# Multi-population mortality models: Fitting, Forecasting and Comparisons

Vasil Enchev, Torsten Kleinow and Andrew J. G. Cairns

*Maxwell Institute, Edinburgh and Department of Actuarial Mathematics & Statistics*
*Heriot-Watt University, Edinburgh, EH14 4AS, UK.*

## Abstract

We review a number of multi-population mortality models: variations of the Li and Lee (2005) model, and the common-age-effect (CAE) model of Kleinow (2014). Model parameters are estimated using maximum likelihood. Although this introduces some challenging identifiability problems and complicates the estimation process it allows a fair comparison of the different models. We propose to solve these identifiability problems by applying two dimensional constraints over the parameters. Using data from six countries, we compare and rank, both visually and numerically, the models' fitting qualities and develop forecasting models that produce non-diverging, joint mortality rate scenarios.

It is found that the CAE model fits best. But we also find that the Li and Lee model potentially suffers from robustness problems when calibrated using maximum likelihood.

*Keywords:* Stochastic mortality model, Multi–population, Li and Lee model, common age effect model

## 1. Introduction

Recent decades have seen significant improvements in mortality in most developed countries. It is natural to take the view that changes over time between different countries will, in some way, be correlated. Worsening mortality can be a consequence of epidemics and war which can easily cross national borders. More importantly, improving mortality is the result of improvements in public health, medical advances, lifestyle changes and government regulation, and, whenever improved practice develops in one country, that best practice will rapidly spread to other countries, causing correlated falls in mortality rates. We acknowledge this fact developing multi-population mortality models that are able to assimilate those relationships by modelling simultaneously more than one population. The application of such models is extensive in areas such as reinsurance and risk hedging.

A variety of different models have been considered previously, such as the Li and Lee (2005) model: a multi-population generalisation of the Lee and Carter (1992) model. Li and Lee propose an additional common factor between the multiple populations, and is one of the models considered here.Other authors have also carried out further work on the Li and Lee model, a notable example being Li (2013), who reviews a variation on the original model. Recent work that considers different models from those in this paper include the extensive analyses of Haberman et al. (2014), Li et al. (2015) and Danesi et al. (2015). Haberman et al. (2014) consider a range of two-population mortality models and assess these against a variety of criteria including the coherence of forecasts.[1] Li et al. (2015) consider two-population variants of seven of the models first considered by Cairns et al. (2009). Their analysis aims for balance between historical quality of fit, qualitative model selection criteria and coherence of forecasts. Danesi et al. (2015) consider ten generalisations of the Lee-Carter model and focus more or a detailed discussion of the historical fit. However, their choice of stochastic mortality models and time series models means that forecasts are not coherent. The model Kleinow (2014) proposes - a common age effect (CAE) multi-population model, we examine in this paper more thoughtfully.

As a whole, research on multi-population mortality models, especially at higher ages, above 65 is pretty limited and there is considerable room for further work in this area. We aim to complement this research by reviewing four multi-population mortality models. To compare and order their value, based on their fitting and forecasting qualities, we need to regard a number of points.

Each of the models was fitted under the assumption that the number of deaths

[1] Coherence, in the sense of Li and Lee, 2005, means that, in the long run, mortality rates at the same age in different populations do not diverge.

in a certain population follows a conditional Poisson distribution. For each model we derived a single global likelihood function, maximised using a Newton-Raphson scheme to obtain the maximum likelihood estimate (MLE). In contrast to a singular value decomposition (SVD) or a multi-step MLE, the solution we apply is highly flexible and can be applied to a variety of models. It also provides a consistent framework for further analysis and comparison of models. Nevertheless, this complicates the iterative estimation process and some non-standard identifiability problems arise. We solve this problem by proposing a set of new constraints over the parameters.

Finally, we develop models for forecasting mortality rates in different populations under which future mortality rates in different populations are non-diverging in the long run. In developing these forecasting models, we allow for the use of different time series processes for each of the mortality models under consideration, to ensure that the most appropriate random process is used for each mortality model. Thus we consider a variety of processes including multivariate random walk with drift, multivariate random walk with a common drift, vector autoregressive process and a variation of the vector autoregressive process.

## 2. Data

We perform our analysis based on data obtained from the *Human Mortality Database*, University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on [October 15, 2014]). The typical data set consists of the numbers of deaths and the central exposure. The age/time period range considered is from 60 up to 89 years (30 consecutive ages in total) and from 1961 up to year 2010 (50 years in total).

Table 1 specifies the six countries that were chosen. They were explicitly selected to have similar population sizes and close geographical location. All six countries have experienced significant mortality improvements since 1961.

| | | Exposure to risk at the age of 60 in 2010 | |
|---|---|---|---|
| $i$ | Country | Male | Female |
| 1 | Austria | 47023.17 | 49526.5 |
| 2 | Belgium | 65344.67 | 66434.33 |
| 3 | Czech Republic | 71575.69 | 71575.69 |
| 4 | Denmark | 34420.17 | 35132.33 |
| 5 | Sweden | 59759.83 | 59742.67 |
| 6 | Switzerland | 46527.67 | 47078.67 |

Table 1: Countries considered

## 3. Notation

Throughout this paper we will adopt the following notation:

- $D(x,t,i)$ is the number of deaths, and
- $E(x,t,i)$ is the central exposed to risk

at age $x$ last birthday in calendar year $t$ in country $i$. We then define the crude death rates $\hat{m}(x,t,i)$ as

$$\hat{m}(x,t,i) = \frac{\text{The number of deaths at specific age, time and population}}{\text{The central exposed to risk}} = \frac{D(x,t,i)}{E(x,t,i)}.$$

The crude death rates are assumed to be roughly equal to the underlying force of mortality denoted by $\mu(x,t,i)$, assuming that the force of mortality remains approximately constant over each age $x$ and during each year $t$.

The set of parameters that we seek to estimate for each model is denoted by the vector $\theta$. The components of $\theta$ depend on the particular model under investigation and will be specified when we introduce the models.

Furthermore, we will use the following notation:

- $N$ is the total number of data points that we use in the model (only the number of deaths are included, so $N$ is the number of ages $\times$ number of years $\times$ number of countries);
- $T$ the final year in the time range;
- $k$ the total number of elements in $\theta$;
- $k^{\text{effective}}$ is the total number of parameters $k$ minus the number of model-specific constraints, which will be specified when the models are introduced in section 4.

## 4. Models

### *4.1. Model specification*

We consider parametric models for the logarithm of the underlying death rates, $\log m(x,t,i)$. All models have global parameters which are common to all six populations and local parameters which are specific for each population. This allows us to identify the global characteristics of death rates over age and time across all countries. Using country-specific parameters, we can adapt the estimated global mortality curve to each country's specific mortality data. In that way we

are able to identify relationships between the mortality of populations of different countries.

We propose, test and compare four models: the Li and Lee model introduced by Li and Lee (2005), two variations of this model, and the common age effect model introduced by Kleinow (2014). The two variations of the Li and Lee model provide a significant reduction in the number of parameters $k$, which results in a considerable reduction of computing time and, also, potentially result in more robust forecasts. The Kleinow model (2014) is the most challenging to deal with, due to its different form. Compared to the other models it does not have a common, global period effect, but rather two country specific ones.

Table 2 summarises the proposed models, with differences between them highlighted in bold.

| | |
|---|---|
| Model 0 | $\log m(x,t,i) = \alpha(x,i) + B(x)K(t) + \beta(x,i)\kappa(t,i)$ |
| Model 1 | $\log m(x,t,i) = \alpha(x,i) + B(x)K(t) + \boldsymbol{\beta(x)}\kappa(t,i)$ |
| Model 2 | $\log m(x,t,i) = \alpha(x,i) + \boldsymbol{B(x)}\left(K(t) + \boldsymbol{\kappa(t,i)}\right)$ |
| Model 3 | $\log m(x,t,i) = \alpha(x,i) + \boldsymbol{\beta^1(x)\kappa^1(t,i)} + \boldsymbol{\beta^2(x)\kappa^2(t,i)}$ |

Table 2: The proposed models

Model 0 is a generalisation of the Lee and Carter (1992) model. The country specific parameters, $\alpha(x,i)$, $\beta(x,i)$ and $\kappa(t,i)$, form two-dimensional arrays, while the global age and period effects, $B(x)$ and $K(t)$, are vectors. The common $K(t)$ explains the evolution of global mortality rates over time. The age effect $B(x)$ tells us which rates decline rapidly and which rates decline slowly in response to changes in $K(t)$. The country specific $\alpha(x,i)$ determine the baseline shape of the mortality curve in each country. The country specific parameters $\beta(x,i)$ and $\kappa(t,i)$ have the same interpretation as the common $B(x)$ and $K(t)$, but they are applied specifically to each population. Li and Lee (2005) proposed that the parameters should be estimated using a two step singular value decomposition, firstly estimating the common parameters $B(x)$ and $K(t)$ from the combined data for all countries, and secondly estimating the rest of the country specific parameters.

The purpose of considering models 1 and 2 is to investigate if the number of parameters in the Li and Lee (2005) model can be reduced without significantly compromising the quality of the fit.

Model 1 is a simplified version of Li and Lee in which we are changing the matrix $\beta(x,i)$ into a common parameter vector $\beta(x)$. Although the fit of the model will be compromised, the reduction in the number of parameters is considerable: 150

fewer elements for the 30 ages considered in this paper.

Model 2 is a further simplification as we propose to reduce the parameter space even more. $\beta(x,i)$ is again turned into a common parameter vector, but now it is set to be equal to the common age effect, $B(x)$. In comparison to Model 1, the reduction is by further 30 elements for the 30 ages considered in this paper.

Model 3, the common age effect (CAE) model, is quite different from the other three models. It has two sets of parameters: one of common age dependent parameters, $\beta^1(x)$ and $\beta^2(x)$; and another of country-specific time-dependent parameters, $\kappa^1(t,i)$ and $\kappa^2(t,i)$. For our time period (50 years), this model has more parameters to estimate than the Li and Lee (2005) model, but the difference between the two depends on the balance between the age and year ranges: if the age range was longer than the time period this model would have fewer parameters.

