from 1956 to 1990 data for the Federal states in former West Germany and those in former East Germany are available only as separate data. All death rates series are issued by the Human Mortality Database (HMD; www.mortality.org) which offers interpolated annual rates (from the life tables published every 2 years by the Federal Statistical Office) for both genders, categorised in one-year age groups as $\{< 1, 1, 2, \ldots, 108, 109, \geq 110 = \omega\}$. Due to the small number of individuals of old age missing and inconsistent values (> 1) occur pretty often in the age groups older than 90. To remove them we use the linear interpolation approach assuming the mortality rate in the age group 110+ equals 1 and replacing all rates between the first missing or inconsistent value and the oldest age group by linear interpolation. For our data, modification of the life table in this way only negligibly alters the projected population over the sample period. The age groups older than 90 built in 2007 less than 1 % of the total population.

Figure 1 shows the logarithmic male and female death rates versus age groups for former West (circles) and East Germany (triangles). Comparing the rates in 1990 and 2006 one can

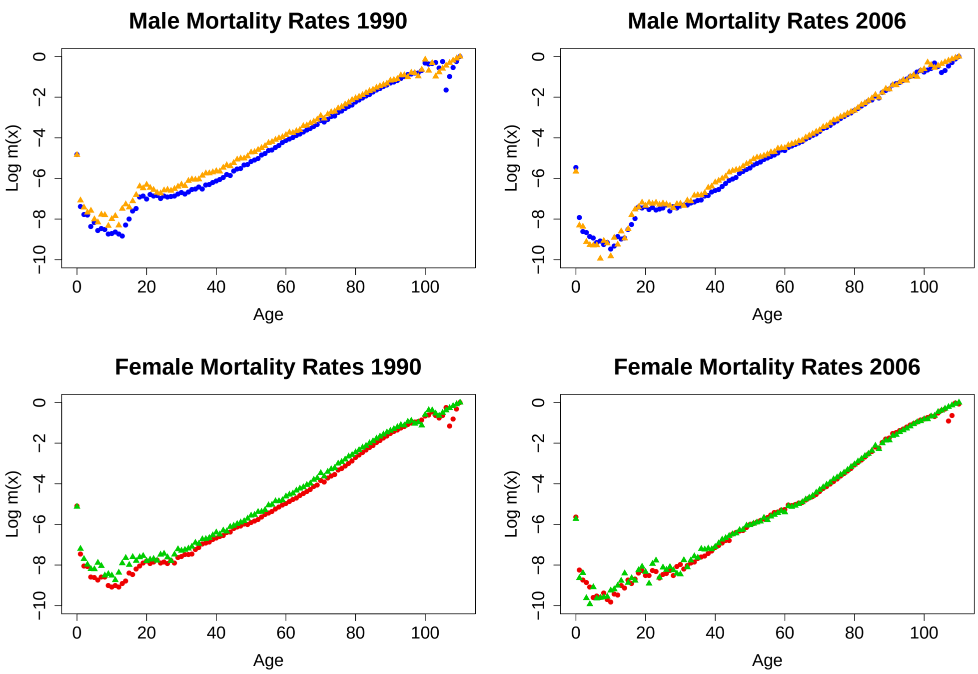

Figure 1: *Logarithmic death rates versus age group(0-110) for males (upper figures) and females (bottom figures) in 1990 and 2006 for former West Germany (circles) and East Germany (triangles).*

observe that the mortality behaviour in East Germany adapted the mortality circumstances seen in West Germany. We take this convergence as a given and hereafter use the West German data only in the model. The projected death rates are assumed to be identical for the East German population. For further comparison of mortality behaviour in West and East Germany see Lipps and Betz (2005).

2.2 Lee-Carter Model for Mortality Data

Mortality varies in time and over age groups. Figure 2 shows a surface of logarithmic death rates versus age groups and calender year for males and females. Strongly declining infant mortality in the last 50 years for both genders can be regarded as well a shift in death rates in all age groups. There is a significant "accident bump" in male mortality between the ages of 18 and 22, a phenomenon also observed by Dinkel and Luy (1999) and Diekmann et al. (2000).

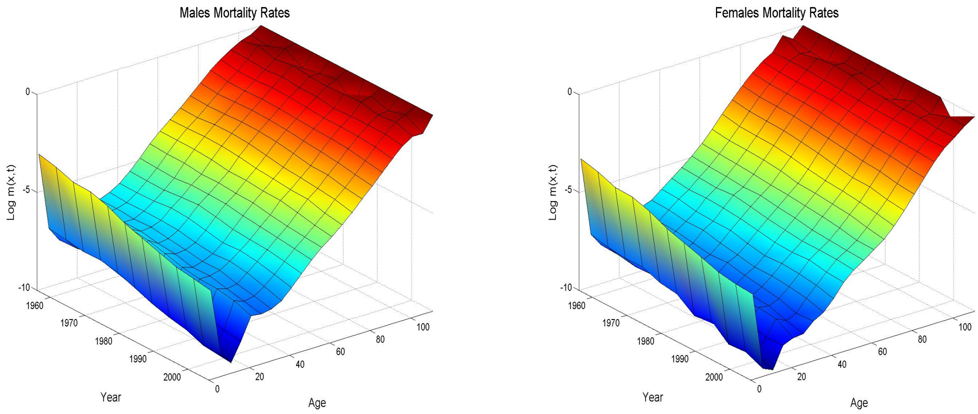

Figure 2: *Logarithmic death rates for males (left figure) and females (right figure) versus age (0-110+) and calendar year (1956-2006).*

To model and forecast the variation of mortality over age and time we use the well-known model proposed by Lee and Carter (1992). Let $m_{x,t}$ be the $(p \times T_m)$-matrix of central death rates for age groups $x = \{< 1, 1, \ldots, 109, \geq 110 = \omega\}$, $p = \omega + 1$, in years $t = \{1956, \ldots, 2006\}$, $T_m = 2006 - 1956 + 1 = 51$. To separate the time dependent part from the age-specific components, we fit the data matrix by the following model:

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

with age specific parameters a_x and b_x and a time varying index k_t. The error term $\varepsilon_{x,t}$ with assumed $\mathsf{E}(\varepsilon_{x,t}) = 0$ and $\mathsf{Var}(\varepsilon_{x,t}) = \sigma^2_\varepsilon$, reflects irregular age-time variations which arise mainly from particular (historical) circumstances. The exponential curve $\exp(a_x)$ describes the general shape of mortality whereas the parameter b_x tells us how fast the rates decline in response to changes in k_t, which follows from the first derivative $\left(\frac{d}{dt}\log(m_{x,t}) = b_x \frac{dk}{dt}\right)$.

As the model (1) is overparametrized one must set restrictions on the parameters to find an unique solution. We assume b_x to sum to unity over age groups and k_t to sum to zero over time. This simply implies for $\widehat{a}_x = \frac{1}{T_m} \sum_{t=1956}^{2006} \log(m_{x,t})$. To derive the factors k_t and b_x the singular value decomposition (SVD) method can be applied on the matrix $M(p \times T)$ of logarithmic death rates after the averages over time have been subtracted. The matrix M can be decomposed as follows:

$$M = [\log(m_{x,t}) - \widehat{a}_x] = \Gamma\ \Lambda\ \Delta^\top\ , \tag{2}$$

where $\Gamma(p \times r)$ and $\Delta(T_m \times r)$ represents column orthonormal matrices with $r = \text{rank}(M) = T_m$ for our data set. $\Lambda = \text{diag}(\lambda_1^{1/2}, \dots, \lambda_r^{1/2})$ is a diagonal matrix with non-zero eigenvalues $\lambda_i\ ,\ i = 1, \dots, r$ of the matrices $M^\top M$ and $MM^\top$, see Härdle and Simar (2007) for further explanation. The $(T_m \times 1)$ vector k_t is the first column vector of matrix Δ, consisting of eigenvectors of matrices $M^\top M$ and $MM^\top$, multiplied with the largest eigenvalue λ_1. The first $(p \times 1)$ vector of Γ after standardisation corresponds to the vector of the age-specific parameter b_x.

Figure 3 plots the estimated indices k_t^m and k_t^f for males and females, respectively (the forecasts are displayed as well). As shown, k declines roughly linearly in the plotted time period for both genders. For a deeper discussion of the reasons and for this decline and a correlation with macroeconomic factors, see Hanewald (2009).

Having developed and fitted the Lee-Carter model, we are now able to perform the forecast of mortality rates. First an appropriate time series model has to be found for k_t applying the Box-Jenkins analysis, see Hamilton (1994). The time series k_t is not stationary so we check whether the process of first differences $\widetilde{k}_t = \Delta k_t$ is a stationary process. In order to do that, the Augmented Dickey-Fuller test (ADF) with an included constant is applied, see Hamilton (1994):

$$\Delta\widetilde{k}_t = \alpha + \rho\widetilde{k}_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta\widetilde{k}_{t-i} + \varepsilon_t\ . \tag{3}$$

The test statistics are for males $A_m = -5.96$ and for females $A_f = -4.76$ with 5% critical value -1.95. We reject the null hypothesis of the unit root $H_0 : \rho = 0$ and hence $\widetilde{k}_t$ for males and females can be assumed as a stationary process. This result can be verified by using the KPSS test:

$$\widetilde{k}_t = c + \mu t + k \sum_{i=1}^{t} \xi_i + \varepsilon. \tag{4}$$

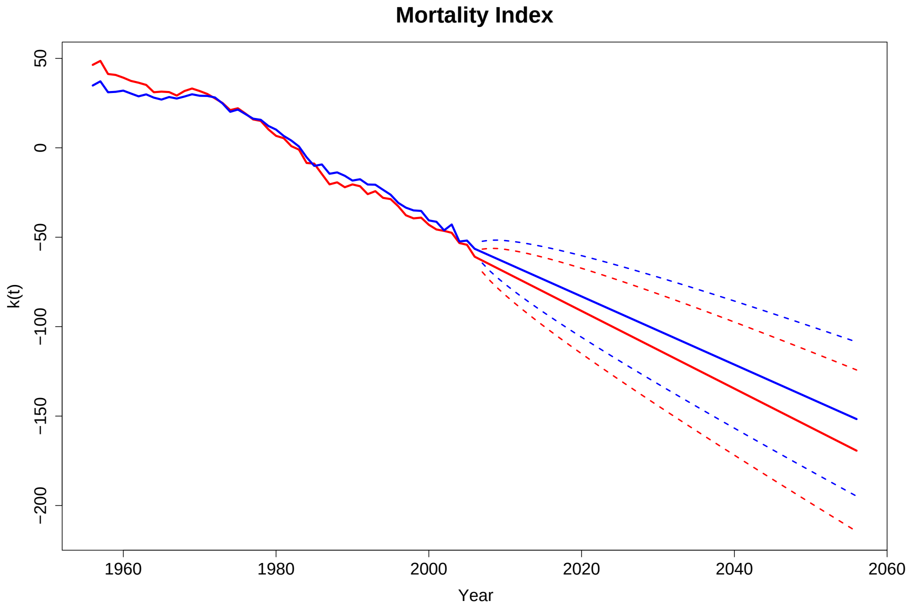

Figure 3: *Time-varying index k_t from 1956 – 2006 and the forecasts to 2056 with 95%-confidence intervals for males (blue) and female (red).*

For both genders, we accept the null hypothesis of stationarity, the test statistics for the constant c are $K_m = 0.38$ and $K_f = 0.26$ with a 5% critical value 0.46. The KPSS test statistics for trend μ are following: $K_m = 0.09$ and $K_f = 0.08$ with the 5% critical value 0.15.