Table 3 reports the number of parameters in each of the models before allowance for identifiability constraints.

| | **Model 0** | **Model 1** | **Model 2** | **Model 3** |
|---|---|---|---|---|
| Number of parameters in $\theta$ | 740 | 590 | 560 | 840 |

Table 3: The total number of parameters to estimate in $\theta$ for each model

*4.2. Identifiability problems*

Each of the models presented in Table 2 exhibits identifiability problems: that is, knowledge of the underlying or theoretical death rate does not uniquely identify the values of the various age and period effects. For example, suppose that the true parameters in Model 0 are $\alpha(x,i)$, $B(x)$, $K(t)$, $\beta(x,i)$ and $\kappa(t,i)$. Then for any scalars $a$, $b \neq 0$, $c$ and $d \neq 0$ the alternative parameter set (1) gives identical death rates.

$$\begin{aligned}
\widetilde{\alpha}(x,i) &= \alpha(x,i) + aB(x) + c\beta(x,i), \\
\widetilde{B}(x) &= \frac{B(x)}{b}, \\
\widetilde{K}(t) &= b(K(t) - a), \\
\widetilde{\beta}(x,i) &= \frac{\beta(x,i)}{d}, \\
\widetilde{\kappa}(t,i) &= d(\kappa(t,i) - c),
\end{aligned} \tag{1}$$

Therefore $K(t)$ and $\kappa(t,i)$ are determined only up to a linear transformation, $B(x)$ and $\beta(x,i)$ are determined only up to a multiplicative constant, and $\alpha(x,i)$ is determined only up to an additive constant. From the reparametrisation (1), it is

clear that it is not possible to uniquely determine the parameters in model 0, and similar issues are found for the other three models. This is known as an identifiability problem, which is very common for mortality models. Such issue normally would not be a problem, because a strict relationship between two parameterisations always exists, therefore we can transfer one set of estimated parameters into another. Usually those relations can even visually be noticed as two parameter sets are slightly tilted or even reversed. The possibility of identifiability, nevertheless, gives rise to two potential problems. First, the parameter estimation process might experience convergence problems if identifiability is not addressed. Second, even if $\theta$ and $\tilde{\theta}$ are parameterisations that give identical *historical* fits, forecast distributions of mortality rates might be different. The second point means that we need to take care when fitting, e.g., a time series model to the period effects to ensure that models allow in a consistent way for the identifiability problem.

The parametrisation (1) is standard in the sense that it is a straightforward extension from the one previously proposed for the Lee and Carter (1992) model. Furthermore, the identifiability problems in Model 0 are completely covered by (1). This is not the case for the other three models, as they are still undetermined.

Due to the more complex form of the models, the period effects, $K(t)$ and $\kappa(t,i)$, are still indistinguishable only up to an additive *time dependent* constant. This issue can be best seen in Model 2. In addition to the parametrisation (1) we can add a scalar, $C(t)$, that would not affect the estimated fitted crude mortality rates, but only the parameter estimates.

Ultimately we can write Model 2 in the form

$$\log m(x,t,i) = \alpha(x,i) + B(x)[\underbrace{K(t) + C(t)}_{\widetilde{K}(t)} + \underbrace{\kappa(t,i) - C(t)}_{\widetilde{\kappa}(t,i)}]. \tag{2}$$

Note that for models 0, 1 and 3 such relationships do not exist.

In what follows, we derive and implement constraints for the parameters that allow us to solve the identifiability issues. To address (1) we normalise the sums of the common age parameter to equate to unity. Additionally, the country specific age dependent parameters should also equate to unity for every population (i.e. the summation of each column $i$ of the matrix should equate to one). Furthermore, the common time dependent parameter $K$ should sum to zero and finally the country specific time dependent parameters $\kappa(.,i)$ should also sum to zero for each population $i$ (i.e. the summation of each row $i$ of the matrix $\kappa(t,i)$ should equate to zero). For (2) the only normalisation we apply is over the country specific time dependent parameter as we set the summation over each year $t$ to be equal to zero (i.e. the summation of each column $t$ of the matrix $\kappa(t,i)$ should equate to zero).

Table 4 summarises all of the constraints applied over the models' parameters. Those constraints fully identify the parameters in the models, but there are some specifics in the application process, which we discuss next.

Model 0 does not require any restrictions over time $t$ for its country specific time dependent parameter $\kappa(t,i)$ and is fully determined. Nevertheless, referring to Table 5, it is evident that the estimation time for this model is considerable. This might mean that this model is close to an identifiability problem, that is, the likelihood function is nearly, but not exactly, flat in some directions. Therefore, it might be possible to derive some additional constraints that would speed up the estimation process.

| **Model 0** | **Model 1** | **Model 2** | **Model 3** |
|---|---|---|---|
| | *Common constraints* | | |
| $\sum_x B(x) = 1$ | $\sum_x B(x) = 1$ | $\sum_x B(x) = 1$ | $\sum_x \beta^1(x) = 1$ |
| $\sum_t K(t) = 0$ | $\sum_t K(t) = 0$ | $\sum_t K(t) = 0$ | $\sum_x \beta^2(x) = 1$ |
| - | $\sum_x \beta(x) = 1$ | - | - |
| | *Country specific constraints* | | |
| For each $i$: | | | |
| $\sum_x \beta(x,i) = 1$ | - | - | $\sum_t \kappa^1(t,i) = 0$ |
| $\sum_t \kappa(t,i) = 0$ | $\sum_t \kappa(t,i) = 0$ | $\sum_t \kappa(t,i) = 0$ | $\sum_t \kappa^2(t,i) = 0$ |
| | *Time specific constraints* | | |
| For each $t$: | | | |
| - | $\sum_i \kappa(t,i) = 0$ | $\sum_i \kappa(t,i) = 0$ | $\sum_i \kappa^2(t,i) = 0$ |
| - | Quasi identifiability constraint | True identifiability constraint | Quasi identifiability constraint |

Table 4: Identifiability constraints for models 0, 1, 2 and 3.

The time specific constraint in Model 2 for $\kappa(t,i)$ is noted as a "True identifiability constraint". The reason for this is that applying the constraint does not change the value of the log-likelihood function.

For models 1 and 3 the time specific constraints in Table 4 are noted as "Quasi identifiability constraint", since they do not solve exact identifiability issues. The optimal set of parameters for those models is unique without imposing the time specific constraints. However, the likelihood function seems to be so flat that the applied Newton-Raphson algorithm converges to different optimal parameter sets depending on the starting value for the algorithm. To address this issue we decided

to impose those "quasi constraints". Therefore, we find a constrained maximum of the likelihood function rather than the optimal solution.

Although this reduces the attained maximum value of the log-likelihood function, the reduction (by around 80) is considered small relative to the differences of the log-likelihood values between the models, see Table 8. The Newton-Raphson algorithm now converges to the same maximum when different starting values are chosen, and therefore, the estimation procedure is numerically more stable. For those reasons we have chosen to incorporate the "quasi-constraints" for models 1 and 3. For the complete implementation of the constraints, see 8.

*4.3. Fitting the models*

As mentioned earlier, Li and Lee (2005) propose a two-step singular value decomposition for Model 0 to estimate parameters. For the other three models, this method cannot be applied. The main reason for this is the existence of the bilinear terms: common parameters multiplied by country specific parameters (the first is a vector and the second is a matrix). In contrast, maximum likelihood based estimation can be applied to all four models and, therefore, gives consistent results. Additionally, it comes with a variety of tools to compare the models such as the Bayesian Information Criterion (BIC) or the likelihood ratio test (for the nested models 0, 1 and 2 only). Using the conditional Poisson assumption, we propose to use single step maximum likelihood.

More precisely, we assume that the number of deaths, $D(x,t,i)$, has a Poisson distribution:

$$D(x,t,i)|\theta \sim Poisson(E(x,t,i)m(x,t,i)).$$

The log-likelihood function: $l(\theta|D(x,t,i),E(x,t,i))$, for all of the models has the form

$$l = \sum_{x,t,i} \left[D(x,t,i)\log(m(x,t,i)) - E(x,t,i)m(x,t,i)\right] + \text{constant}. \tag{3}$$

The constant term in (3) is independent of $\theta$ and therefore it does not influence the estimation of the parameters, but it is used in the calculations of the BIC. Its form is

$$\text{constant} = \sum_{x,t,i} \left[D(x,t,i)\log(E(x,t,i)) - \log(D(x,t,i)!)\right].$$

The function (3) is optimised for $\theta$ using a standard Newton-Raphson method with appropriate starting points for the parameters. The process iterates to the optimal values until the log–likelihood value converges. Each iteration cycles sequentially through each element of a model's parameter vector, $\theta$, subject to each model's identifiability constraints.

The calculation time is an important consideration. Since the Newton-Raphson algorithm is an iterative method, it is highly dependent on the chosen starting points. The further they are from the optimal solution the longer it takes to converge. Distant starting values might cause convergence to be slow, whereas a well chosen, close starting value could result in much faster convergence. In general, though, the choice of model has a much bigger influence on the speed of convergence (see Table 5).

Using the constraints in Table 4, we tested a range of different starting points and Table 5 provides the typical observed number of iterations until the convergence level was reached for each model.

| | **Model 0** | **Model 1** | **Model 2** | **Model 3** |
|---|---|---|---|---|
| Number of iterations | 500 | 10 | 5 | 40 |

Table 5: Typical number of iterations until the convergence level was reached from different initial values of $\theta$ for each model.

Comparing the number of iterations in Table 5 with the number of elements to estimate in Table 3, it is evident that the size of $\theta$ is not the determinant factor for the speed of convergence, since Model 3 has the highest number of elements and Model 2 the smallest. The most likely reason for the slow convergence of Model 0 is that its log-likelihood function is flat or almost flat in certain dimensions and since the Newton-Raphson method is based on the tangent, it takes considerably higher number of iterations to reach the maximum. This issue could be addressed by implementing "quasi-constraints" for model 0.

## 5. Comparison of the fitted models

Several aspects of the parameter estimates and the model fits are of interest: the shape of the parameters, the quality of the fit, patterns in the residuals, calculation speed (number of iterations until convergence) and robustness of the models. In the following, we compare the four models based on those properties.

The fitted parameter shapes strongly depend on the constraints imposed, which becomes more obvious in the graphs that follow. This implies that the constraints are a determining factor when deciding what time series process to use for generating mortality scenarios.

In a multi-population model, we expect a certain behaviour from its parameters. In a perfect situation the common parameters should be able to capture the true global mortality trend, in both age and time, amongst the populations as a whole. If the underlying philosophy of the model is correct, then we would expect that the country specific period effects all fluctuate around some constant level in the long term. Significant differences from this level, in either range or shape would mean that this particular population is somehow different from the other populations and there should be a reason for this behaviour. The maximum likelihood estimation is highly influenced by the population size and the countries with bigger size should have a higher impact when estimating the common mortality parameters.

The estimated parameters are shown in Figure 1. Due to the different magnitude of the parameters, we did not try to apply the same scale for every graph.

We notice in Figure 1 that the parameters for models 1 and 2 look similar: there are some differences between the common $\beta(x)$ in Model 1 and $B(x)$ in Model 2, but not enough to give rise to significant differences between the $\kappa(t,i)$. Evidently the proposed change for Model 2 is not significant in comparison to Model 1 and it hardly effects the estimated parameters. Essentially this would lead to small differences in the log-likelihood values between the models.

In all four models, due to the constraints applied (e.g. $\sum_x B(x) = 0$), the country specific parameter $\alpha(x,i)$ is approximately equal to the average over time of the crude mortality rates i.e.

$$\alpha(x,i) \approx \log \frac{\sum_t D(x,t,i)}{\sum_t E(x,t,i)}.$$

Differences exist, because of the non-linear form of the log-likelihood function, but these are not easily visible.

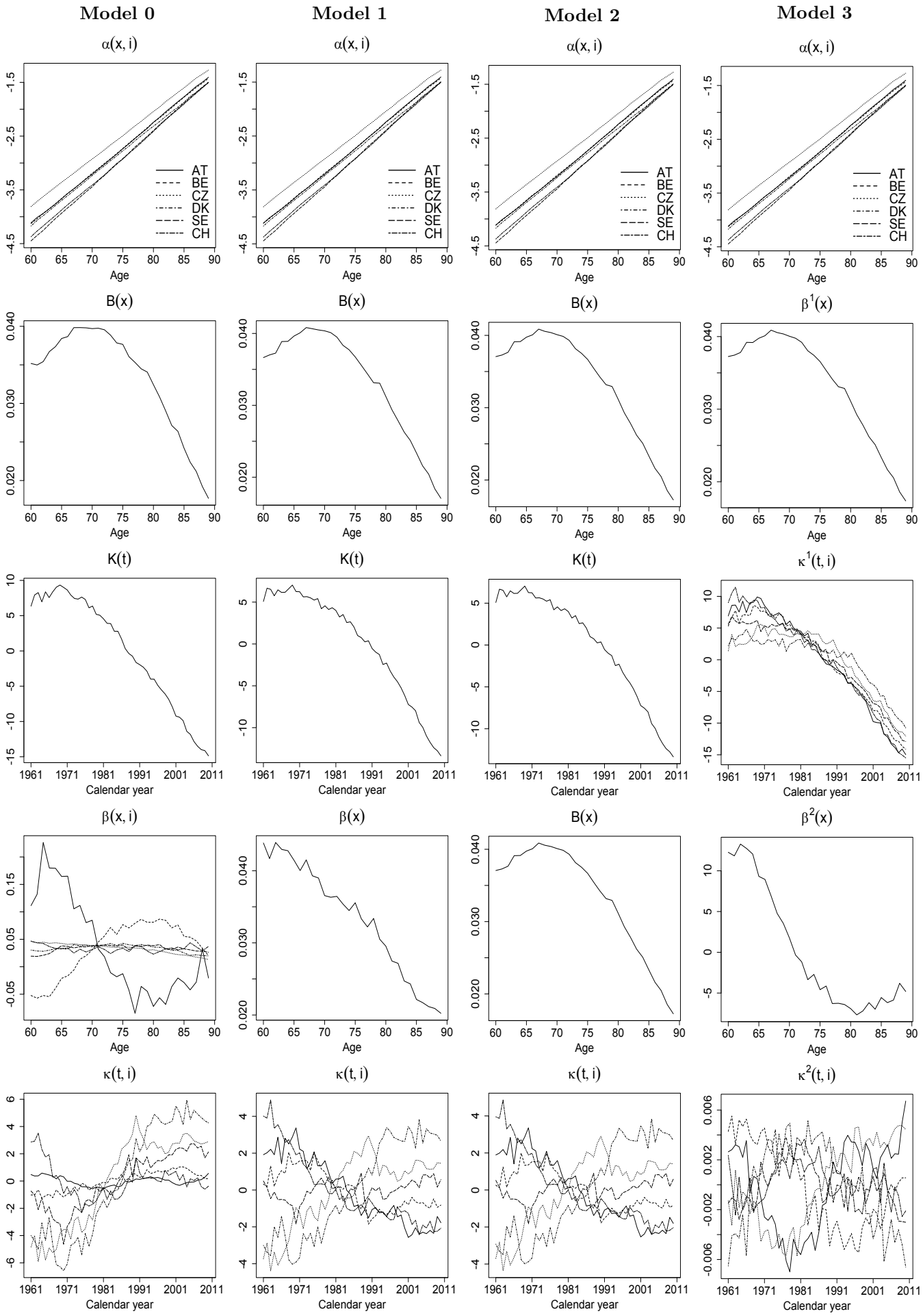


Figure 1: The estimated parameters for all of the models 0 - 3 [Males]

The common time parameter $K(t)$ for models 0, 1 and 2 is strictly decreasing after 1970, showing consistent improvement in the global mortality of the populations. Model 3 does not have a common time dependent parameter, but the country specific $\kappa^1(t,i)$ evidently behave in a similar way to $K(t)$. Comparing models 0 and 3, we also note that $B(x)$ in Model 0 and $\beta^1(x)$ in Model 3 are quite similar. Alongside this, the country-specific period effects $\kappa^2(t,i)$ in model 3, while individually different, have a similar shape and trend to the common period effect, $K(t)$ in model 0. We conclude, therefore, that $K(t)$ in Model 0 picks up most of the improvements in the six countries.

Additionally, the shape of $B(x)$ and $\beta^1(x)$ in the models is consistent with the empirical observation that the mortality rates have declined more rapidly at younger ages than older ages.

Models 1, 2 and 3 have another common age parameter, decreasing over the ages. For Model 1, $\beta^1(x)$ is quite similar in shape to $B(x)$. However, the "value added" through having the additional age effect relative to Model 2 comes through the difference in the shape of the second age effect below age 70. Also, Model 3's $\beta^1(x)$ is negative after the age of 72 indicating that the mortality at those ages tends to increase when falling at other ages. In this sense, therefore, $\beta^1(x)$ in Model 3 is associated with changes in the overall level of mortality, while $\beta^2(x)$ links in to changes in the slope of the mortality curve.

As expected, all of the common parameters behave similarly, which is an indication that the models capture the true global trend.

For Model 0, we see a country-specific behaviour of the $\beta(x,i)$. Austria and Belgium have country specific age effects that change sign and exhibit much more variability than the other four countries. However, the latter is an artefact of the constraint that $\sum_x \beta(x,i) = 1$ which acts in a different way if the $\beta(x,i)$ change sign. Indeed, we see that the $\kappa(t,i)$ are less variable for Austria and Belgium to compensate.

More generally, we can note that the country-specific period effects are all varying around zero in all four models, in part reflecting the choice of identifiability constraints.

We also find for model 3 that the $\kappa^2(t,i)$ exhibit the typical behaviour of an autoregressive mean reverting process for each country $i$. The scale is considerably smaller in comparison to the other model's $\kappa(t,i)$ and to compensate for this its $\beta^2(x)$ parameter is substantially bigger than $\beta(x,i)$, $\beta(x)$ and $B(x)$ for models 0, 1 and 2.

### *5.1. Explanation ratios*

We compare the quality of the fit numerically by calculating, for each population, in percentage the common explanation ratio – $R_C$ and the country specific explanation ratio – $R_{AC}$, defined by Li and Lee (2005) and shown in Table 6.

| | |
|---|---|
| $R_C$ | $100\left[1-\frac{\sum_{x,t}[\log\hat{m}(x,t,i)-\alpha(x,i)-B(x)K(t)]^2}{\sum_{x,t}[\log\hat{m}(x,t,i)-\alpha(x,i)]^2}\right]$ |
| $R_{AC}$ | $100\left[1-\frac{\sum_{x,t}[\log\hat{m}(x,t,i)-\alpha(x,i)-B(x)K(t)-\beta(x,i)\kappa(t,i)]^2}{\sum_{x,t}[\log\hat{m}(x,t,i)-\alpha(x,i)]^2}\right]$ |

Table 6: Explanation ratios for model 0.

The notation used in Table 6 is only for Model 0, but it is easy to derive the formulas for the other models.[2] The better the quality of the fit the better the models explain the historical mortality data. We aim for a high explanatory percentage and values below 90% might be considered weak. Such populations, with weak explanation ratios, are very specific and they should be reviewed separately. Additionally we do not try to correct the jump-off bias at the end of the fit, as Lee and Miller (2001) did for example, because this would mean re-estimating the time dependent parameters and this would drift from the assumption of Poisson distributed deaths.

In Table 7 we show the calculated explanation ratios for all models. The denominator (Table 6) already includes the baseline, country-specific age effect, $\alpha(x,i)$, so $R_C$ and $R_{AC}$ provide us with information on how much of an improvement there is in the fit of the model over all years relative to a static age-dependent mortality model.

| | **Model 0** | | **Model 1** | | **Model 2** | | **Model 3** | |
|---|---|---|---|---|---|---|---|---|
| | $R_C$ | $R_{AC}$ | $R_C$ | $R_{AC}$ | $R_C$ | $R_{AC}$ | $R_C$ | $R_{AC}$ |
| Austria | 96 | 97 | 92 | 96 | 92 | 96 | 96 | 97 |
| Belgium | 95 | 97 | 94 | 96 | 86 | 94 | 96 | 97 |
| Czech Republic | 57 | 94 | 81 | 93 | 73 | 81 | 92 | 95 |
| **Denmark** | **5** | **89** | **58** | **88** | **42** | **88** | **89** | **91** |
| Sweden | 87 | 97 | 96 | 97 | 93 | 96 | 97 | 97 |
| Switzerland | 95 | 97 | 91 | 97 | 91 | 97 | 97 | 97 |

Table 7: The explanation ratios (%) for the models

[2] Specifically, $R_C$ includes in the numerator the common age-period effect for models 1 and 2, and the first factors, $\beta^1(x)\kappa^1(t,i)$, for model 3. $R_{AC}$ includes all country specific effects in the numerator.

We can see from Table 7 that Model 3 performs best across all countries. In some countries, the inclusion of the country-specific component does not seem to add much (e.g. Austria). However, for others, the country-specific component is essential for getting a good fit.

Specifically, through all of the models, the explanation ratios for Denmark are smaller than the rest of the countries. One reason for this could be the smaller exposure (see Table 1). Additionally, although its mortality rates have been decreasing over time, the pace of improvement has been much slower in comparison to the other populations. A major reason for this slow mortality improvement might be the excessively high number of deaths from non-transmissible disease such as cancer and heart disease, see for example the Health at a Glance 2011: OECD Indicators (2011).

Also note that the explanation ratios $R_C$ for models 1 and 2 would be identical if a two step estimation method had been used to find the optimal parameters. However, since we maximise the full likelihood function, the changes made to the second term in those models has an effect on the estimated values of $B(x)$ and $K(t)$.

*5.2. Standardized residuals*

We also compare the quality of fit visually, by calculating the standardized residuals $Z(x,t,i)$

$$Z(x,t,i) = \frac{D(x,t,i) - E(x,t,i)m(x,t,i)}{\sqrt{E(x,t,i)m(x,t,i)}}$$

and plotting their heat maps. As stated by Cairns at al., (2011), if the model fits the data well, then the standardized residuals should be independent of each other, meaning that the heat plot should exhibit a high degree of randomness, with no discernible patterns.

Heat plots of the residuals for each model and country are shown in Figure 2. We find that Switzerland and Sweden come the closest to what we hope to find as noted above. In particular, Model 3 gives a very random looking plot for Switzerland. However, in contrast to what we would hope to see, we can observe patterns in most of the heat plots. Most obviously there is a significant cohort effect present in several countries: Austria, Belgium and Czech Republic.

Additionally, less visible patterns can also be recognised. As a whole, the models produce very robust, very similar heat plots. We can also note that, where patterns can be detected, they tend to be more pronounced in models 1 and 2, which have fewer parameters than the other two models. As noted above, Sweden and Switzerland produce the most random plots, and it is not surprising, therefore, that these two countries also have a lower empirical variance for the standardized residuals.