The sample Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the integrated series $\widetilde{k}_t$ are close to zero, see Appendix 8, which denotes a white noise process of the differenced series. For that reason an Autoregressive Integrated Moving Average (ARIMA) process of order (0,1,0) is chosen as suitable to model both indices k^m and k^f. In the next step we provide a simple t-test whether a constant term δ should be added to the model. The absolute values of the test statistics (6.13 and 5.07) are larger than the 95%-quantile of the t-distribution with $T_m - 2 = 49$ degrees of freedom ($t_{0.95;49} = 1.68$) so that we reject the null hypothesis $H_0 : \delta = 0$ for both genders.

Therefore, the fitted model for the time-varying indices is a random walk with drift:

$$k_t = \delta + k_{t-1} + u_t \ , \tag{5}$$

with u_t being the white noise and δ denoting the drift parameter. The estimated model over the time period is:

$$\begin{array}{rclclcl} k_t^m & = & -1.83 + k_{t-1}^m + u_t^m & \text{with } \widehat{\sigma}_{u^m} & = & 3.11 & \text{for males,} \\ k_t^f & = & -2.14 + k_{t-1}^f + u_t^f & \text{with } \widehat{\sigma}_{u^f} & = & 3.26 & \text{for females.} \end{array} \tag{6}$$

The constant terms -1.83 and -2.14 represent the average annual change in k^m and k^f, respectively. $\widehat{\sigma}_u$ denotes the standard deviation of the white noise process u_t.

The forecast for k in the year $l + T_m$ follows a straight line:

$$k_{T_m}(l) = c_T + l \cdot \delta \qquad l = 1, 2, \dots\ , \tag{7}$$

with the constant c_T depending on the starting point T and the trend term in the ARIMA process δ. We choose the last observed point of the mortality index to be the starting point c_T for the forecast. The estimated indices k^m and k^f to 2056 with their 95%-confidence intervals are shown in Figure 3. After computing the future values for k we can now generate forecasts of the central death rates. For purposes of clarity life expectancy at birth to 2056 for both genders has been estimated, see Figure 4. Life expectancy shows an increasing trend, it grows in mean to 83.2 years for men and 89.1 years for women in 2056 compared to 77.2 and 82.3 years in 2006, respectively. As expected the forecasted values for females lie with a high probability above the forecasted male values.

3 Fertility

Fertility in Germany passed through a dynamic development in the last century. Concerning Total Fertility Rate (TFR), which represents the average number of children that a woman is expected to bear during her child-bearing years, Figure 8 shows that its maximum lies in 1964 with 2.5 children per woman and drops to its minimum of 1.3 children in 1985. Two structural breaks can be observed in TFR progression: the baby boom period of very high fertility from 1954 to 1966 followed by the strong baby bust starting in 1968 as the pill took effect.

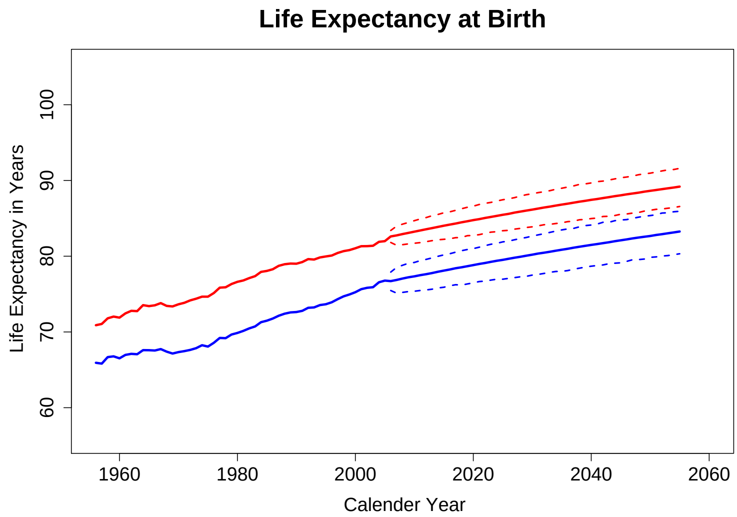

Figure 4: *Life expectancy from 1956 – 2006 for males(blue) and female(red) and the forecasts to 2056 with 95%-confidence intervals generated with 5000 simulations.*

3.1 The Historical Fertility Data for Germany

To measure fertility we use the annual age-specific fertility rates (ASFRs) defined as the number of births from mothers at the age x per 1000 women at the same age, per one calender year. The sum of the ASFRs over age x related to one women gives the TFR described above. As for the mortality data, the annual ASFRs for the entire German population are available for years 1991 to 2006. Only separate data for former West and former East Germany are available for the calender years from 1950. All fertility rates were obtained from the Federal Statistical Office which provides fertility rates for mothers from the age of 15 to the age of 49. As the data are calculated from the actual numbers of births and the population size the data sets do not contain any missing data.

Figure 5 shows the fertility rates versus age for former West (blue line) and East Germany (red line). Comparing the rates in 1990 and 2007 one can observe the fertility behaviour of East German women adapted in these 15 years the fertility behaviour of West German women: the women in East Germany have their children at an older age and the number of children declines. We assume this adaptation to continue in the future as well and use

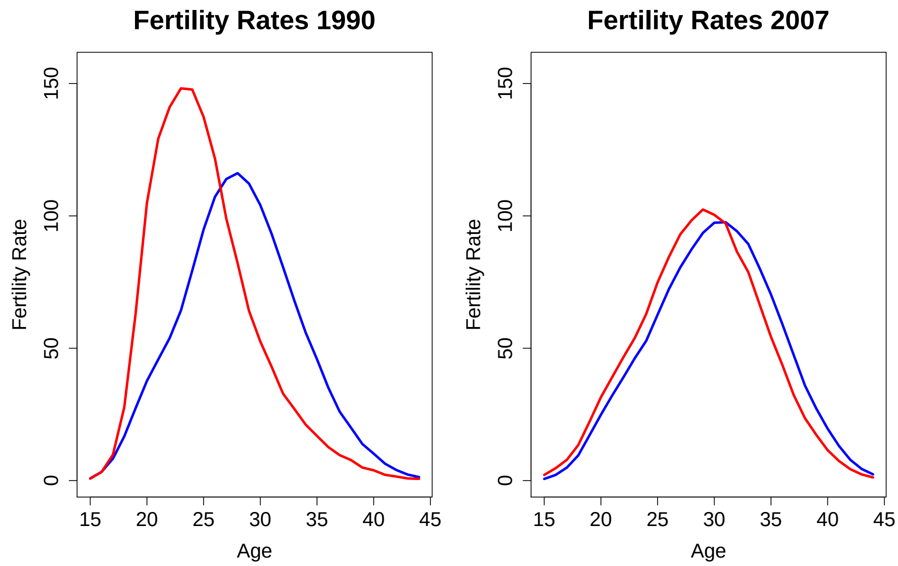

Figure 5: *Age-specific fertility rates versus mothers age (15-44) in 1990 (left figure) and 2007 (right figure) for former West Germany (blue) and East Germany (red).*

hereafter in the model the West German data only. The forecasted fertility rates are assumed to be identical for both parts of the German population.

3.2 Lee-Carter Model for Fertility Data

As shown in Figure 6 the fertility rates change in time and over the mothers age. A large peak of the baby boom children could be seen in the late 50s and and 60s as well the maximum of children born shifting to an older age of the mother since 1990.

To describe the variation in fertility over mothers age and time, we apply the Lee-Carter model discussed in Subsection 2.2 to the $(q \times T_f)$ matrix of ASFR $f_{x,t}$ for mothers of age $x = \{15, 16, \dots, 49\}$; $q = 49 - 15 + 1 = 35$ in calender years $t = \{1950, 1951, \dots, 2007\}$; $T_f = 2007 - 1950 + 1 = 58$. The Lee-Carter model for fertility is:

$$f_{x,t} = a_x + b_x f_t + \varepsilon_{x,t} \tag{8}$$

with identical assumptions as in Subsection 2.2. For fertility rates, one does not use logarith-

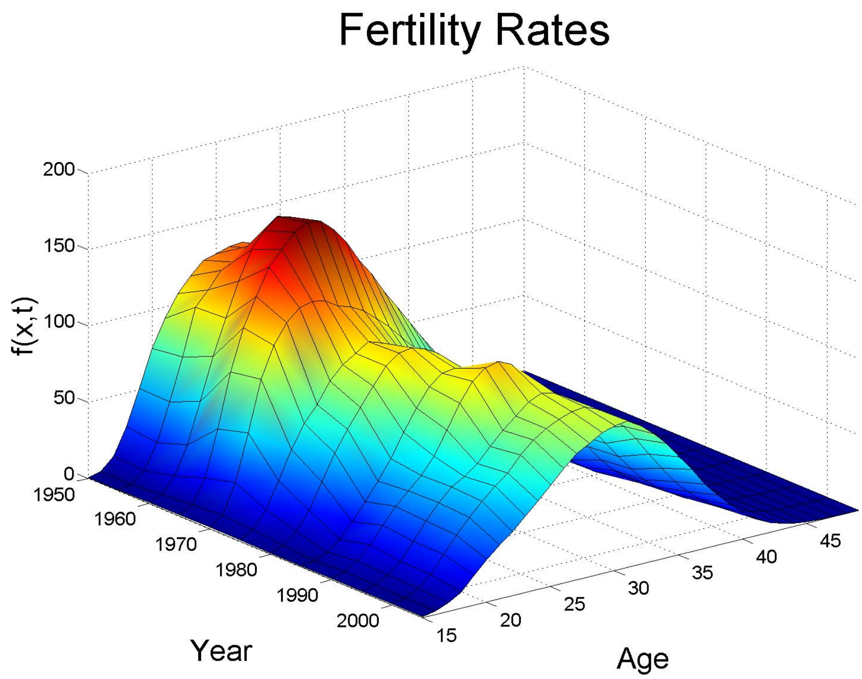

Figure 6: *Fertility rates versus mothers age (15-49) from 1950-2007 for former West Germany.*

mic rates in the model, as the fertility rates achieve large values between 0.2 (for mothers at age 47 or more) and 178 (for mothers at bearing age). The derivation of the age-specific parameter b_x and fertility index f_t is provided by the SVD of the $(q \times T_f)$ matrix $N = f_{x,t} - \widehat{a}_x$, after estimating $\widehat{a}_x$ as the mean value of ASFR over the time period T_f, see Subsection 2.2.

Summing both sides of the equation (8), one gets: $TFR_t = A + f_t + E_t$, where $A = \sum_{x=15}^{49} \widehat{a}_x$ and $E_t = \sum_{t=1950}^{2007} \varepsilon_{x,t}$. Thus, the fertility index f_t can be interpreted as a deviation of the TFR in period t from its long term average A.

To provide time series analysis it will be convenient to work with the fitted value of the TFR: $F_t = A + f_t$, where f_t fluctuates around zero. To ensure a demographically plausible forecast of fertility, Lee (1993) suggests incorporating pre-specified lower and upper bounds on the TFR directly into the modelling process. Denoting L and U as the lower and upper bounds, respectively, let us define a transformed fertility process g as follows:

$$g_t = \log\left(\frac{F_t - L}{U - F_t}\right) \ . \tag{9}$$

After the series g was modelled and forecasted one obtains the forecast for F_t by inverse transformation from g:

$$F_t = \frac{U \cdot \exp(g_t) + L}{1 + \exp(g_t)} \ . \tag{10}$$

It is obvious that as g goes to infinity, F goes to the upper bound U; as g goes to negative infinity, F goes to the lower bound L. Due to this characteristic of the logistic transform, the forecast and its confidence interval falls within these limits.

For the German data, the bounds were set on TFR to lie between 0 and 5. Furthermore, we do not assume any structural breaks for the fertility process in the future, for the analysis we consider the fertility rates only after the pill took effect, i.e. from 1976 to 2007. Realising the Box-Jenkins analysis for the transformed index g (see Figure 7), we first test whether g_t is a stationary process providing the ADF test including a drift term, see (3). The test statistic equals -2.81 with a 10% critical value -2.60, so that we reject the null hypothesis of unit root. We can validate our result by the KPSS test for stationarity, see (4), which has the test statistics 0.16 for trend μ and 0.13 for the constant c. Both values are close to zero so that we cannot reject the null hypothesis of stationarity. Analyzing the sample ACF and PACF functions of series g_t and rejecting the t-test for zero intercept δ (see Appendix 8) a simple Autoregressive Moving Average model (ARMA) of order (1,1) was chosen as the appropriate model. The general formula of an ARMA(1,1) process is:

$$g_t = \delta + \phi \ g_{t-1} + \theta \ u_{t-1} + u_t \ , \tag{11}$$

with intercept δ, AR-parameter ϕ, MA-parameter θ and innovations u_t following the white noise process: $u_t \sim (0, \sigma_u^2)$. The fitted model for g is then:

$$g_t = -0.96 + 0.51 \ g_{t-1} + 0.33 \ u_{t-1} + u_t \ , \tag{12}$$

with $u_t \sim (0, \widehat{\sigma}_u^2 = 9.22 \cdot 10^{-4})$. The forecast of an ARMA process decays geometrically at the rate of the AR-parameter toward the unconditional mean $\widehat{\mu} = \dfrac{\delta}{1-\phi}$. The point forecast for g with its 95% confidence interval is shown in Figure 7.

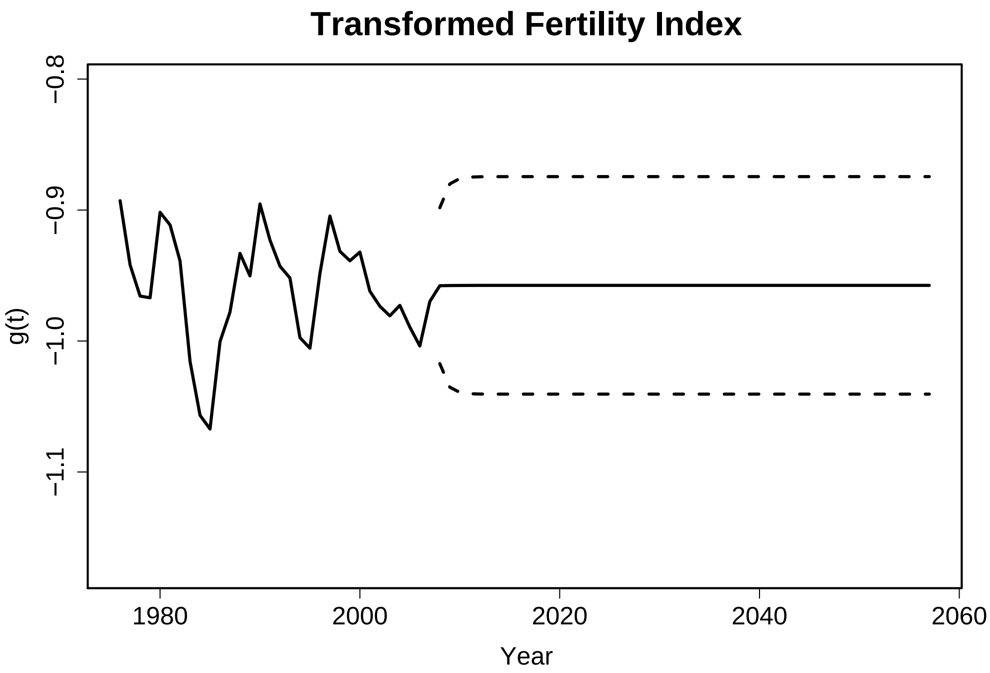

Figure 7: *Transformed fertility index from 1976 – 2007 and the forecast to 2058 with the 95%-confidence interval.*

Using equation (10), the forecast of g_t can be easily converted into the forecast for F_t or $f_t = F_t - A$, respectively. Figure 8 shows the TFR with the forecast and its 95%-confidence interval.

4 Migration

The third demographic variable involving the structure and size of the population is population migration. The number and age of immigrating and emigrating people is influenced on the one hand by many political, economic, demographic and ecological factors in migration countries. On the other hand, the number of immigrants is affected by the migration policy and the development of the labour market in Germany as a destination country for migration.

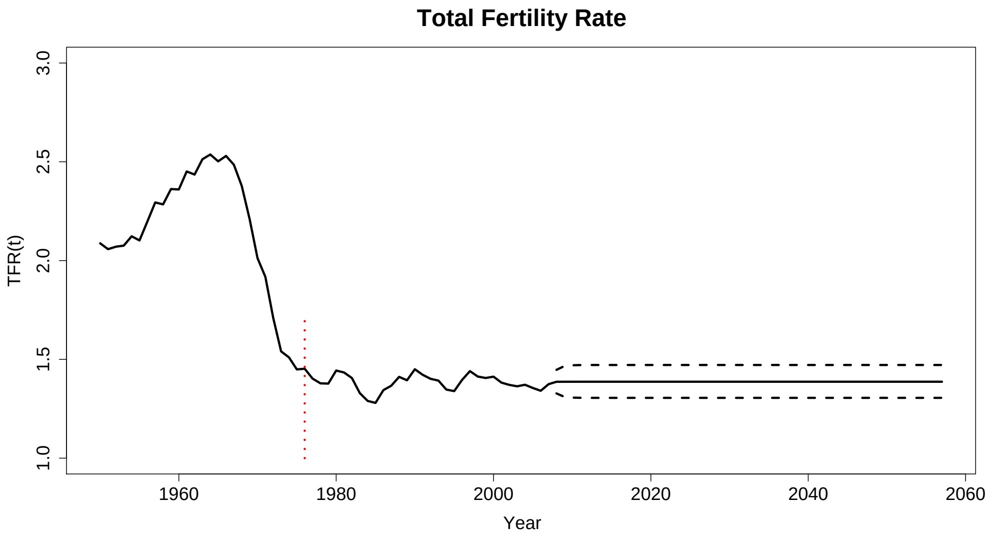

Figure 8: *Total Fertility Rates from 1950 – 2007 and the forecast to 2057 concerning data from 1976 only (dotted red line) with the 95%-confidence interval.*

4.1 Historical Migration Data for Germany

After the World War II and foremost after the construction of the Berlin Wall 1961 until its fall in 1989, migration from and into the former German Democratic Republic was strongly regulated. Indeed, official statistics do not show the realistic, larger numbers of emigrants from East Germany to West Europe or to the USA. For this reason, we consider migration data before 1990 from the former West Germany only. Since the German unification in 1990 until 2007, we use the migration information for whole of Germany. The numbers of migrants are available for age groups $\{< 1, 1, \ldots, 89, 90+\}$. The data was provided by the German Federal Statistical Office.