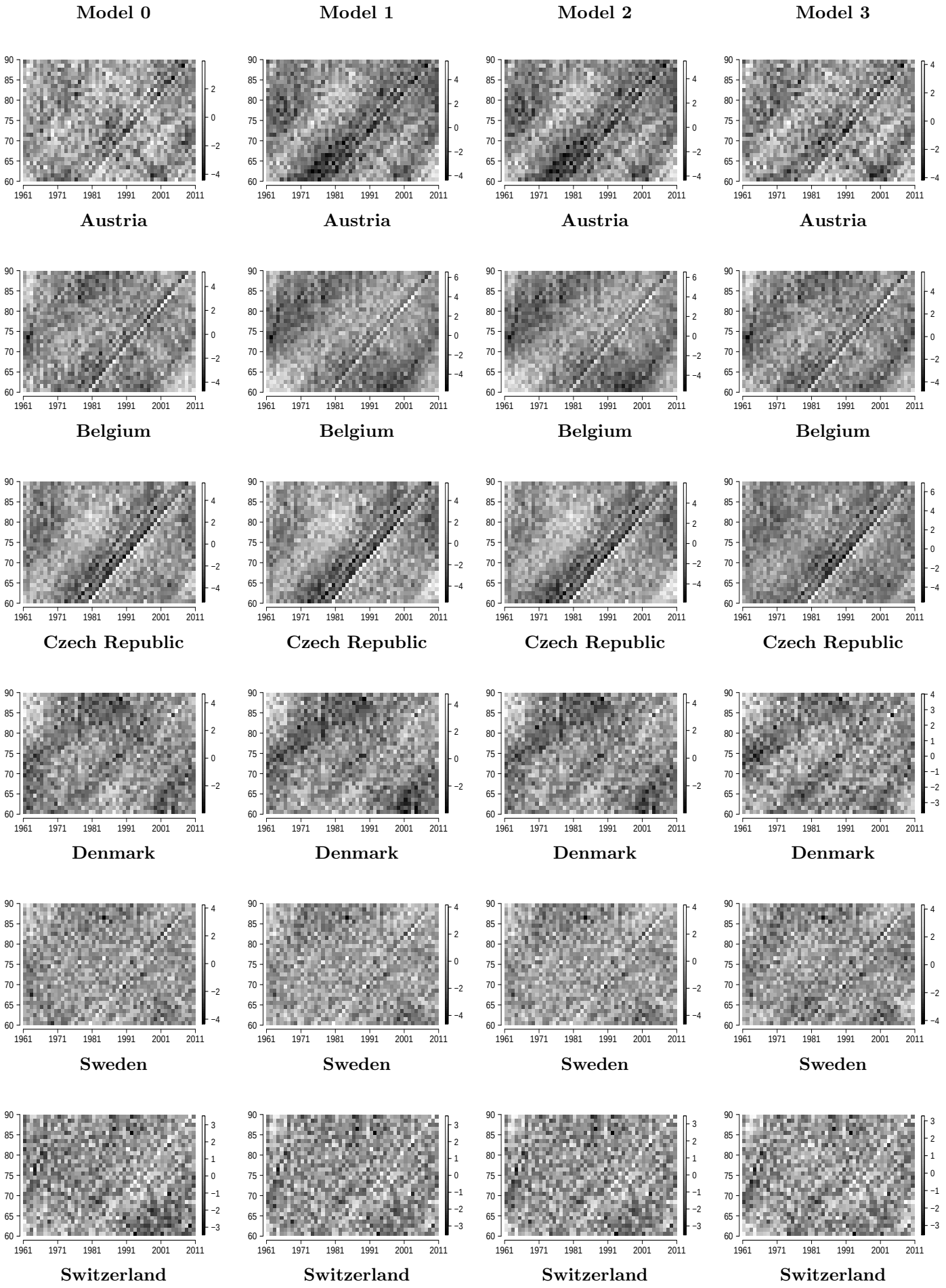


Figure 2: The standardized residuals heat plots for models 0 - 3 [Male]

*5.3. Model selection*

Finally, we rank the models based on their Bayesian Information Criterion value defined as

$$\text{BIC value} = -2[\text{log-likelihood value}] + \log(N)k^{\text{effective}}.$$

In Table 8 we estimate the BIC and rank the models.

| | N | k | $k^{\text{effective}}$ | log-likelihood | BIC | rank |
|---|---|---|---|---|---|---|
| **Model 0** | 9000 | 740 | 726 | -46716.43 | 100043.08 | ( 2 ) |
| **Model 1** | 9000 | 590 | 531 | -48396.82 | 101628.38 | ( 4 ) |
| **Model 2** | 9000 | 560 | 502 | -48477.21 | 101525.12 | ( 3 ) |
| **Model 3** | 9000 | 840 | 776 | -46369.96 | 99805.38 | ( 1 ) |

Table 8: The rank of the models based on their BIC value

Based on the BIC definition, the smaller its value the higher we rate the model. Therefore the best model, based on the fitting qualities, is Model 3. Additionally, as noted previously, the log-likelihood value of Model 1 and Model 2 is close and this confirms the results from the parameter plots in Figure 1. We have tested the four models with different data (for females and with an extra country that is significantly different from the rest). The results were consistent and conclusive as the ranks of the models never changed.

Note that the values of the log-likelihood function for models 1 and 3 are reduced due to the "quasi-constraints" applied for those models as we mentioned in section 4.2. Without the "quasi-constraints" the BIC values for those models would be even lower showing that the common age effect model (model 3) has indeed a significantly lower BIC value than the Li and Lee model (Model 0) for the data in this study.

## 6. Forecasting mortality

Ultimately, we aim to generate joint scenarios for the mortality rates of all six countries by extrapolating both the common and the country specific time dependent parameters of the models. Models 0, 1 and 2 are very similar in those terms as they all have a common $K(t)$ parameter and another country specific $\kappa(t, i)$ parameter. Model 3 presents a different challenge, due to its two country specific parameters $\kappa^1(t, i)$ and $\kappa^2(t, i)$.

*6.1. Initial investigation*

Li and Lee propose to extrapolate the common and the country specific period parameters in Model 0, respectively with a random walk with drift and an autoregressive process. Our initial analysis is focused on whether those processes are appropriate for the models considered in this paper (Table 2). More precisely, in the next Table 9, we summarise the applied initial time series processes for each model and their parameters.

| **Model** | **Parameter** | **Formula** | **Process** |
|---|---|---|---|
| Model 0 | $K(t)$ | $K(t) = d + K(t-1) + \mathcal{E}(t)$ | Random walk with drift |
| Model 1<br>Model 2 | $\kappa(t,i)$ | $\kappa(t,i) = \varphi(i)\kappa(t-1,i) + \mathcal{Z}(t,i)$ | Autoregressive process reverting to 0 |
| Model 3 | $\kappa^1(t,i)$ | $\kappa^1(t,i) = \varphi(i)\kappa^2(t-1,i) + \mathcal{Z}^1(t,i)$ | Autoregressive process reverting to 0 |
| | $\kappa^2(t,i)$ | $\kappa^2(t,i) = \varphi(i)\kappa^2(t-1,i) + \mathcal{Z}^2(t,i)$ | Autoregressive process reverting to 0 |

Table 9: The initial time series processes used for the common and the country specific time dependent parameters for models 0, 1, 2 and 3.

In an ideal world there would be no correlations between those processes residuals structure. In other words we expect that the correlations of $\mathcal{A}$=corr($\mathcal{E}(t)$,$\mathcal{Z}(t,i)$) for models 0, 1 and 2 and $\mathcal{B}$=corr($\mathcal{Z}^1(t,i)$,$\mathcal{Z}^2(t,i)$) for Model 3 should be very weak. This, in practice, is not the case – for our data we observe a considerable linear dependence – and the possibility of correlation between the random innovations is potentially important in terms of its impact on the correlations in the short and long term between mortality improvements in different countries. For Model 0 the correlations $\mathcal{A}$ are very high – above 80%, for Model 1 and Model 2 they are at an average level – around 50% and finally for Model 3 the correlations $\mathcal{B}$ are below 30%, which is considered weak.

*6.2. Scenarios for Period Effects*

Choosing a suitable time series process, is an essential step in the forecasting. Clearly, we wish to use time series models that are strongly influenced by the historical shape of the parameter estimates. However, the shape (Figure 1) is highly influenced by the constraints that we have applied, see Table 4.

Clearly, from Figure 1, the common $K(t)$ parameters for models 0, 1 and 2 exhibit a significant downwards trend after 1972 pointing to strong improvements in mortality in all of the populations. Therefore we do not expect that there would be any appreciable changes in the improvement rate (consistent, for example, with the ideas of Oeppen and Vaupel (2002)),although some researchers state that

future improvement rates might be weaker (see, for example, Olshansky et al., (2005) ).

For models 0, 1 and 2 we model the $K(t)$ process as a random walk with drift, meaning that the central forecast would be defined by this strong historical trend. For Model 3 the country specific $\kappa^1(t,i)$ processed seem, in general, to follow the trend observed in the other models. Therefore we model it as a multivariate random walk, which, additionally, has a common drift parameter to ensure that the different populations do not diverge too quickly.

Modeling the multivariate country specific time dependent parameters presents a more interesting challenge for all of the models. We want to ensure that forecasts are coherent between the populations and do not diverge significantly over time. We can achieve this by applying multivariate time series processes, that are able to capture the correlations between the countries. After some initial investigation, the time series processes chosen for the models are specified in Table 10.

| **Model** | **Parameter** | **Formula** | **Process** |
|---|---|---|---|
| Model 0 | $K(t)$ | $K(t) = d + K(t-1) + \sigma\mathcal{E}(t)$ | Random walk with drift |
| | $\kappa(t,i)$ | $\kappa(t) = \Phi\kappa(t-1) + C\mathcal{Z}(t)$ | Vector Autoregression process reverting to 0 |
| Model 1 | $K(t)$ | $K(t) = d + K(t-1) + \sigma\mathcal{E}(t)$ | Random walk with drift |
| | $\kappa(t,i)$ | $\kappa(t) = d_\kappa + \kappa(t-1) + C\mathcal{Z}(t)$ | Multivariate random walk with drift |
| Model 2 | $K(t)$ | $K(t) = d + K(t-1) + \sigma\mathcal{E}(t)$ | Random walk with drift |
| | $\kappa(t,i)$ | $\kappa(t) = d_\kappa + \kappa(t-1) + C\mathcal{Z}(t)$ | Multivariate random walk with drift |
| Model 3 | $\kappa^1(t,i)$ | $\kappa^1(t) = d_c + \kappa^1(t-1) + C\mathcal{Z}^1(t)$ | Random walk with common drift term |
| | $\kappa^2(t,i)$ | $\kappa^2(t) = \Phi\kappa^2(t-1) + C\mathcal{Z}^2(t)$ | Vector Autoregression process reverting to 0 ($\Phi$ is a diagonal matrix) |

Table 10: The final time series processes used for the common and the country specific period effects for models 0, 1, 2 and 3. $\kappa(t) = (\kappa(t,1), \ldots, \kappa(t,6))'$ is a vector of country-specific period effects, $\Phi$ is an autoregression matrix, $C$ is a volatility matrix, $d$ is the scalar drift for the common period effect, $d_\kappa$ is a vector of drifts, $d_c$ is a vector of common drifts (i.e. all components are equal), $\mathcal{E}(t)$ is a scalar sequence if i.i.d. standard normal innovations, and $\mathcal{Z}(t)$ is a sequence of i.i.d. standard multivariate normal innovations that are independent of the $\mathcal{E}(t)$.

In Figure 3 we plot for each model the extrapolated country specific time dependent parameters using the specified time series processes.

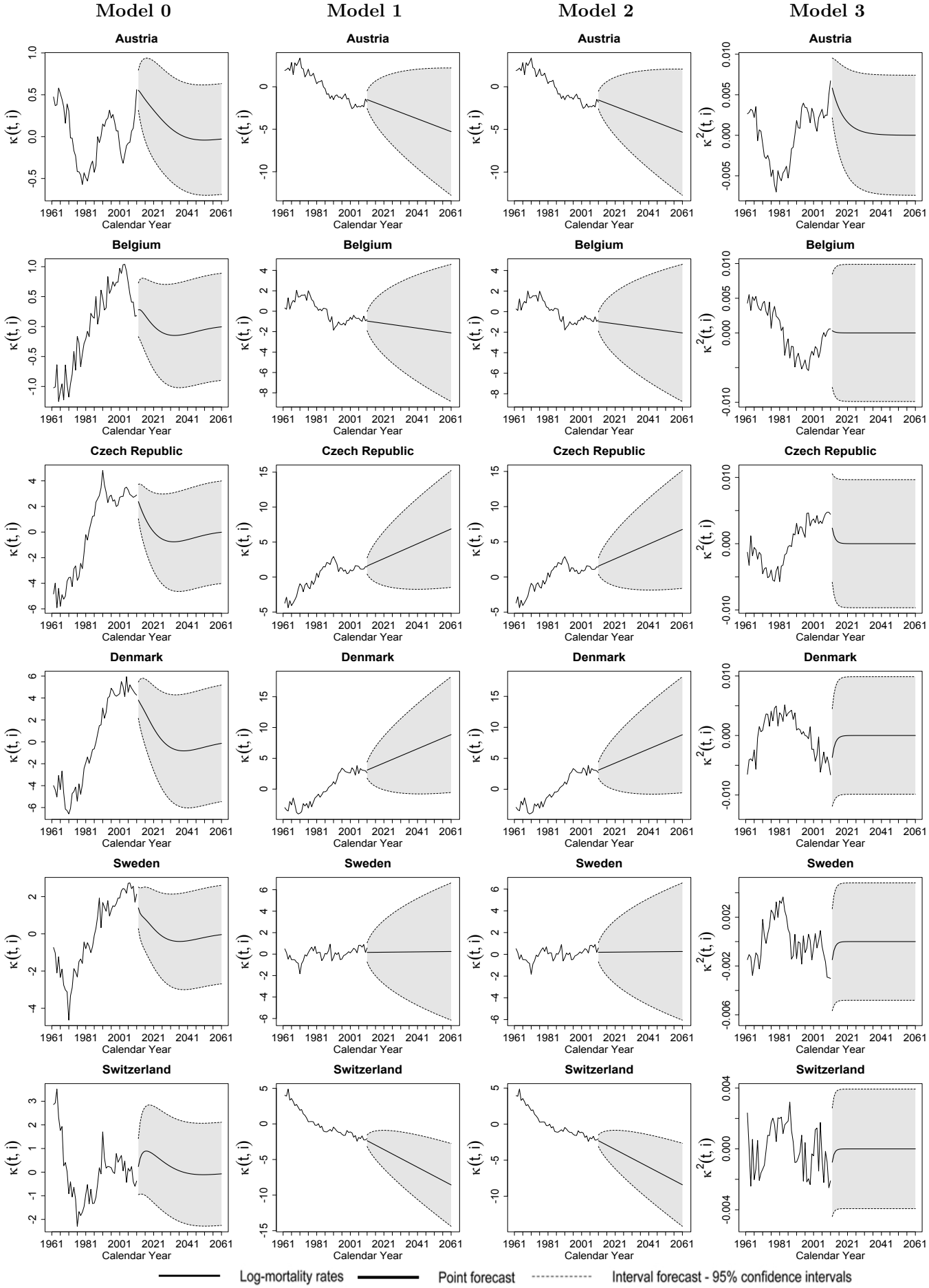


Figure 3: Forecasted country specific time dependent parameters for 50 years for Model 0 - Vector autoregression process, Model 1 - Multivariate random walk with drift, Model 2 - Multivariate random walk with drift, Model 3 - Vector autoregression process ($\Phi$ is diagonal matrix)

In Figure 3, the first column represents individually the country specific parameters for each population under Model 0. We observe a lot of variation between the historical country-specific period effects, with no strong common trends. The choice of time series model (a multivariate mean reverting VAR(1) process) means that we, in a simple way, achieve the desired property that the period effects fluctuate around zero in the future. Specifically, it allows us to produce coherent non-diverging mortality rate forecasts for all of the populations. Additionally, if we calculate the eigenvalues for the $\Phi$ matrix in the VAR(1) process we can see that all are less than one, so the process is stationary. Therefore, the evolution over time of the individual variances in the forecast mortality rates, will mainly be dominated in the long run by the random walk for the common period effect, $K(t)$. Those properties and the fairly easy implementation of the VAR(1) process makes it a natural choice for our data.

The central line forecasts for some countries, like Denmark, Czech Republic or Switzerland are significantly curved, reflecting the structure of the autoregressive matrix, $\Phi$, and the consequent interaction of the other country effects. The 95% confidence intervals reflect the interaction between $\Phi$ and the volatility matrix, $C$.

The focus is now on the second and third column on Figure 3. As we mentioned before, Model 1 and Model 2 behave in a very similar way and this is why we review them together in this section. Going back to Figure 1 we notice that the most significant difference between the parameter estimates for those models are observed for the age parameters $\beta(x)$ and $B(x)$ and, even those roughly follow the same trend. Therefore, the forecasted mortality rates will look very similar.

In figure 3 we find that all of the curves historically seem to follow a specific trend, which is an important factor when considering the suitable time series process. In comparison to Model 0, where the curves were just fluctuating around the zero, this behaviour is very different. The trend is strongly down for Austria and Switzerland; and strongly up for the Czech Republic and Denmark. It is these differences that led to the use of a multivariate random walk with country specific drifts rather than the VAR(1) model and this is reflected in the central trends and wider fans in the long run. The model, therefore, allows for diverging mortality rates, conflicting with the usual criterion (see, for example, Li and Lee, 2005, or Cairns et al., 2011) the mortality rates in different populations should not diverge over time.

Finally, the time specific constraints that we applied for those models (Table 4) mean that for any country $i$ the country-specific period effect $\kappa(t, i)$ is a linear combination of the $\kappa$ processes of the other five countries, which reduces the rank of the $6 \times 6$-dimensional variance-covariance matrix of the joint $\kappa(t)$ process to five. Therefore one of its eigenvalues is always close to zero, so the Cholesky decomposition applied in the estimation process might fail due to some rounding

errors. We do not want to give any special treatment to a single country and to be consistent, we propose to always use a variation of the Cholesky decomposition. We decompose the covariance matrix$_{(6\times6)}$ into the product of a unique lower triangular matrix$_{(6\times5)}$ and its transpose in contrast to the original Cholesky decomposition where the unique lower triangular matrix is $6 \times 6$. More details are provided in 8.

For Model 3, the historical $\kappa^2(t,i)$ (Figure 3, right hand column) exhibit more consistent mean reversion in all countries compared to Model 0. This means that the model can capture effectively the main improving trend in each country using the first set of parameters, $\beta^1(x)$ and $\kappa^1(t,i)$. The second bilinear term, $\beta^2(x)\kappa^2(t,i)$, then captures the small variations around that trend. We have chosen a VAR(1) process, since it can account for those factors and ensures coherent forecasts between the populations. Additionally, we propose to simplify the VAR(1) process by reducing the number of parameters to estimate. More precisely, we set the $\Phi$ matrix to be diagonal and the values on the diagonal are not all equal. Effectively, this reduces the parameters to estimate by 30 for this process. The extrapolated curves closely resemble those we would obtain when applying individual AR(1) processes for each country, but due to the covariance matrix in the VAR(1) process we can preserve the coherent forecast. This process is stationary. The eigenvalues of the variance-covariance matrix, for this model and for our data, are all higher than zero, therefore it has full rank of 6. The "Quasi identifiability" constraints do not lead to problems like those encountered in Model 1 and Model 2, but if for a different data set the rank is lower, the partial case of the Cholesky decomposition will effectively solve this issue.

### *6.3. Mortality Scenarios*

Based on the scenarios for the period effects we can now generate scenarios for future mortality rates. In Figure 4 we plot forecast intervals for the log-mortality rates for each model and each country. The future mortality forecasts are plotted from the end of the fitted curves rather than the historical mortality values. For all four models we find that the higher the age the narrower is the forecast confidence interval. This reflects the age effects $B(x)$, $\beta(x,i)$ etc., which all tend to decrease in magnitude as a function of age $x$.

The mortality forecasts for Model 1 and Model 2, unsurprisingly, are very similar. The future log-mortality rates they predict are strongly influenced by the country specific parameters. The strong local upward trend in $\kappa(t,i)$ (Figure 3) for countries like Denmark and the Czech Republic are affecting the forecast shifting it up, worsening the mortality improvements over time, while for the other countries the opposite can be observed. As a whole Model 1 and Model 2 forecasts are very wide and do not seem to follow the general in sample mortality trend.