The total number of immigrants, emigrants and their difference are shown in Figures 9 for males and in Figure 10 for females. Both immigration processes are defined by strong past fluctuations. One can observe two big peaks in the immigration processes in 1990 and 1992 corresponding to the large number of Russian and Rumanian Germans which immigrated into Germany after the Iron Curtain fell in Eastern Europe. More than 7 million people came to Germany between 1989 and 1993 as an effect of the open borders in Eastern Europe. In 1993

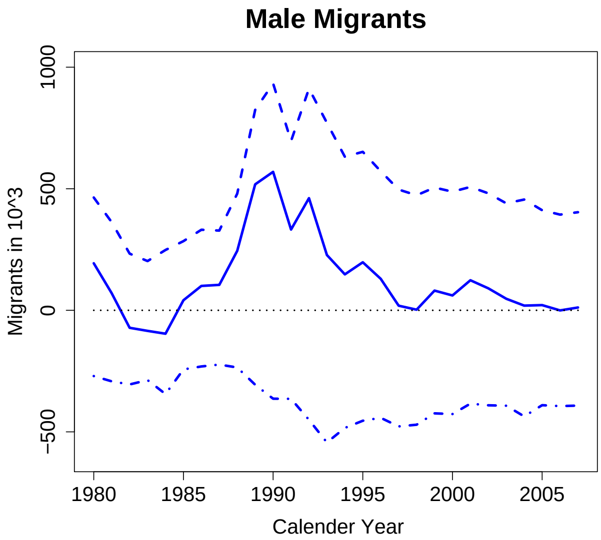

Figure 9: *Level of net migration (solid line), immigration (dashed line) and emigration (dotdashed line) for males from 1980 – 2007.*

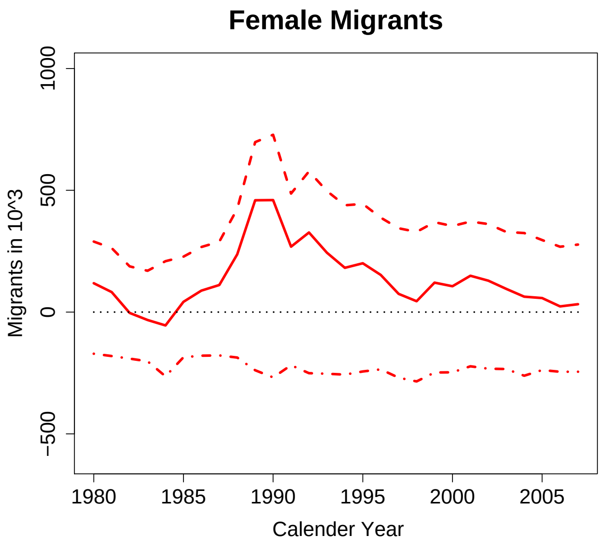

Figure 10: *Level of net migration (solid line), immigration (dashed line) and emigration (dotdashed line) for females from 1980 – 2007.*

German policy reacted with stricter asylum laws which rapidly brought down the number of immigrants. After 1998 the level of immigrants stabilised at around 500 thousands for males and 350 thousands for females becoming a downward trend since 2001. In our model, we assume the validity of the asylum laws from 1993 in the future and herewith consider the consistent immigration process only from 1994 to 2007.

In contrast, the emigration process behaves more constantly for both genders. There is an increase in emigrating males after 1990 so that net migration decreased to closed to zero in 1998 and 2006. The development of emigrating females fluctuates around its mean value of 230 thousands people with a small bump of more than 280 thousands in 1998. The behaviour of the stable number people moving out of a country over a time period is called a "base emigration".

With the decreasing immigration level and a stable "base emigration" one observes a slight decreasing trend in net migration since 2001.

4.2 Modelling of Migration Data

In our model of the migration process we treat the in-migration and the out-migration for both genders separately and assume them to follow a stochastic process. Beside the levels of the immigrating and emigrating population, we focus on the age structure of individuals moving. In contrast to the mortality and fertility process, there is no trend in the age structure of the moving population in the observed time period. Figure 11 shows the estimated age density of immigrating people from 1994 until 2007, Figure 12 displays the estimated density for emigrants between 1980 and 2007. The kernel density estimator $\widehat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$, with bandwidth h, number of observations n and kernel function K, was used to derive the density in each calender year of the time period. The Gaussian kernel function was chosen as appropriate, the bandwidth selection followed the Silverman's rule of thumb, see Härdle and Werwatz (2004).

Neither in the in-moving nor in the out-moving population any significant change in the age of the migrants can be found, as shown in Figures 11 and 12. The majority of immigrating men are between 20 and 40 years old. The variance of age in immigrating women is smaller, most of them are between 19 and 30 years old.

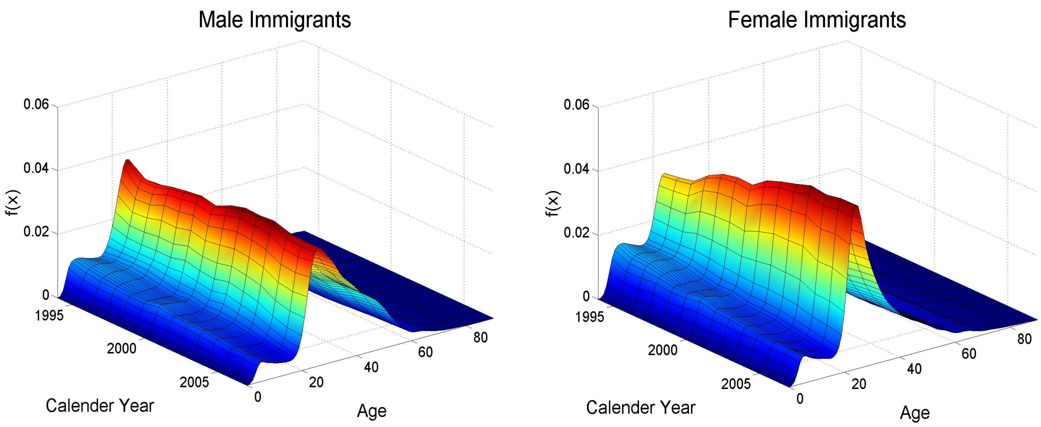

Figure 11: *Estimated density of age of male (left figure) and female (right figure) immigrants in the time period 1994-2007.*

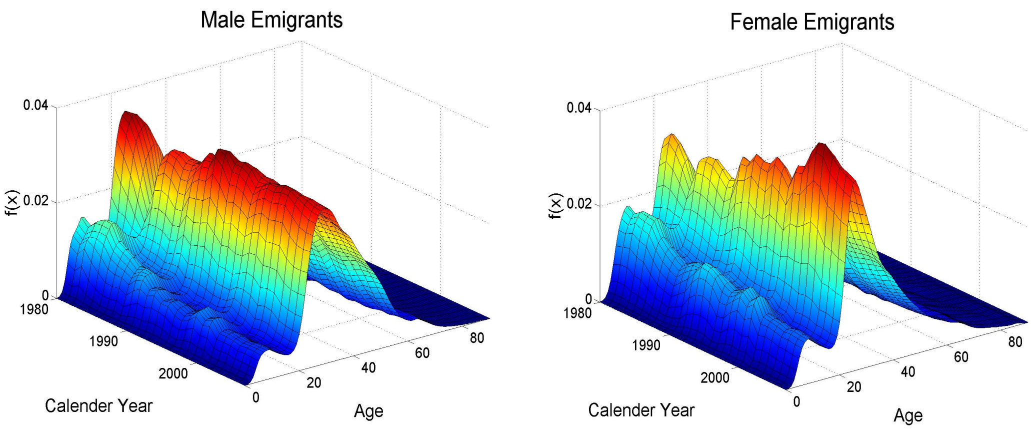

Figure 12: *Estimated density of age of male (left figure) and female (right figure) emigrants in the time period 1980-2007.*

The age of the emigrating population has a stable distribution over the time period, out-moving men are mainly between 20 and 45 years, most of the out-moving women are between 20 and 30 years.

To forecast the migration process we first model the level of the immigrating and emigrating people with a suitable time series model. After the estimation of model parameters we are able to construct a forecast for the number of in-moving and out-moving people. In each forecasted calender year the number of immigrating and emigrating people in the single age group is estimated by the kernel density estimator of the migrants' ages from 2007. The number of net migrants is then calculated as the difference between the immigrants and emigrants in each age group.

Series	p-value
i_t^m	< 0.01
i_t^f	0.04
e_t^m	0.57
e_t^f	< 0.01

Table 1: *p-values for the ADF unit root test including an intercept and a drift term.*

The first step of the analysis consists of a verification of the stationarity of the time series. Table 4.2 shows the p-values of ADF unit root test, see (3). KPSS test statistics for constant τ and trend μ with corresponding critical values are shown in Table 4.2. Hence, we can assume that the immigrating process for both genders and the emigrating process for females are stationary. In the case of emigrating men, the null hypothesis of unit root was not rejected but the KPSS test accept the null hypothesis of stationarity. For that reason we can assume stationarity for this process as well.

	KPSS for τ		KPSS for μ	
Series	$\text{KPSS}_{\text{stat}}$	$\text{KPSS}_{\text{CV}}(\alpha)$	$\text{KPSS}_{\text{stat}}$	$\text{KPSS}_{\text{CV}}(\alpha)$
i_t^m	0.200*	0.216 (1%)	0.396*	0.347 (10%)
i_t^f	0.206*	0.216 (1%)	0.417*	0.463 (5%)
e_t^m	0.116*	0.146 (5%)	0.287*	0.347 (10%)
e_t^f	0.133*	0.146 (5%)	0.351*	0.463 (5%)

Table 2: *KPSS test statistics ($\text{KPSS}_{\text{stat}}$) for the constant τ and for the trend μ with corresponding critical values on the significance level α (KPSS_{CV}). The symbol $*$ denotes the accepting of stationarity hypothesis in KPSS test.*

Analysing the sample ACF and the PACF of the time series i_t and e_t and testing estimated parameters against zero, see Appendix 8, the autoregressive process (AR) of order 1 was chosen as the appropriate time series model for the immigration and emigration processes for both genders. The AR(1) process has a general form:

$$y_t = \delta + \phi y_{t-1} + \varepsilon_t \ ,$$

with intercept δ, parameter ϕ and innovations ε_t which follows a white noise process: $\varepsilon_t \sim (0, \sigma_\varepsilon^2)$.