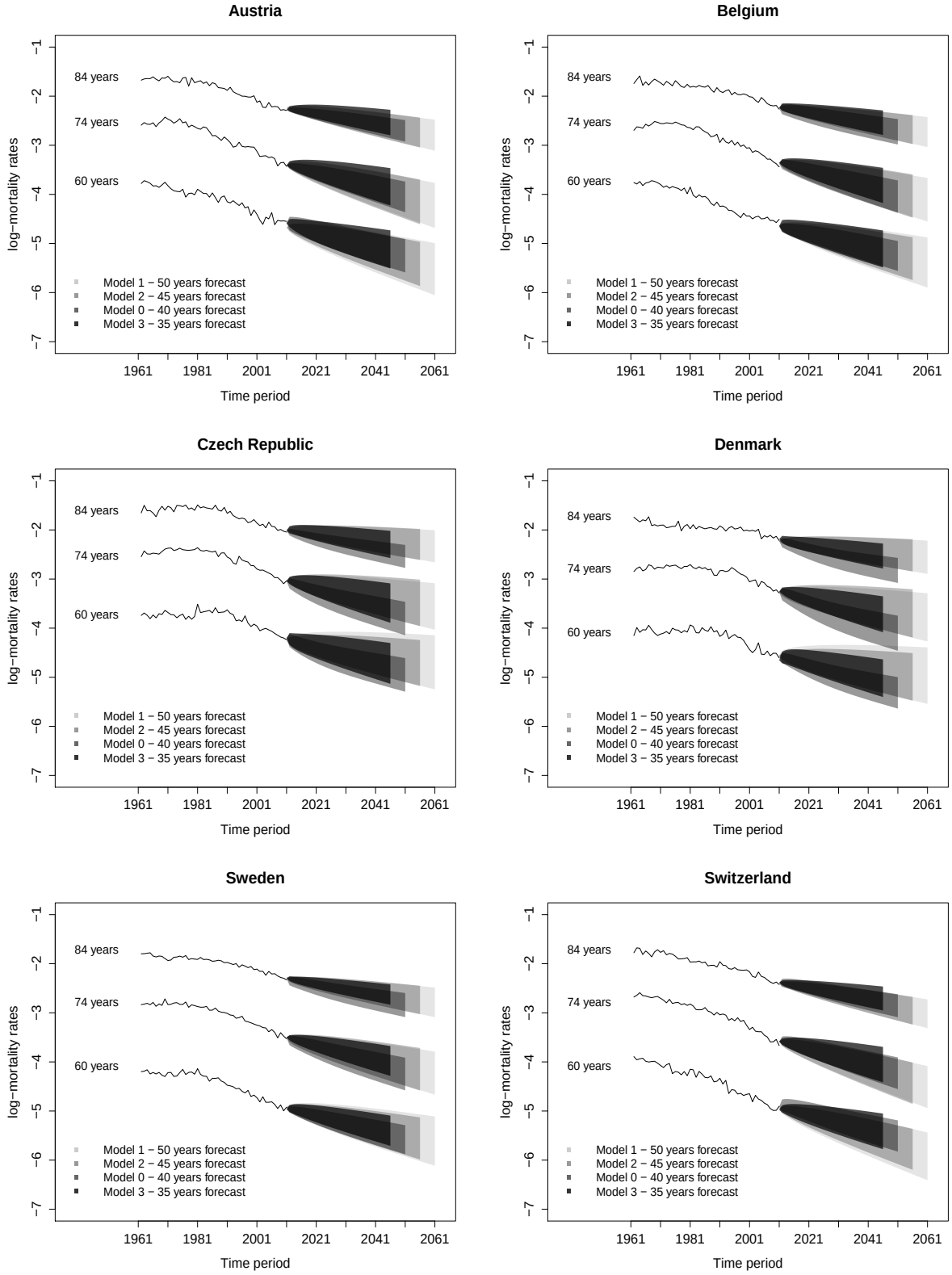


Figure 4: Male log-mortality rates - $\log m(x,t,i)$ for Austria, Belgium, Czech Republic, Denmark, Sweden and Switzerland forecasted into the future using Model 0 (40 years period), Model 1 (50 years period), Model 2 (45 years period) Model 3 (35 years period)

In contrast to models 1 and 2, Model 0 produces forecasts that seem to follow the in sample mortality trend better, and its confidence intervals are smaller due to the stationary VAR(1) process. It can be seen for Denmark and the Czech Republic that the log-mortality rates are a bit curved due to the irregular shape of its country specific parameters.

Model 3's forecasts also seem to follow the in sample mortality trend, but in comparison to Model 0, it tends to predict more conservative mortality improvements in the future. The major reason for this is the common drift applied in the multivariate random walk (Table 9), when extrapolation the $\kappa^1(t,i)$ parameter. Its cumulative value is higher when for example using country specific drifts. This effectively shifts up the future parameter estimates resulting in more conservative mortality forecasts. The forecasts reflect the structure of the diagonal autoregressive matrix, $\Phi$, and, since there is no interaction of the other country effects, they are very straight. Also the 95% confidence intervals are typically bigger than the forecast intervals for Model 0.

## 7. Robustness

In this section we discuss how robust the estimates of the parameters are in relation to changes in the time period used to fit the models. Our analysis is based on data for the countries in Table 1 with the time period being reduced by 10 years. Therefore we now fit the models to data of 30 ages (from 60 to 89) and 40 years (from 1961 to 2000). Figures 6, 7 and 8 present the estimated parameters from the full data set and the reduced data set.

For all models we have estimated the following cases outlined in more detail in 8:

(A) one step MLE using the complete (non–reduced) dataset,

(B) one step MLE using the reduced dataset,

(C) one step MLE using the reduced dataset where the initial parameter values are the final estimates from 2 step based MLE.

One characteristic of a robust model is that estimates of historical age and period effects should not be too sensitive to changes in the range of ages or years used to calibrate the model. Parameter estimates should change a bit because we have different amounts of data to work with, but they should not jump to a solution that is qualitatively quite different. Based on this limited experiment, Models 1, 2 and 3 appear to be robust (Figures 7 and 8). Due to the similarities between Model 1 and Model 2, their robustness plots are very similar and they can not be visually distinguished. Therefore we only consider ones the plot for

both of them - Figure 7. For Model 3, for example, (Figure 8) when we compare the full dataset (solid line) with the reduced dataset (dots) we see that all of the estimated age and period effects qualitatively have a similar shape. Some systematic differences can be seen that are, in reality, the consequences of the identifiability constraints being applied over a different range of years (e.g. the $\kappa^1(t,i)$). Also, it needs to be noted that the estimated $\beta^2(x)$ has been rescaled in the plot with a counterbalancing rescaling of the $\kappa^2(t,i)$. The reason for this is that, because $\beta^2(x)$ changes sign, the constraint that the $\beta^2(x)$ sum to 1 makes it more sensitive to small changes in the underlying shape. A different constraint could be applied (e.g. $\sum_x \left(\beta^2(x)\right)^2 = 1$) that might be less sensitive, but this was not felt to be necessary and does not affect the quality of fit or have any impact on the forecasts of mortality rates.

In contrast, Model 0 did exhibit robustness problems. First, focus on the solid and dashed lines in each on the sub-plots in Figure 6: Cases A and B using the same one-step MLE algorithm. Apart from the $\alpha(x,i)$, all of the plots exhibit significantly and qualitatively different shapes for the parameter estimates based on the two datasets.

One notable feature is that, for Case B, the $\beta(x,i)$ are all quite similar to the global $B(x)$ plot. If, in fact, the $\beta(x,i)$ were all equal to $B(x)$ we would have a troubling identifiability problem: we could take an arbitrary function $\varepsilon(t)$ and replace $\kappa(x,i)$ by $\tilde{\kappa}(t,i) = \kappa(t,i) + \varepsilon(t)$, and $K(t)$ by $\tilde{K}(t) = K(t) - \varepsilon(t)$ with no impact on the fitted values of $\log m(x,t,i)$. Thus we see that the distinctive shape of $K(t)$ for the reduced dataset is almost cancelled by the shapes of the $\kappa(t,i)$.

To investigate this robustness problem further we tried estimating the Model 0 parameters using an alternative 2-step approach (see 8). The second step of this maximises the same log-likelihood function as Case B and so should converge to the same solution. However, for the reduced dataset it does not, it converges to a quite different solution (Figure 6, dotted curves) that is much more like the solid curve (Case A, full dataset).[3] A possible explanation for this is that the full likelihood function for the Li and Lee model has a problem with multiple maxima: a global maximum in Case B that has a log-likelihood that is 155 higher than the alternative local maximum in Case C.

To investigate the multiple maximum question further, let $\theta_B$ and $\theta_C$ be the two vectors of parameter estimates and define

$$l(\rho) = \text{log-likelihood}\left(\theta_B + \rho(\theta_C - \theta_B)\right). \tag{4}$$

We then plot the log-likelihood as it evolves along a straight line between the

[3] Note that, we also applied the 2-step procedure to the full dataset, but this produced the same parameter estimates as Case A.

two local maxima. $l(\rho)$ is shown in Figure 5, and we can see a very substantial difference in the likelihood between $\theta_B$ and $\theta_C$. The shape observed is consistent with the function having multiple maxima, although it does not prove it. But it does point to the need for additional quasi-identifiability constraints for Model 0.

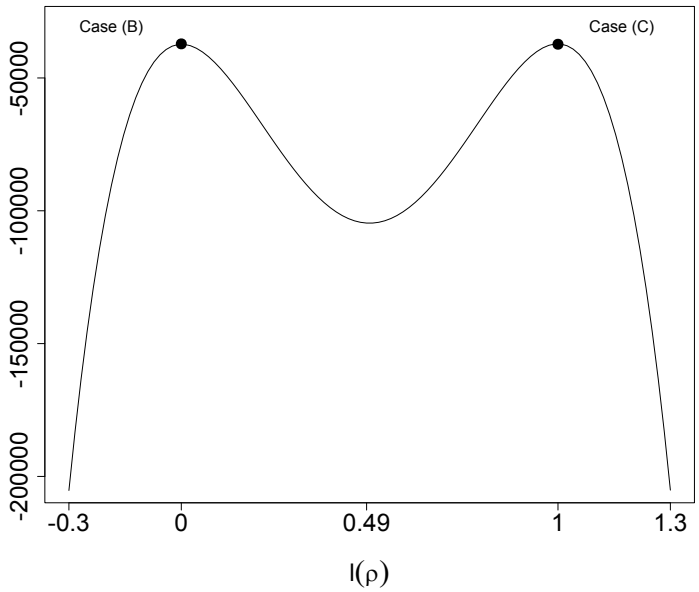


Figure 5: The log-likelihood function for Model 0, $l(\rho)$ along the straight line connecting $\theta_B$ (Case B) and $\theta_C$ (Case C).