The fitted model for the process of immigrants number (in thousand), denoted with i_t is then:

$$\begin{aligned} i_t^m &= 562.85 + 0.94\, i_{t-1}^m + \varepsilon_t^m \quad \text{with} \quad \widehat{\sigma}_{u^m} &= 52.52 \quad \text{for males,} \\ i_t^f &= 377.93 + 0.94\, i_{t-1}^f + \varepsilon_t^f \quad \text{with} \quad \widehat{\sigma}_{u^f} &= 31.18 \quad \text{for females.} \end{aligned}$$

Analogous, the fitted model for the process e_t of number of emigrants (in thousands) follows:

$$\begin{aligned} e_t^m &= 359.76 + 0.86\, e_{t-1}^m + \varepsilon_t^m \quad \text{with} \quad \widehat{\sigma}_{u^m} &= 42.70 \quad \text{for males,} \\ e_t^f &= 227.55 + 0.60\, e_{t-1}^f + \varepsilon_t^f \quad \text{with} \quad \widehat{\sigma}_{u^f} &= 26.07 \quad \text{for females.} \end{aligned}$$

Fitted models confirm the facts one could observe in Figures 9 and 10: the emigration level is lower than the immigration one and the variance of the emigration process is substantially smaller then the variance of the strong fluctuating immigration process, considered for separated genders.

The forecast of the stationery AR(1) process decays geometrically with increasing forecast horizon toward its unconditional mean $\mu = \dfrac{\delta}{1-\phi}$.

5 Population Projection

After estimating of models for the demographic processes of mortality, fertility and migration it is possible to construct the stochastic population forecast. The projection is performed by a cohort-component-method which creates a matrix of survival probabilities and fertility rates (so-called Leslie matrix, named after Leslie (1945) who invented the matrix representation in the population forecasting) and a vector of net migrants for each forecasted calender year, see Diekmann et al. (2000).

5.1 Cohort-Component-Method

The definition of variables for the matrix notation of the cohort-component-method is as follows:

- $N^M_{x,t}$ or $N^F_{x,t}$ as the male or female population count, respectively, of the age group x in year t,
- $F_{x,t}$ as the age-specific fertility rate for mother of age x in year t which corresponds to the ASFR related to one women,
- $P^M_{x,t}$ or $P^F_{x,t}$ as the probability for men or women, respectively, of the age x to reach the next year $t+1$, defined as: $p_{x,t} = \frac{(1-q_{x,t})\cdot(1-\frac{q_{x+1,t+1}}{2})}{1-\frac{q_{x,t}}{2}}$ with the probability of death: $q_{x,t} = \frac{2m_{x,t}}{2+m_{x,t}}$.
- $NI^M_{x,t}$ or $NI^F_{x,t}$ as the number of male or female net migrants, respectively, of the age x in year t
- $s = 0.4854$ as the sex ratio at birth taken as 100 newborn girls to 106 newborn boys, which corresponds to the historical average value for Germany.

Given the population in the calender year t one calculates the female population on the next year $t+1$ as follows:

$$\begin{pmatrix} N^F_{1,t+1} \\ N^F_{2,t+1} \\ \vdots \\ \vdots \\ N^F_{x,t+1} \\ \vdots \\ \vdots \\ N^F_{k,t+1} \end{pmatrix} = \begin{pmatrix} 0 & \dots & s\cdot F_{15,t} & \dots & s\cdot F_{49,t} & 0 & \dots & 0 \\ P^F_{1,t} & 0 & \dots & \dots & \dots & 0 & \dots & 0 \\ 0 & P^F_{2,t} & \ddots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & \ddots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & \dots & \dots & 0 & P^F_{k-1,t} & 0 \end{pmatrix} \cdot \begin{pmatrix} N^F_{1,t} \\ N^F_{2,t} \\ \vdots \\ \vdots \\ N^F_{x,t} \\ \vdots \\ \vdots \\ N^F_{k,t} \end{pmatrix} + \begin{pmatrix} NI^F_{1,t} \\ NI^F_{2,t} \\ \vdots \\ \vdots \\ NI^F_{x,t} \\ \vdots \\ \vdots \\ NI^F_{k,t} \end{pmatrix}.$$

The calculation of the future male population proceeds analogously; indeed there are no fertility rates in the Leslie matrix and the male newborns are calculated as:

$$(1-s)\sum_{x=15}^{49} \cdot F_{x,t} \cdot N_{x,t}^{F} \ .$$

In our model, we assume that all births, deaths and migrations happen on one day at the end of each year. As our interest lies in the forecast for a larger time span, this assumption only alters the results negligibly. In addition, our cohort-component-model assumes that the immigrating population adapts directly to the mortality and fertility conditions prevalent in Germany. We are compelled to adhere to this assumption due to the lack of available fertility and mortality data on foreign immigrants. As a starting population, we use the population structure in Germany from the 01.01.2007 obtained from the Human Mortality Database so that we are able to apply the estimated mortality rates for 2007 and combine them with known fertility rates and migration level. From 2008 on, we use the forecasted vital rates only for the estimation.

5.2 Results

In the following section the results of the population forecast till 2057 are presented. The mean projection of the population size with its 95% confidence intervals is listed in Table 3. The population size decays in mean from 83 m. in 2007 to 76 m. in 2055. The distribution of the forecasted population size in years 2010, 2025, 2040 and 2055 is shown by histograms in Figure 13. The forecasting interval grows up with the increasing forecasted time period. So the 95% interval amounted to 1.5 m. in 2015 increases to almost 8 m. people in 2055.

The rapidly changing structure in the German population is reflected in Figure 14 by the population pyramids in years 2010, 2025, 2040 and 2055. Population pyramids (with the red field for female and blue field for male population) display the decreasing number of children and increasing number of the elderly. For comparison, the population structure from 2007 is shown as the grey pyramid on the background.

In Figure 15, we compare our results with the deterministic forecast of the Federal Statistical Office, see Stat.Bundesamt (2006b). The lower and upper bound of the middle scenario are

Year	5%-Quantile	Mean	95%-Quantile
2010	81.89	82.13	82.35
2015	81.21	82.00	82.75
2020	80.48	81.81	83.14
2025	79.50	81.44	83.32
2030	78.46	80.88	83.24
2035	77.35	80.20	82.96
2040	76.14	79.44	82.64
2045	74.87	78.52	81.99
2050	73.52	77.36	81.06
2055	71.93	76.02	79.86

Table 3: *Projection and its 95% confidence intervals of the population size in selected years.*

denoted with the red dash-dotted lines, the upper bound lies in the 95% confidence interval of our forecast. As well, the extreme scenario with the largest population size (blue dash-dotted line) falls into our confidence interval. The extreme scenario with smallest population size and the lower bound of the middle scenario expect a smaller population size than our model. The reason for that may lie in the assumptions of a low immigration level of 100 000 people per year in both scenarios. The range of the 2 extreme scenarios grows very fast in the forecasted period and achieves its maximum of 12.5 m. (between 67 and 79.5 m.) citizen in 2050. For the same forecasted period the range of the confidence interval in our model amounts to 7.5 m. and indeed, one can assign the probability of 95% to this interval, as for all estimated demographic factors.

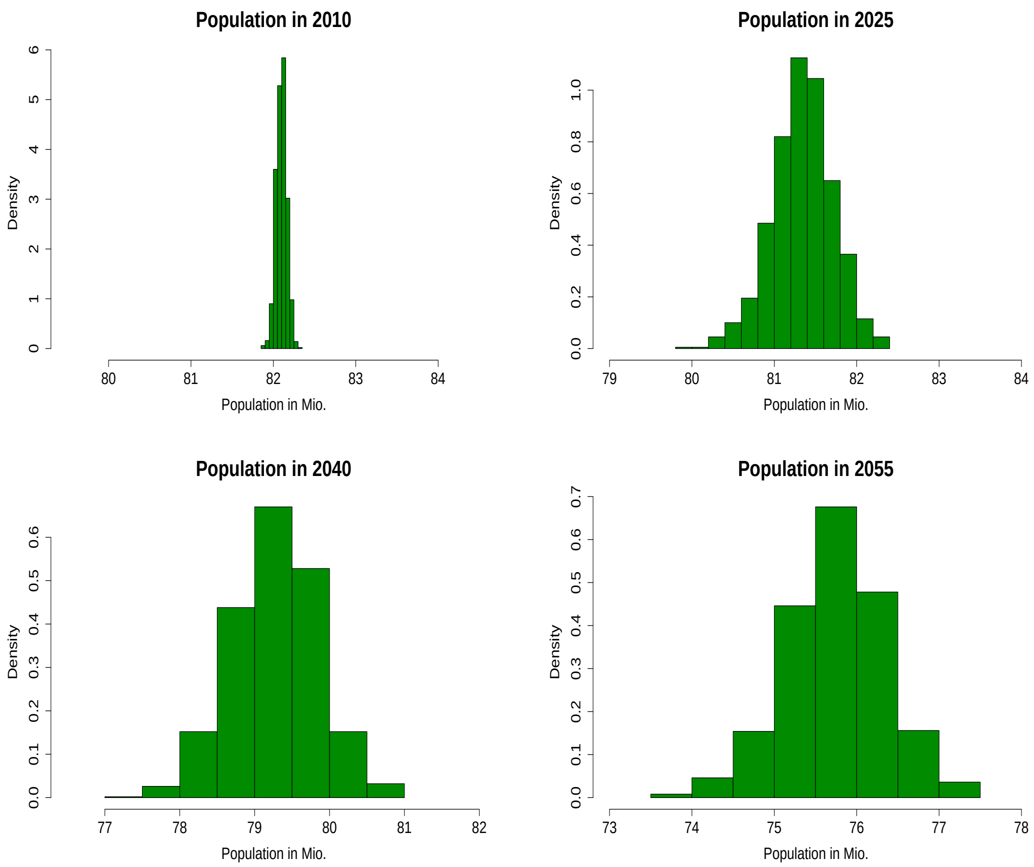

Figure 13: *Histograms of the population size in forecasted years 2010, 2025, 2040 and 2055.*

6 Consequence for the German Pay-as-you-go Pension System

After the World War II, the German pension system was reorganized to a fully pay-as-you-go financing. Since that time, the German society passed many changes including the German unification in 1990 after which all citizens from the former German Democratic Republic were included in the pension system of the old West German states.