**Model 0**

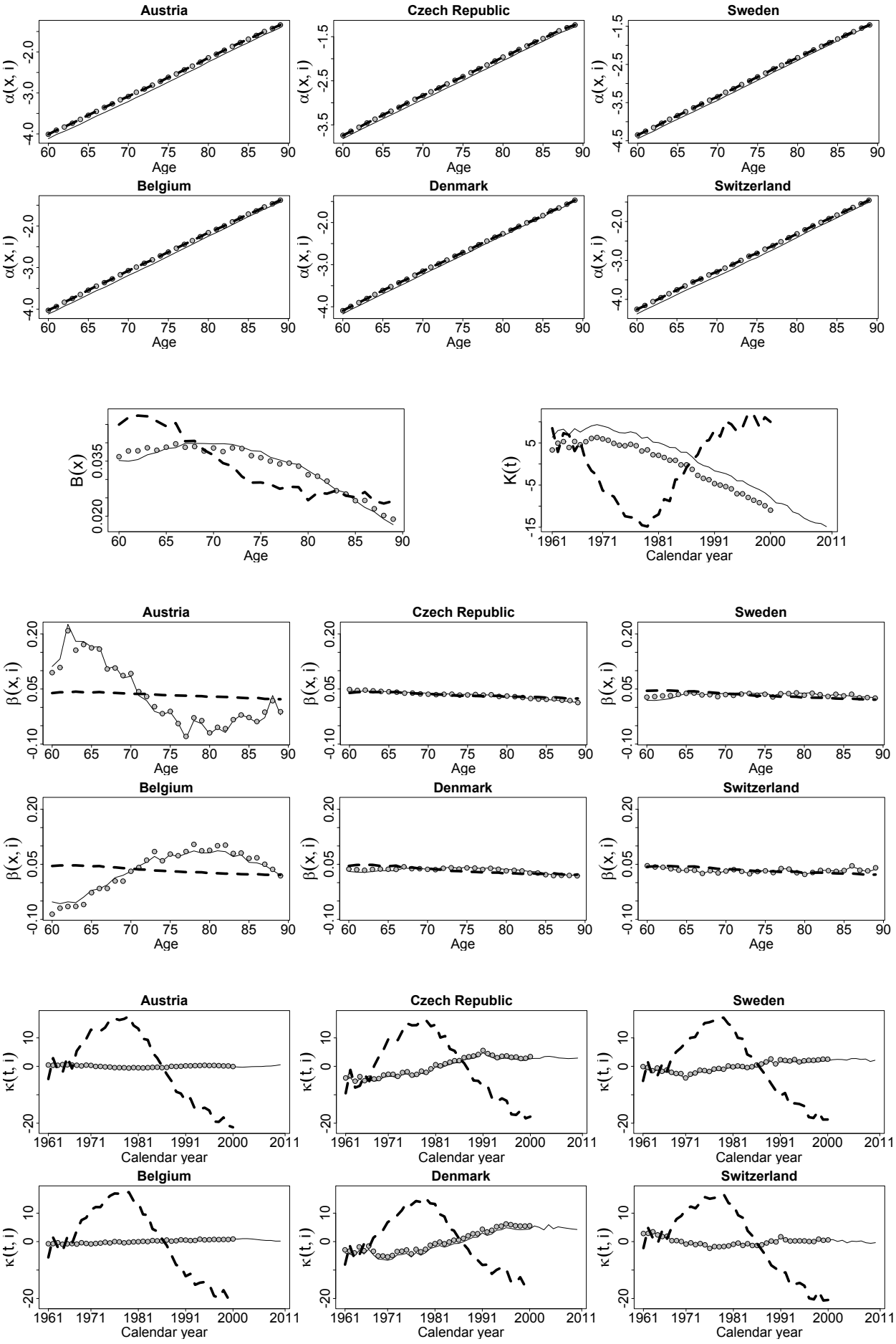


Figure 6: Estimated parameters for Model 0: Case (A) – the solid line; Case (B) – the dashed line; Case (C) – the dotted line. No scaling is applied.

**Model 1**

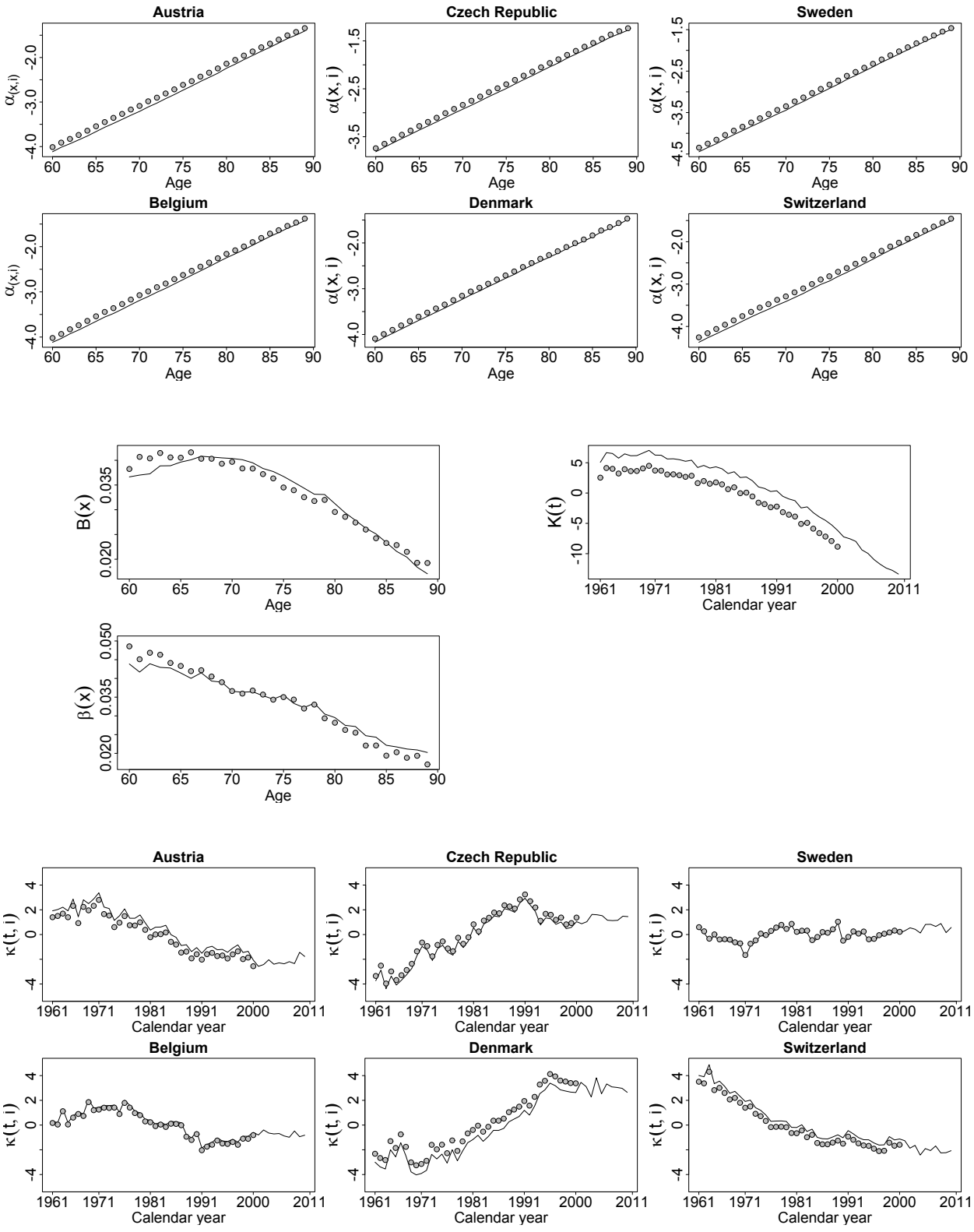


Figure 7: Estimated parameters for Model 1: Case (A) – the solid line; Case (B) – the dotted line. No scaling is applied.

**Model 3**

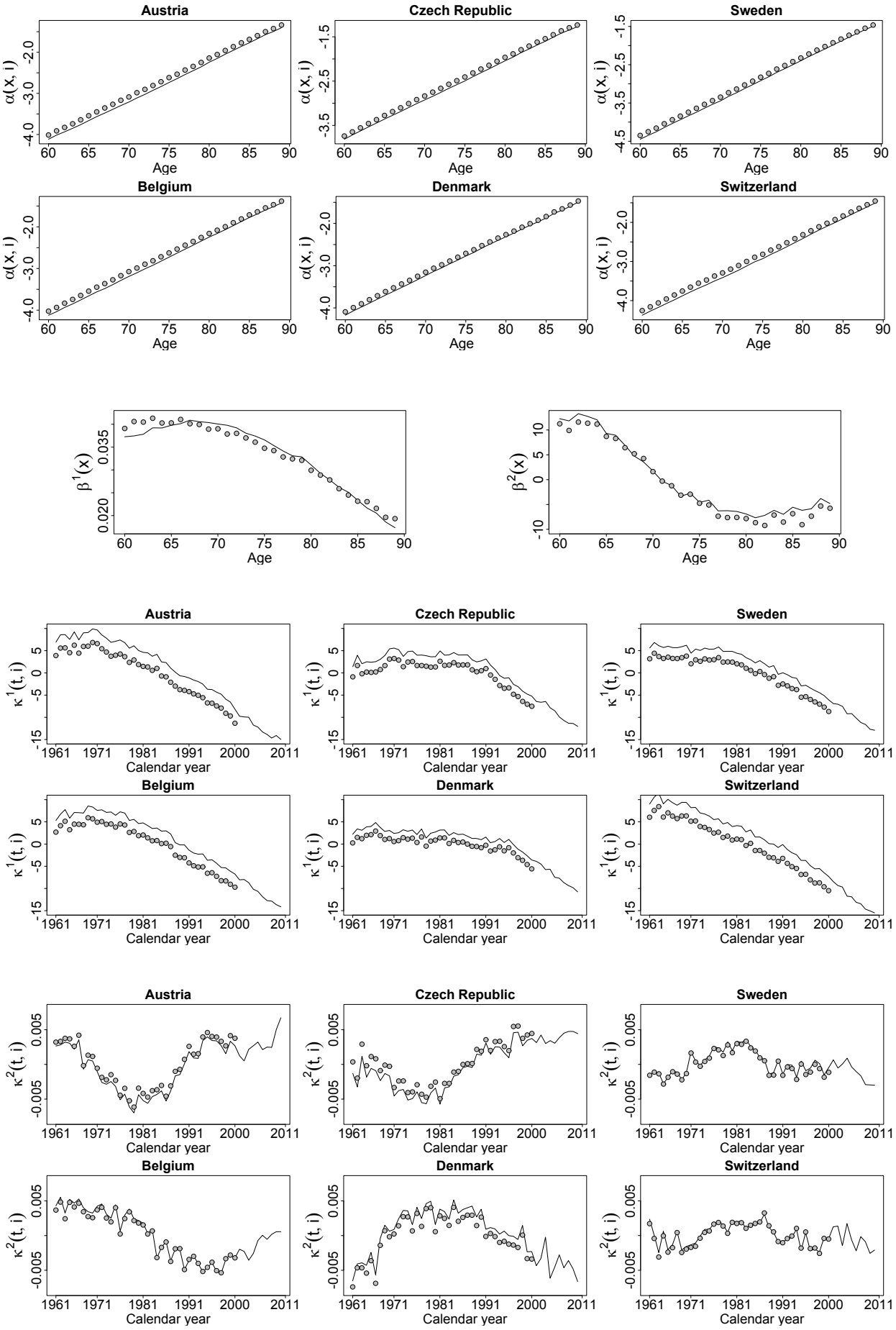


Figure 8: Estimated parameters for Model 3: Case (A) – the solid line; Case (B) – the dotted line, where $\kappa^2(t,i)$ is multiplied by -1/26 and $\beta^2(x)$ multiplied by -26.

## 8. Conclusion

We compared the multi-population mortality model by Li and Lee (2005) and two of its variants with the Common Age Effect (CAE) model by Kleinow (2014) using mortality data from six countries. All models have identifiability problems that have been addressed by applying exact and quasi identifiability constraints. The use of maximum likelihood estimation allows a fair comparison of the models on the basis of the Bayesian Information Criterion as well as graphical diagnostics. We find that the simplified special cases of the Li and Lee model do not perform well in sample, while the full Li and Lee and CAE models produce reasonable results, with the CAE model performing in a more satisfactory way against several of the criteria. Plots of residuals point to a need for additional cohort effects in some countries.

The Li and Lee model fitted the data for the six countries quite well, but exhibited robustness problems along with associated problems with slow convergence. This points to a need for caution in the use of that model and, potentially, for the introduction of some quasi identifiability or other constraints.