The key factor for an efficient pay-as-you-go pension system is the old-age dependency ratio: the ratio of the elderly population with entitlement to a state pension to the active population

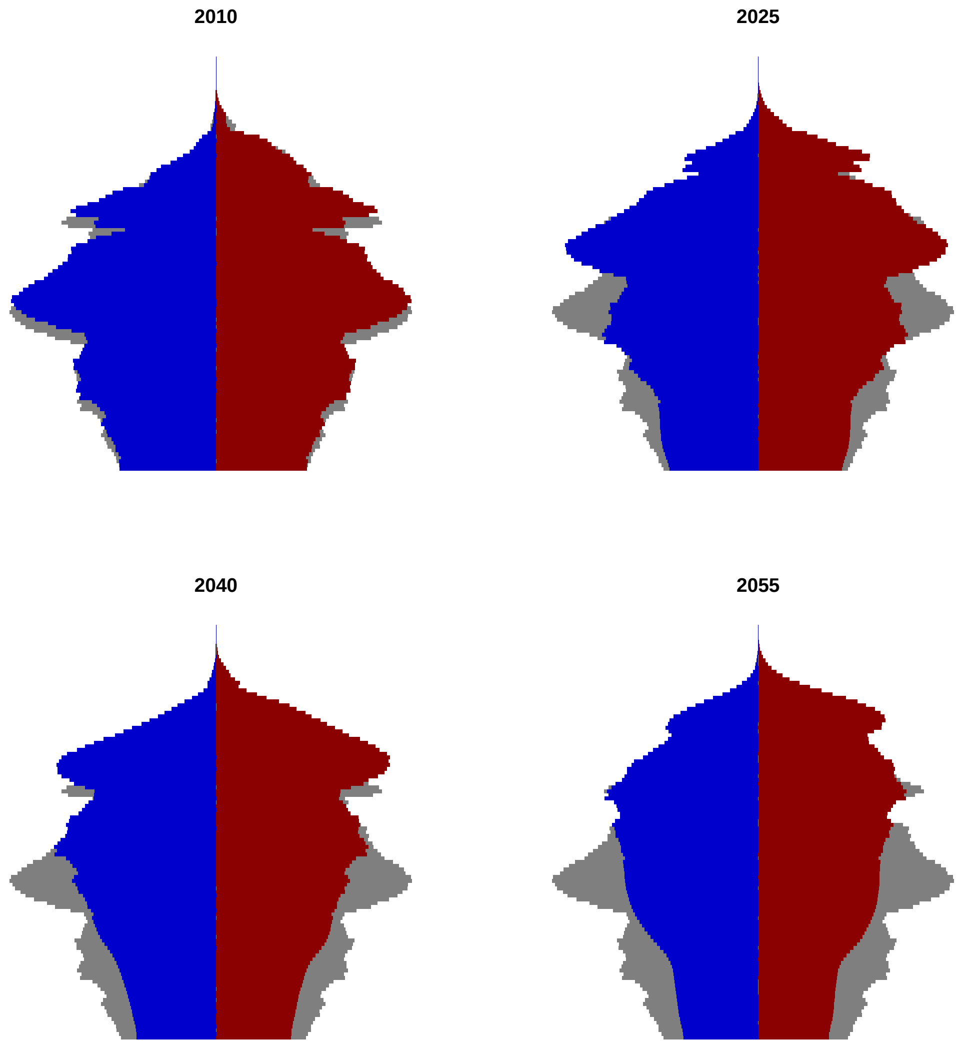

Figure 14: *Population pyramids in forecasted years 2010, 2025, 2040 and 2055. The blue line denotes male population, the red line female population, dashed lines correspond to the 95%-confidence intervals.*

over 20. Figure 16 shows two estimated old-age dependency ratios with their 95% confidence intervals. The black lines correspond to the ratio of elderly population from 65, the green lines are corresponding to the ratio with an age limit of 67 years, since the age of retirement in Germany will be increased to 67 years from 2012. In our model, both ratios increase

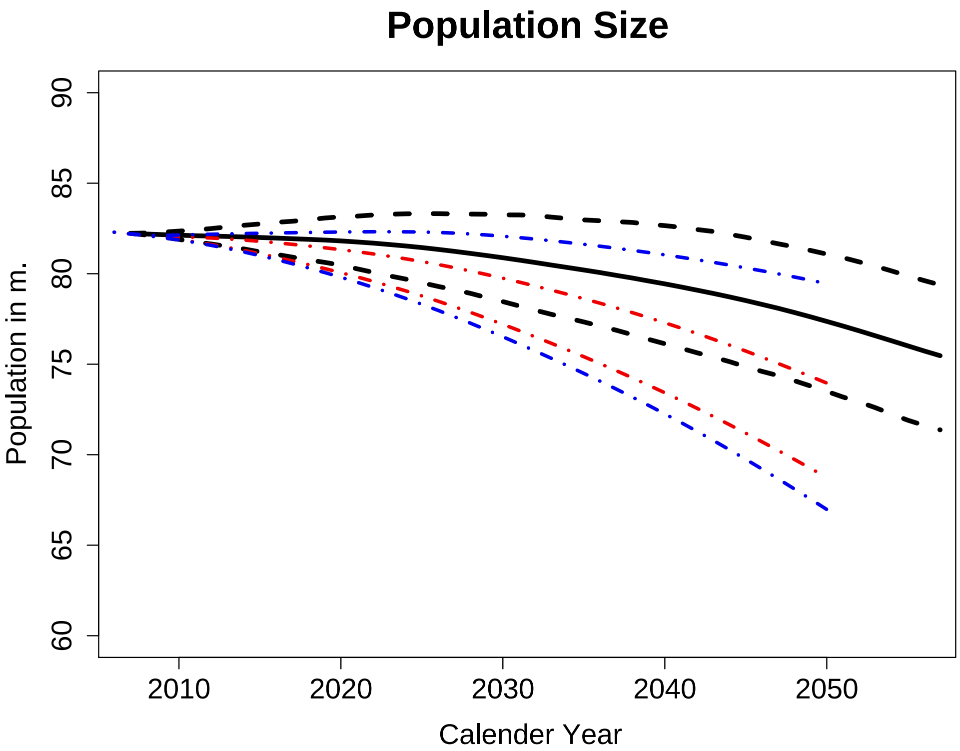

Figure 15: *Population size from 1990 – 2006 and its estimate until 2056 (solid line) with 95%-confidence intervals (dashed lines), compared to 4 scenarios of Federal Statistical Office (red lines denotes the limits of middle scenario, blue lines denotes the maximal and minimal population size by the combination of all scenarios).*

rapidly in the next 50 years. The old-age ratio with the limit 65 years rises from 32% in 2007 into an interval between 54% and 69% in year 2057. The old-age ratio with the limit of 67 years shows a similar development; it rises from 27% in 2007 to an interval between 46% and 60%. This means that with 95% probability at least one person in retirement age falls on one person between 20 and 64 years or 66 years, respectively.

To keep the system functioning one of the financial sources – the social system premium rate, the benefit level or the government subventions into the system – has to change. In Figure 17, we show estimated minimal required premium rates in case the actual average benefit level of 720 EUR per months and actual government subventions in the amount of

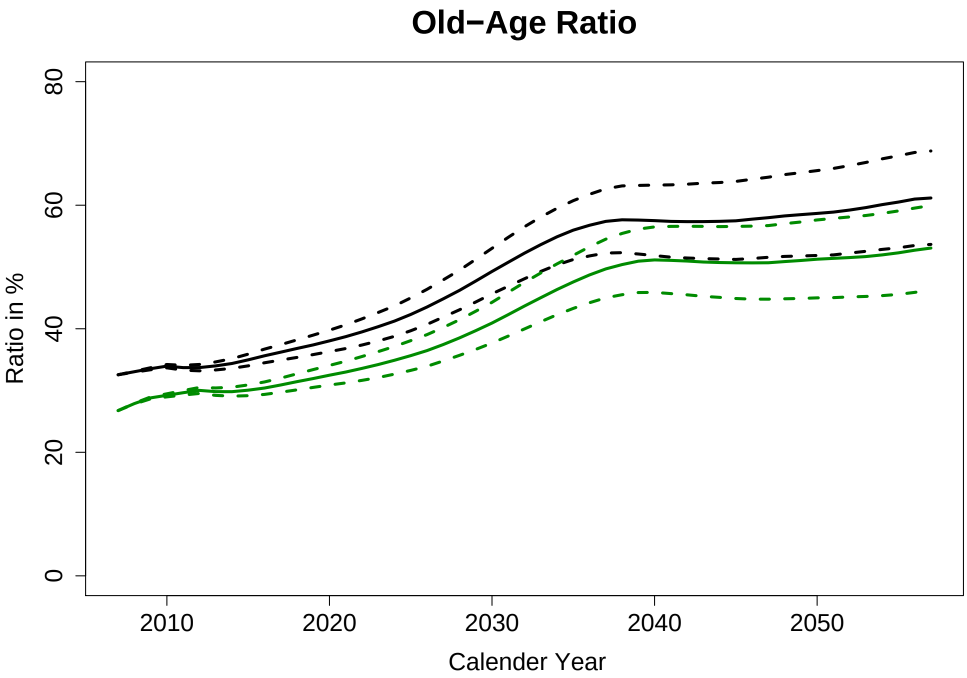

Figure 16: *Old-age dependency ratio with age limit 65 years (black) and with age limit of 67 years (green) with their 95% forecast intervals.*

25% of total revenues of the pension system are maintained. The estimated premium rate rises from 19.9% in 2007 to the maximal values between 26.0 and 30.1 % in 2040. The slight decline of the premium rate until 2015 is caused by the increase of the age of retirement to 67 years.

Estimated average benefit level in percent of the average level in 2007 (720 EUR per month) with its confidence interval is shown in Figure 18. The benefit level drops in mean by 28% until 2040 if the actual premium rate of 19.9 % and actual government subventions keep maintained. In our simulations we assumed that the actual circumstances of salary and employment levels will be maintained. A deeper discussion of the impact of demographic uncertainty on the public finances can be found by **?**.