We also developed forecasting models with satisfactory results for both the Li and Lee and CAE models. The different models produced somewhat different forecasts in terms of central trajectories and the amount of uncertainty, pointing to model risk as an important consideration in an overall assessment of future levels of mortality.


## Acknowledgements

Vasil Enchev acknowledges financial support from Heriot-Watt University through the award of a James Watt scholarship and from the Actuarial Research Centre of the Institute and Faculty of Actuaries. Torsten Kleinow and Andrew J. G. Cairns acknowledge financial support from Netspar under project LMVP 2012.03.


## Appendix A

## Calculation notes on the models

### I step MLE

### Initial values

The starting points chosen for $B(x)$, $\beta(x)$, $\beta(x,i)$, $\beta^1(x,i)$ and $\beta^2(x,i)$ should be different from zero. Any other set of values, close to the solution is acceptable for the rest of the parameters.

| **Using the Newton-Raphson method** | **Model** |
|---|---|
| 1. Estimate $\alpha(x,i)$ | [ All ] |

| | | | |
|---|---|---|---|
| 2. | Dependent on the model | • Estimate the global time parameter $K(t)$ | [ 0,1,2 ] |
| | | • Estimate the time/country-specific $\kappa^1(t,i)$ | [ 3 ] |
| 3. | Dependent on the model | • Estimate the global age parameter $B(x)$ | [ 0,1,2 ] |
| | | • Estimate the global age parameter $\beta^1(x)$ | [ 3 ] |
| 4. | Dependent on the model | • Estimate the time/country-specific $\kappa(t,i)$ | [ 0,1,2 ] |
| | | • Estimate the time/country-specific $\kappa^2(t,i)$ | [ 3 ] |
| 5. | Dependent on the model | • Estimate the age/country-specific $\beta(x,i)$ | [ 0 ] |
| | | • Estimate the global $\beta(x)$ parameter | [ 1 ] |
| | | • Estimate the age/country-specific $\beta^2(x,i)$ | [ 3 ] |

**Apply the constraints for the country specific parameters**

| | | | |
|---|---|---|---|
| 6. | Update the values of $\alpha(x,i)$ | | [ All ] |
| 7. | Dependent on the model | • Apply the constraints for $\kappa(t,i)$ | [ 0,1,2 ] |
| | | • Apply the constraints for $\kappa^1(t,i)$ | [ 3 ] |
| | | • Apply the constraints for $\kappa^2(t,i)$ | [ 3 ] |
| 8. | Dependent on the model | • Apply the constraints for $\beta(x,i)$ | [ 1 ] |

**Apply the time specific constraints**

| | | | |
|---|---|---|---|
| 9. | Dependent on the model | • Apply the new constraints for $\kappa(t,i)$ | [ 1,2 ] |
| | | • Apply the constraints for $\kappa^2(t,i)$ | [ 3 ] |

*Note: Every time a constraint is applied we remove or divide by something, but to preserve the same model, we need to multiply or add immediately by the same thing. Up to this point those changes were absorbed by the other parameters in the models. The new constraint does not require to correct the change occurred by dividing or removing value to preserve the model. In fact it has no effect whether we add those values to the other parameters (Model 2) or it is not possible to add them (models 1 and 3).*

**Apply the constraints for the global parameters (only once)**

| | | | |
|---|---|---|---|
| 10. | Dependent on the model | • Apply the constraints for $K(t)$ | [ 0,1,2 ] |
| | | • Apply the constraints for $B(x)$ | [ 0,1,2 ] |
| | | • Apply the constraints for $\beta(x)$ | [ 1 ] |

- Apply the constraints for $\beta^1(x)$ [ 3 ]
- Apply the constraints for $\beta^2(x)$ [ 3 ]

11. Return to step 1 until the log-likelihood value converges.
    - Convergence is deemed to have occurred when the change in the log-likelihood is less than 0.0001.

*Note: Also the order of the steps is essential and it does matter.*

### Two Step MLE

The two step based maximum likelihood estimation can be applied only for Model 0 and it is impossible for the other models due to the presence of the bilinear terms: common parameter multiplied by a country specific one.

1. Using the Lee and Carter model, define:

$$\log m(x,t) = A(x) + B(x)K(t)$$

and then estimate the parameters $A(x)$, $B(x)$ and $K(t)$ for the combined data set of all of the populations. The log-likelihood function $l$ of the model is

$$l = \sum_{x,t} [D(x,t)\log(m(x,t)) - E(x,t)m(x,t)] + \text{constant}$$

which is then optimised using the Newton-Raphson iterative scheme. The $A(x)$ parameter is used only for the estimation process and it does not affect the estimation any further. Also the standard constraints are imposed i.e. $B(x)$ should sum to unity and $K(t)$ to zero.

2. Using the estimated $B(x)$ and $K(t)$ as fixed, constant values, we apply MLE for Model 0 **Using the Newton-Raphson method**

   2.1 Estimate $\alpha(x,i)$

   2.2 Estimate $\kappa(t,i)$

   2.3 Estimate $\beta(x,i)$

   **Apply the constraints for the country specific parameters**

   2.4 Update $\alpha(x,i)$

   2.5 Apply the constraints for $\kappa(t,i)$

   2.6 Apply the constraints for $\beta(x,i)$

   Return to step 2.1 until the log-likelihood value converges (threshold level is set to 0.0001).

## Appendix B

**Cholesky decomposition (partial case)**

The time specific constraints (Table 4) over the local $\kappa(t,i)$ parameter for Model 1 and Model 2 diminishes the rank of the variance–covariance matrix by one, and, therefore, one of the eigenvalues is zero. Therefore, the standard Cholesky decomposition can not be applied. For our data we have a $6 \times 6$ variance–covariance matrix, but its rank is 5. The solution that we apply is a partial Cholesky decomposition as we decompose the variance–covariance matrix into a product of a unique lower triangular matrix of dimension $6 \times 5$ and its transpose. In the general case ($\text{matrix}_{n\times n}$) we have:

$$V_{n\times n} = \begin{bmatrix} v_{1,1} & v_{1,2} & \dots & v_{1,n} \\ v_{2,1} & v_{2,2} & \dots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n,1} & v_{n,2} & \dots & v_{n,n} \end{bmatrix} = L_{n\times n-1} L^t_{n-1\times n} \equiv$$

$$\equiv \begin{bmatrix} \boldsymbol{l_{1,1}} & & & & \\ l_{2,1} & \boldsymbol{l_{2,2}} & & \mathbf{0} & \\ l_{3,1} & l_{3,2} & \boldsymbol{l_{3,3}} & & \\ \vdots & \vdots & \vdots & \ddots & \\ l_{n-1,1} & l_{n-1,2} & l_{n-1,3} & \dots & \boldsymbol{l_{n-1,n-1}} \\ l_{n,1} & l_{n,2} & l_{n,3} & \dots & l_{n,n-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{l_{1,1}} & l_{2,1} & l_{3,1} & \dots & l_{n-1,1} & l_{n,1} \\ & \boldsymbol{l_{2,2}} & l_{3,2} & \dots & l_{n-1,2} & l_{n,2} \\ & & \ddots & \vdots & \vdots & \vdots \\ & \mathbf{0} & & \boldsymbol{l_{n-2,n-2}} & l_{n-1,n-2} & l_{n,n-2} \\ & & & & \boldsymbol{l_{n-1,n-1}} & l_{n,n-1} \end{bmatrix}.$$

Therefore by multiplying $L_{n\times n-1} L^t_{n-1\times n}$ and equating the result to $V$, we can derive relationships for the elements of the $L$ matrix.

We can see that for the diagonal elements ($l_{i,i}$) there is a calculation pattern:

$$l_{i,i} = \sqrt{v_{i,i} - \sum_{j=1}^{i-1} l^2_{i,j}},$$

for the elements below the diagonal ($l_{i,k}$, where $i > k$) there is also a calculation pattern:

$$l_{i,k} = \frac{1}{l_{k,k}} \left( v_{i,k} - \sum_{j=1}^{k-1} l_{i,j} l_{k,j} \right).$$

The only different calculation pattern is for the $l_{n,n-1}$-th element:

$$l_{n,n-1} = \sqrt{v_{n,n} - (l^2_{n,1} + l^2_{n,2} + \ldots + l^2_{n,n-2})}.$$

## References


[1] Brouhns, N., Denuit, M. and Vermunt, J. K. (2002), 'A poisson log-bilinear regression approach to the construction of projected lifetables', *Insurance: Mathematics and Economics* **31**, 373393.

[2] Cairns, A. J. G., Blake, D. and Dowd, K. (2006), 'A two-factor model for stochastic mortality with parameter uncertainity: theory and calibration', *The Journal of Risk and Insurance* **73**(4), 687–718.

[3] Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A. and Balevich, I. (2009), 'A quantative comparison of stochastic mortality model using data from england and wales and the united states', *North American Actuarial Journal* **13**(1).

[4] Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D. and Khalaf-Allah, M. (2011), 'Stochastic mortality modelling for two populations', *ASTIN Bulletin* **41**(4), 29–55.

[5] Danesi, I. L., Haberman, S. and Millossovich, P. (2015), 'Forecasting mortality in subpopulations using leecarter type models: A comparison', *Insurance: Mathematics and Economics* **62**, 151–161.

[6] Haberman, S., Kaishev, V., Millosovich, P., Villegas, A., Baxter, S., Baxter, S., Gunnlaugsson, S. and Sison, M. (2014), 'Longevity basis risk: A methodology for assessing basis risk', *Sessional meeting of the Institute and Faculty of Actuaries* .

[7] Kleinow, T. (2014), 'A common age effect model for the mortality of multiple populationns', *Insurance: Mathematics and Economics* .

[8] Lee, R. D. and Carter, L. R. (1992), 'Modeling and forecasting u.s. mortality', *Journal of the American Statistical Association* **87**, 659–675.

[9] Lee, R. and Miller, T. (2001), 'Evaluating the performance of the lee-carter method for forecasting mortality', *Demography* **38**(4), 537–549.

[10] Li, J. (2013), 'A poisson common factor model for projecting mortality and life expectancy jointly for females and males', *Population Studies: A Journal of Demography* **67**(1), 111–126.

[11] Li, J. S.-H., Zhou, R. and Hardy, M. (2015), 'A step-by-step guide to building two-population stochastic mortality models', *Insurance: Mathematics and Economics* **63**, 121–134.

[12] Li, N. and Lee, R. (2005), 'Coherent mortality forecasts for a group of populations: an extension of the lee-carter method', *Demography* **42**(3), 575594.

[13] OECD (2011), 'Health at a glance 2011'.

[14] Oeppen, J. and Vaupel, J. (2002), 'Broken limits to life expectancy', *Science* **296**, 1029–1031.

[15] Olshansky, S. J., Passaro, D. J., Hershow, R. C., Layden, J., Carnes, B. A., Brody, J., Hayflick, L., Butler, R. N., Allison, D. B., and Ludwig, D. S. (2005), 'A potential decline in life expectancy in the united states in the 21st century', *New England Journal of Medicine* **352**(11), 1138–1145.