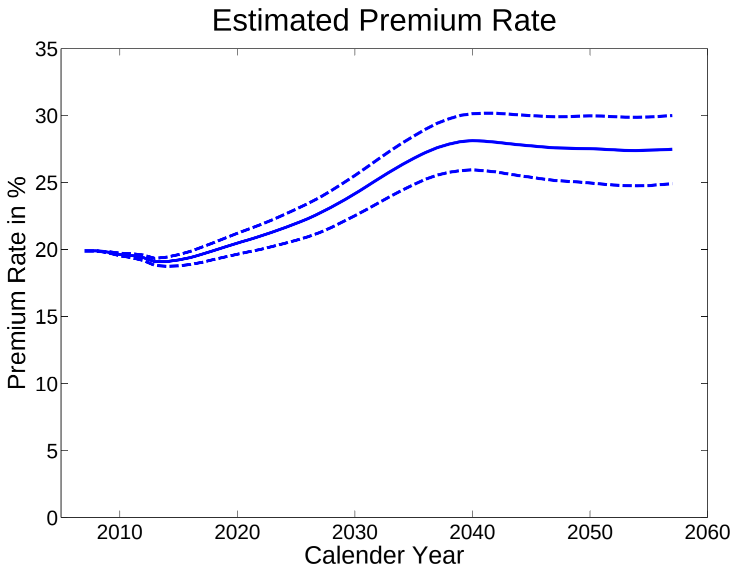

Figure 17: *Estimated required premium rate in the German Social System with 95% confidence interval.*

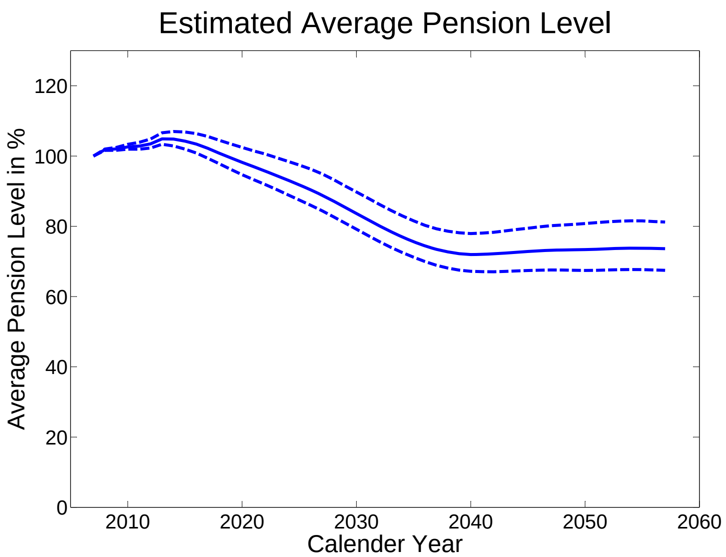

Figure 18: *Estimated averaged benefit level in percent (2007=100) with 95% confidence interval.*

7 Conclusion

We have provided a population forecast for Germany using the actual data of age-specific death rates, fertility rates and migration level. In our model of fertility and mortality, we have combined the classical approach from Lee and Carter with a time series analysis of the time-dependent factors of these two demographic variables. To model the migration, we have combined the appropriate time series models for processes of gender specific immigration and emigration with a nonparametric age density estimation. We have determined the forecast of population size and its age structure indicated among others by the old-age ratio. The consequence for the pay-as-you-go financed pension system is shown on predicted future premium rate and averaged pension level. For our forecasted factors we can also produce prediction intervals which take into account all sources of variation and estimate the distribution of our forecast.

Acknowledgment
Both authors would like to thank Prof.Dr.Hans-Peter Schwintowski for his constructive ideas.

References

Jensen S.E.H. Alho, J.M. and J. Lassila, editors. *Uncertain Demographics and Fiscal Sustainability*. Cambridge University Press, 2009.

J.M. Alho and B.D. Spencer. Uncertain population forecasting. *Journal of the Statistical Association*, 80:306–314, 1985.

B. Babel. *Bevölkerungsvorausberechnungen unter besonderer Berücksichtigung von Unsicherheiten*. EUL Verlag, Lohmar - Köln, 2007.

V. D'Amato and Russolillo M. Lee carter error matrix simulation: heteroschedasticity impact on actuarial valuations. *Presented at the International Conference MAF 2008: Mathematical and Statistical Methods for Actuarial Sciences and Finance.*

DRV Deutsche Rentenversicherung Bund. *Aktuelle Daten 2007*. Deutsche Rentenversicherung Bund, 2006a.

DRV Deutsche Rentenversicherung Bund. *Rentenversicherung in Zeitreihen*. Deutsche Rentenversicherung Bund, 2006b.

A. Diekmann, U. Mueller, and B. Nauck. *Handbuch der Demographie*. Springer, Heidelberg, 2000.

R.H. Dinkel and M. Luy. Natur oder Verhalten? Ein Beitrag zur Erklärung der männlichen Übersterblichkeit durch einen Vergleich von Kloster- und Allgemeinbevölkerung. *Zeitschrift für Bevölkerungswissenschaft*, 24,2:105–132, 1999.

F. Girosi and G. King. *Demographic Forecasting*. Princeton University Press, New Jersey, 2008.

J.D. Hamilton. *Time Series Analysis*. Princeton University Press, New Jersey, 1994.

K. Hanewald. Lee-carter and the macroeconomy. *Discussion Paper, SFB 649 Humboldt University Berlin*, 008, 2009.

Müller M. Sperlich St. Härdle, W. and A. Werwatz. *Nonparametric and Semiparametric Models*. Springer Verlag, Heidelberg, 2004.

W. Härdle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer Berlin, 2nd edition, 2007.

R.J. Hyndman and H. Booth. Stochastic population forecasts using functional data models for mortality, fertility and migration. *Monash Econometrics and Business Statistics Working Papers, Monash University*, 14/06, 2006.

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

N. Keilman. Why population forecasts should be probabilistic - illustrated by the case of norway. *Demographic Research*, 6:409–454, 2002.

R.D. Lee. Modeling and forecasting the time series of US fertility: Age distribution, range, and ultimate level. *International Journal of Forecasting*, 9:187–202, 1993.

R.D. Lee. Probabilistic Approaches to Population Foeracting. *Population and Development Review*, 24:156–190, 1998. Issue Supplement: Frontiers of Population Forecasting.

R.D. Lee and L.R. Carter. Modeling and Forecasting U.S. Mortality. *Journal of the American Statistical Association*, 87:659–671, 1992.

R.D. Lee and S. Tuljapurkar. *Population Forecasting for Fiscal Plannning: Issues and Innovations, in Demography and Fiscal Policy.* Cambridge University Press, 2000.

R.D. Lee and S. Tuljapurkar. Stochastic population projections for the u.s.: Beyond high, medium and low. *Journal of the American Statistical Association*, 89:1175–1189, 1994.

P.H. Leslie. On the use of matrices in certain population mathematics. *Biometrika*, 33: 183–212, 1945.

O. Lipps and F. Betz. Stochastische Bevölkerungsprojektion für West- und Ostdeutschland. *Zeitschrift für Bevölkerungswissenschaft*, 30,1:3–42, 2005.

W. Lutz and S. Scherbov. Probabilistische Bevölkerungsprognose für Deutschland. *Zeitschrift für Bevölkerungswissenschaft*, 23(2):83–109, 1998.

M. Luy. Leben frauen länger oder sterben männer früher? *Publi Health Forum*, 14(50): 18–20, 2006.

P. Moral and P. González. *Univariate Time Series Modelling in XploRe Application Guide.* Springer Verlag Heidelberg, 2000.

P. Pflaumer. Confidence intervals for population projections based on Monte Carlo methods. *International Journal of Forecasting*, 4:135–142, 1988.

P. Pflaumer. Forecasting US population totals with the Box-Jenkins approach. *International Journal of Forecasting*, 8:329–338, 1992.

A. Renshaw and S. Haberman. Lee-Carter mortality forecasting: a parallel generalized linear modelling approach for England and Wales mortality projections. *Applied Statistics*, 52 (1):119–137, 2003.

Rentenversicherungsbericht. *Bericht der Bundesregierung über die gesetzliche Rentenversicherung, insbesondere über die Entwicklung der Einnahmen und Ausgaben, der Nachhaltigkeitsrücklage sowie des jeweils erforderlichen Beitragssatzes in den künftigen 15 Kalenderjahren gemäß§154 SGB VI.* Bundesministerium für Arbeit und Soziales, 2005.

Rürup-Komission. *Abschlussbericht der Komission für die Nachhaltigkeit in der Finanzierung der Sozialen Sicherungssysteme unter Vorsitz von B. Rürup.* Bundesministerium für Gesundheit und Soziale Sicherung, Berlin, 2003.

Stat.Bundesamt. *Statistisches Jahrbuch 2006, Für die Bundesrepublik Deutschland.* Statistisches Bundesamt Wiesbaden, 2006a.

Stat.Bundesamt. *Bevölkerung Deutschlands bis 2050, Ergebnisse der 11. koordinierten Bevölkerungsvorausberechnung.* Statistisches Bundesamt, Wiesbaden, 2006b.

S. Tuljapurkar. *Population Forecasts, Fiscal Policy, and Risk.* The Levy Economics Institute of Bard College, Working Paper No.471, 2006.

J.R. Wilmoth, K. Andreev, D. Jdanov, D.A. Glei, C. Boe, M. Bubenheim, D. Philipov, V. Shkolnikov, and P. Vachon. *Methods Protocol for the Human Mortality Database*, version 5 edition, May 2007.

K. Wolfsdorf. *Versicherungsmathematik, Teil 1 Personenversicherung.* B.G. Teubner Stuttgart, 1997.

8 Appendix

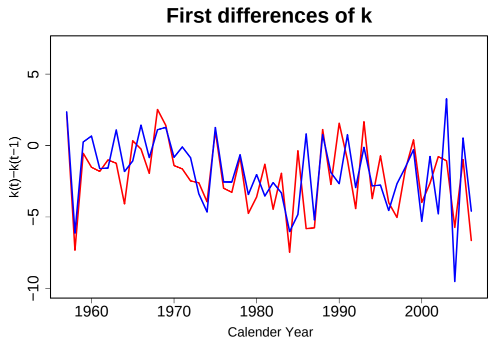

Figure 19: *Process of the first differences of the index k_t for males (blue) and females (red line).*

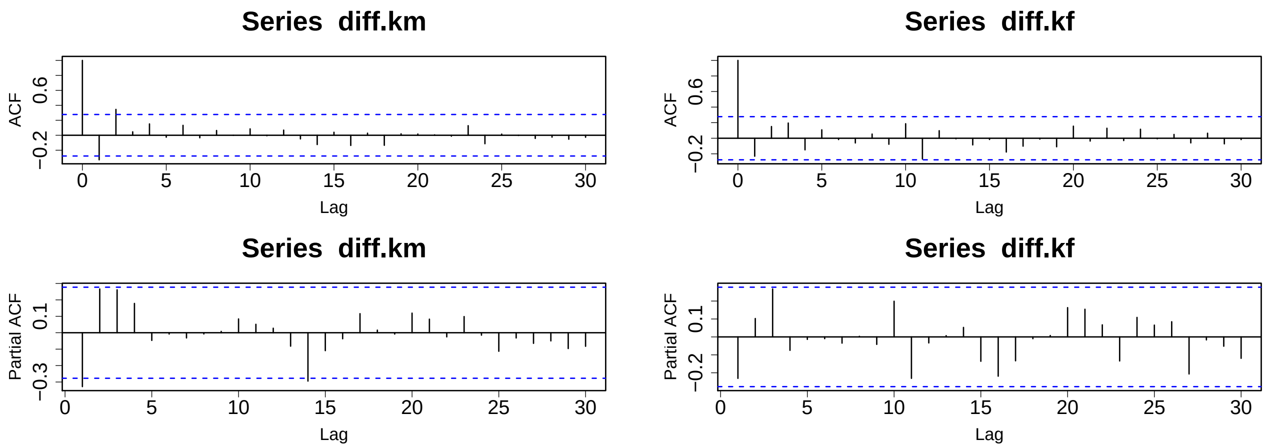

Figure 20: *ACF (top) and PACF (bottom) for the differentiated processes $\widetilde{k}_t^m$ (left) and $\widetilde{k}_t^f$ (right).*

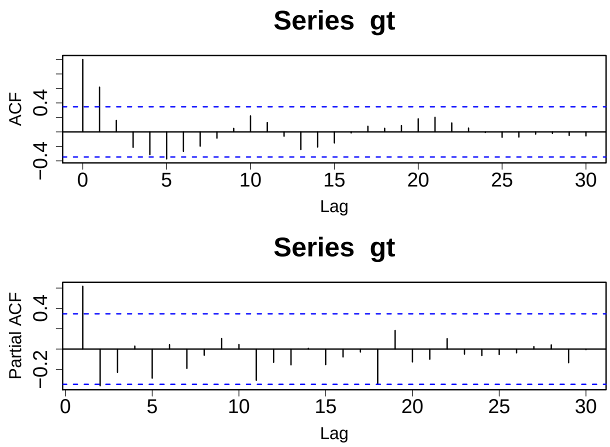

Figure 21: *ACF (top) and PACF (bottom) for g_t.*

Coeff.	Estimate	SE	t-statistic	P(> |t|)
ϕ	0.511	0.185	2.755	0.006 *
θ	0.326	0.161	2.023	0.043 *
δ	-0.958	0.014	-68.066	< 0.001 *

Table 4: *Estimated parameters for g_t as an ARMA(1,1) model.*

Coeff.	Estimate	SE	t-statistic	P(> |t|)
ϕ^m	0.940	0.077	12.194	< 0.001*
ϕ^f	0.938	0.080	11.739	< 0.001*
δ^m	562.848	130.115	4.326	< 0.001*
δ^f	377.927	74.972	5.041	< 0.001*

Table 5: *Estimated parameters for i_t^m and i_t^f as an AR(1) model.*

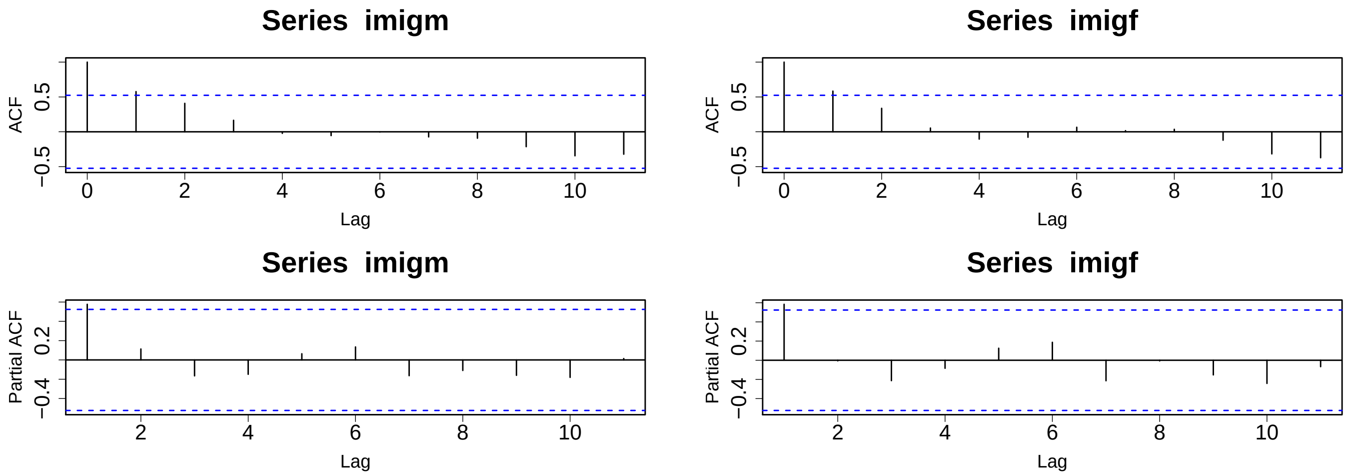

Figure 22: *ACF (top) and PACF (bottom) for immigration processes of males (left) and females (right).*

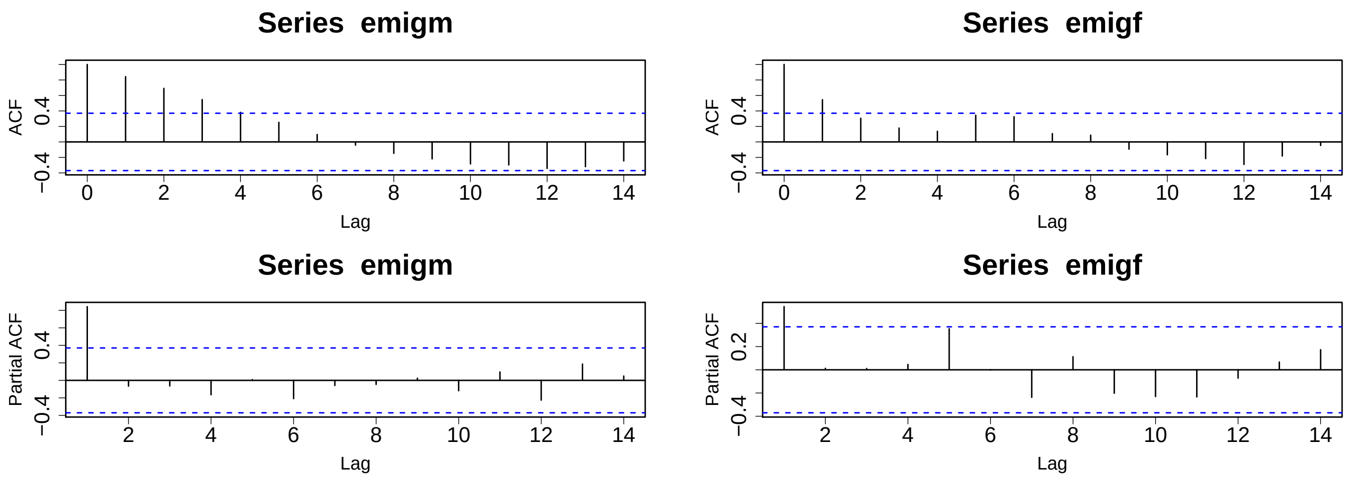

Figure 23: *ACF (top) and PACF (bottom) for emigration processes of males (left) and females (right).*

Coeff.	Estimate	SE	t-statistic	$P(> \|t\|)$
ϕ^m	0.859	0.087	9.843	< 0.001*
ϕ^f	0.600	0.158	3.809	< 0.001*
δ^m	359.757	48.059	7.486	< 0.001*
δ^m	227.554	11.771	19.332	< 0.001*

Table 6: *Estimated parameters for e_t^m and e_t^f as an AR(1) model.*

SFB 649 Discussion Paper Series 2009

For a complete list of Discussion Papers published by the SFB 649, please visit http://sfb649.wiwi.hu-berlin.de.

001 "Implied Market Price of Weather Risk" by Wolfgang Härdle and Brenda López Cabrera, January 2009.
002 "On the Systemic Nature of Weather Risk" by Guenther Filler, Martin Odening, Ostap Okhrin and Wei Xu, January 2009.
003 "Localized Realized Volatility Modelling" by Ying Chen, Wolfgang Karl Härdle and Uta Pigorsch, January 2009.
004 "New recipes for estimating default intensities" by Alexander Baranovski, Carsten von Lieres and André Wilch, January 2009.
005 "Panel Cointegration Testing in the Presence of a Time Trend" by Bernd Droge and Deniz Dilan Karaman Örsal, January 2009.
006 "Regulatory Risk under Optimal Incentive Regulation" by Roland Strausz, January 2009.
007 "Combination of multivariate volatility forecasts" by Alessandra Amendola and Giuseppe Storti, January 2009.
008 "Mortality modeling: Lee-Carter and the macroeconomy" by Katja Hanewald, January 2009.
009 "Stochastic Population Forecast for Germany and its Consequence for the German Pension System" by Wolfgang Härdle and Alena Mysickova, February 2009.

SFB 649, Spandauer Straße 1, D-10178 Berlin
http://sfb649.wiwi.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk".