KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Structural changes in mortality rates

with an application to Dutch and Belgian data

Frank van Berkum, Katrien Antonio, Michel Vellekoop

AFI_1379

Structural changes in mortality rates

with an application to Dutch and Belgian data

Frank van Berkum* †, Katrien Antonio†,‡, and Michel Vellekoop†

†Faculty of Economics and Business
University of Amsterdam
Amsterdam, The Netherlands

‡Faculty of Economics and Business
KU Leuven
Leuven, Belgium

April 2013

Abstract

We focus on a collection of stochastic mortality models, applied to two age buckets (20-89 and 60-89) of Dutch and Belgian mortality data. Recent literature relies on the standard ARIMA-framework (in particular: a random walk with drift) to project mortality rates. As a result the projections can be highly sensitive to the calibration period. We present a modelling strategy for the time-dependent parameters that allows for objective, statistical detection of one or more structural changes in the time series. By comparing projections based on different calibration periods and different time series specifications, we show that the proposed methodology leads to more robust mortality projections with respect to the calibration period used.

Key words: Stochastic mortality; structural changes; mortality forecasting; back testing

*Corresponding author. Faculty of Economics and Business, Valckenierstraat 65, 1018 XE Amsterdam, The Netherlands. Email: f.vanberkum@uva.nl

1 Introduction

Mortality rates have improved substantially during the last century. For pricing and reserving purposes it is important that improvements in mortality rates are adequately monitored and predicted. For Solvency II purposes, it is also important to quantify the uncertainty in future mortality rates.

Constructing mortality rate projections consists of two parts, namely (i) estimating a mortality model on historical data, and (ii) forecasting time dependent parameters from the estimated model. The current mortality modelling literature has mainly focused on the first part by creating more extensive mortality models allowing for specific factors such as a cohort effect or mortality improvements of specific age groups.

The seminal paper by Lee and Carter (1992) introduces a first stochastic mortality model which allows for mortality improvements. This is a factor model with age and period effects. Several extensions have been made to the Lee-Carter model, like the introduction of a cohort effect (Renshaw and Haberman (2006); Currie (2006)), using functional forms to specify the age effects to limit the number of parameters (Cairns et al. (2006) and Cairns et al. (2009)), and the introduction of age-group specific and quadratic effects (Plat (2009) and O'Hare and Li (2011)). Some of the models are designed for pensioner ages only (Cairns et al. (2009)), while others are applicable to the whole age range due to the inclusion of effects for specific age groups.

The predictive power of the models has been back tested extensively for multiple countries, for example England & Wales (Cairns et al. (2009), Dowd et al. (2010), Haberman and Renshaw (2011)), and Finland & Sweden (Lovász (2011)). Brouhns et al. (2002) consider the Lee-Carter model for Belgian data, and Plat (2009) considers the model introduced in Plat (2009) for Dutch data. As a first contribution our paper estimates the models mentioned above and evaluates their predictive power for Dutch males.

The modelling of time-dependent effects in mortality models is underexposed in recent literature. The time-dependent effects (i.e. period and cohort) are often modelled using ARIMA-models. However, when structural changes are present, the time-dependent effects cannot always be caputured by standard ARIMA-models. As a result, mortality forecasts following from projections using standard ARIMA-models are possibly highly sensitive towards the calibration period. The second and main contribution of this paper is that we introduce a strategy to test for structural changes in the period effects, and we date the structural changes in an objective manner (Bai and Perron (1998); Zeileis et al. (2003)). Using the BIC criterion the optimal number of structural changes is selected. Mortality forecasts using the proposed method are more robust towards the calibration period.

The remainder of this article is organised as follows. First, in Section 2 the mortality models are introduced and estimation results are presented. Next, in Section 3 we introduce the method used that allows for structural changes within the modelling of time-dependent effects. The mortality models are back tested in Section 4 the mortality models are back tested and Section 5 concludes.

2 Calibrating mortality models

2.1 Data

The HMD[1] provides data on the number of deaths during calendar year t aged x at death, $D_{t,x}$, and the average population during calendar year t aged x, $E_{t,x}$. The crude central mortality rate, $m_{t,x}$ is defined by

$$m_{t,x} = \frac{D_{t,x}}{E_{t,x}}. \tag{1}$$

[1]Human Mortality Database, see www.mortality.org

The probability that a person aged exactly x at the beginning of calendar year t dies within the next year is called the initial mortality rate $q_{t,x}$, and the force of mortality $\mu_{t,x}$ is the instantaneous death rate at exact time t for individuals ages exactly x at time t. If we assume that $\mu_{t,x}$ is constant in the interval $[t, t+1)$, the initial mortality rate is linked to the crude central mortality rate by the approximation [2]

$$q_{t,x} \approx 1 - e^{-m_{t,x}}. \tag{2}$$

We will estimate the mortality rates using age effects ($\beta_x^{(i)}$), period effects ($\kappa_t^{(i)}$), and cohort (year of birth) effects (γ_c, with $c = t - x$). Mortality models may include several age and period effects, hence the superscript (i) for the β's and κ's.

We use data from the Netherlands and Belgium for the years 1950 to 2008. Earlier data is excluded such that there are no world wars in the sample. Further, we consider the ages 20-89 as these are most important for insurers and pension funds.

2.2 Model structures

Following Brouhns et al. (2002) the number of deaths within a year follows a Poisson distribution of the form $D_{t,x} \sim \text{Poisson}(E_{t,x} m_{t,x})$. The various specifications for $m_{t,x}$ are listed in Table 1[3] .

Model	HR names	Formula	Original paper
M1	LC	$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}$	Lee and Carter (1992)
M1A	LC2	$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\kappa_t^{(3)}$	Renshaw and Haberman (2003)
M2	M	$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x}$	Renshaw and Haberman (2006)
M2A	-	$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\kappa_t^{(3)} + \gamma_{t-x}$	
M3	APC	$\log m(t,x) = \beta_x^{(1)} + \kappa_t^{(2)} + \gamma_{t-x}$	Currie (2006)
M5	CBD	$\text{logit } q(t,x) = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)}$	Cairns et al. (2006)
M6		$\text{logit } q(t,x) = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + \gamma_{t-x}$	Cairns et al. (2009)
M7		$\text{logit } q(t,x) = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + b(x)\kappa_t^{(3)} + \gamma_{t-x}$	Cairns et al. (2009)
M8		$\text{logit } q(t,x) = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (x_c - x)\gamma_{t-x}$	Cairns et al. (2009)
M9	M6*	$\log m(t,x) = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + + (\bar{x}-x)^+\kappa_t^{(3)} + \gamma_{t-x}$	Plat (2009)
M10	M5*	$\log m(t,x) = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + (\bar{x}-x)^+\kappa_t^{(3)}$	Haberman and Renshaw (2011)
M11	M7*	$\log m(t,x) = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + + (\bar{x}-x)^+\kappa_t^{(3)} + b(x)\kappa_t^{(4)} + \gamma_{t-x}$	Haberman and Renshaw (2011)
M12	M8*	$\log m(t,x) = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + + (\bar{x}-x)^+\kappa_t^{(3)} + (x_c - x)\gamma_{t-x}$	Haberman and Renshaw (2011)
M13	Expl.YM	$\log m(t,x) = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + + c(x)\kappa_t^{(3)} + \gamma_{t-x}$	O'Hare and Li (2011)

Table 1: Model specifications used in this paper.

Here, $b(x) = \left((x-\bar{x})^2 - \frac{1}{k}\sum_{i=x_1}^{x_k}(i-\bar{x})^2\right)$ where x_1 is the youngest and x_k the oldest age included in the data set, $c(x) = (\bar{x}-x)^+ + [(\bar{x}-x)^+]^2$, $\bar{x}$ is the average of the ages included in the sample, and x_c is

[2] See Cairns et al. (2009)

[3] The column 'HR names' shows the names used in Haberman and Renshaw (2011) to facilitate comparison.

		The Netherlands		Belgium	
Model	Ages	Males	Females	Males	Females
M8	60-89	60	60	60	60
M12	60-89	60	89	89	89
M12	20-89	20	89	20	26

Table 2: Optimal values for x_c in M8 and M12 when $x_c \in \{60, \ldots, 89\}$ or $x_c \in \{20, \ldots, 89\}$, based on the calibration period 1950-2008.

a constant which can be chosen upfront or can be estimated; in this paper we estimate this parameter[4]. For each of these models we specify the likelihood and apply standard Newton-Raphson methods to optimise this likelihood. However, most models experience identification issues. We therefore apply parameter constraints similar to those implemented in recent literature. Appendix A gives an overview.

The models M5 to M8 use the linearity of the age effects for the pensioner ages. That linearity does not hold for lower and higher ages, and these models are therefore appropriate for the pensioner ages only (60-89). We therefore calibrate the models M5-M8 only on the ages 60-89, and the other models are calibrated both for the ages 20-89 and the ages 60-89.

Plat (2009) considers two cases for the cohort effect: 1) estimate the cohort effect for all cohorts available, and 2) estimate the cohort effect only for cohorts older than 1946. The idea is that the cohort effect is most prominent for higher ages, and cohort effects estimated on younger cohorts are therefore not appropriate. For M9 and M13 we set cohort effects equal to zero when there are no observations available related to age 60 or higher, conform the idea in Plat (2009).

2.3 Estimation results

Ages 20 to 89

We present the results for the models M1-M3 and M9-M13 calibrated on ages 20-89 and period 1950-2008. Given the model specification in Table 1, the importance of the cohort effect for different ages depends on the parameter x_c. For ages close to x_c the cohort has only little effect, whereas for ages further away from x_c the cohort effect becomes more important. Table 2 shows the results for optimisation over the parameter x_c. We conclude that for the age range 20-89 x_c is often chosen rather low.

Table 3 shows the effective[5] number of parameters that is estimated in each of the models, and the corresponding BIC[6]. The models have the highest BIC if a cohort effect is included and interaction between age and period effects is allowed for. Further, the ranking of the models for Dutch male mortality rates is similar to the ranking of the models for Belgian male mortality rates.

Ages 60 to 89

Table 4 shows similar results for the ages 60-89. For illustration purposes only, we present the parameter estimates for M2 in Figure 1. Though the model is the same, the two calibrations on different age ranges result in completely different parameter estimates. This is due to the parameter constraints imposed and

[4] We estimate the model for all $x_c \in \{x_1, \ldots, x_k\}$, and the value of x_c is chosen such that the likelihood is optimal.

[5] The effective number of parameters is the total number of parameters that is included in the model minus the number of parameter constraints that are used to identify the model.

[6] The BIC is defined as $\text{BIC} = \log L - \frac{1}{2}k \cdot \log n$, where $\log L$ is the loglikelihood, n is the number of observartions, and k is the effective number of parameters. Higher is better.

	The Netherlands			Belgium		
Model	Number of parameters	BIC	Rank	Number of parameters	BIC	Rank
M1	197	-21,866.15	10	197	-22,662.75	10
M1A	324	-20,559.09	8	324	-21,146.23	8
M2	393	-19,675.04	5	393	-20,378.66	6
M2A	521	-20,095.07	7	521	-20,653.21	7
M3	254	-19,757.53	6	254	-20,348.68	5
M9	332	-19,412.83	1	332	-19,939.70	1
M10	244	-20,676.05	9	244	-22,189.76	9
M11	430	-19,624.63	4	430	-20,176.91	4
M12	372	-19,472.53	2	372	-20,023.38	2
M13	372	-19,578.75	3	372	-20,094.02	3

Table 3: Estimation results for Dutch and Belgian male mortality rates, estimated on the age range 20 to 89 and calibration period 1950-2008.

	The Netherlands			Belgium		
Model	Number of parameters	BIC	Rank	Number of parameters	BIC	Rank
M1	117	-10,589.19	14	117	-10,436.08	14
M1A	204	-9,761.63	13	204	-10,222.81	12
M2	233	-9,442.14	4	233	-9,636.54	4
M2A	321	-9,721.16	11	321	-9,877.67	9
M3	174	-9,424.88	3	174	-9,584.90	2
M5	118	-9,667.92	10	118	-10,234.93	13
M6	204	-9,297.87	1	204	-9,500.78	1
M7	262	-9,458.79	5	262	-9,662.22	5
M8	206	-9,363.17	2	206	-9,602.30	3
M9	292	-9,557.78	8	292	-9,748.52	7
M10	204	-9,465.34	6	204	-9,905.13	10
M11	350	-9,735.46	12	350	-9,929.52	11
M12	292	-9,589.46	9	292	-9,807.80	8
M13	292	-9,554.26	7	292	-9,745.66	6

Table 4: Estimation results for Dutch and Belgian male mortality rates, estimated on the age range 60 to 89.

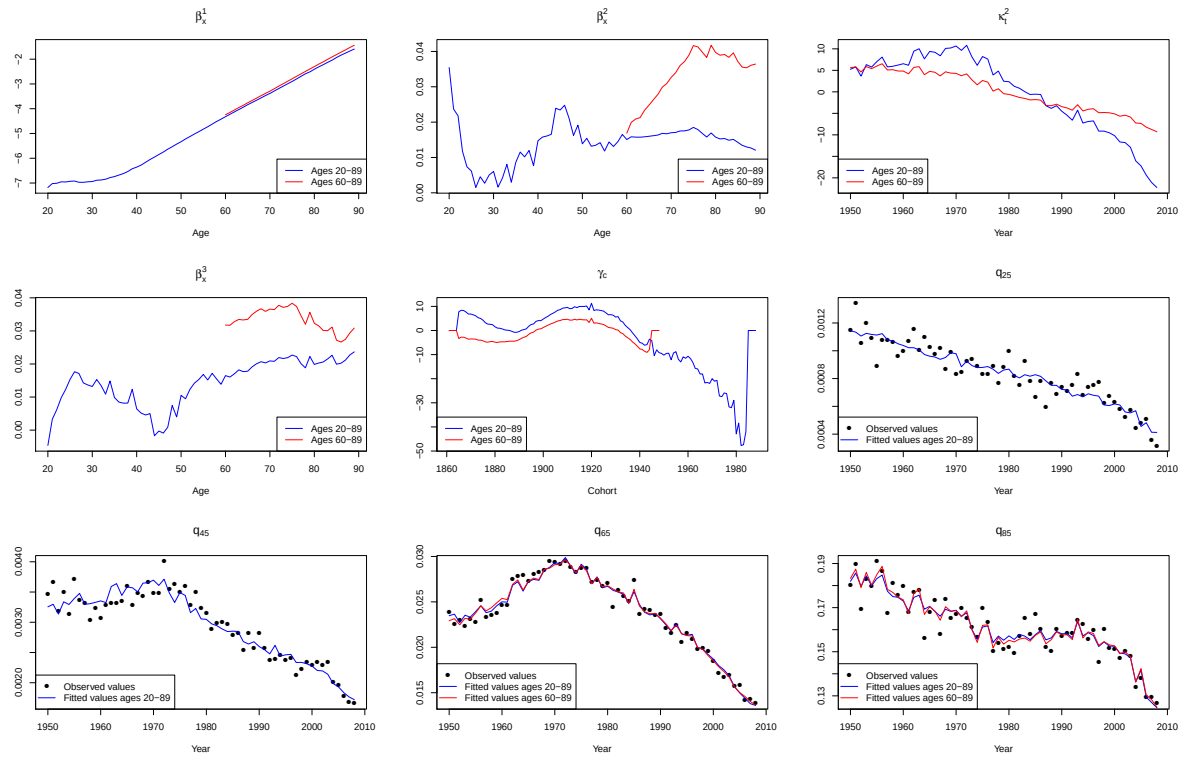


Figure 1: Parameter estimates for M2. Top row is calibrated on the age range 20 to 89, the middle row is calibrated on the age range 60 to 89. The bottom row shows realised mortality rates (dots), and fitted mortality rates for $x = \{25, 45, 65, 85\}$ (calibrated on ages 20-89 and ages 60-89), calibration period 1950-2008.

the different mortality dynamics in the two data sets. In order to project mortality, the parameter $\kappa_t^{(2)}$ needs to be projected into the future. Further, for new cohorts we will also need to project the cohort effect. This is mainly the case for the younger ages in the specific age range.

In what follows we show results only for the models M1, M2, M3 and M9 for the ages 20-89 and for the ages 60-89 we also consider the models M5 and M6, because these models represent the different mortality model structures (inclusion of a cohort effect, age-specific factors, interaction between factors or not).

3 Forecasting mortality

3.1 ARIMA models used in literature on mortality forecasting

Dowd et al. (2010) consider models M1 to M7. These models are fitted to England and Wales data from 1971 to 2006. They fit a (uni/multi)variate random walk with drift for the period effects, and either a mean reverting process (AR(1)) or an ARIMA$(1, 1, 0)$ process for the cohort effects.

Plat (2009) introduces M9 and includes it in a comparative study of mortality models fitted to data from the United States (1961 to 2005), England & Wales (1961 to 2005), and the Netherlands (1951 to 2005). In his approach the first period effect ($\kappa_t^{(1)}$ in Table 1) is the main effect, and a random walk with drift is used to project this factor. For the other period effects ($\kappa_t^{(2)}$ and $\kappa_t^{(3)}$ in Table 1), a non-stationary ARIMA process like a random walk with drift is not used for projection, as this *"could*

lead (in some scenarios) in projected scenarios where the shape of the mortality curve over ages is not biologically reasonable". Therefore, a mean reverting process is fitted with non-zero mean (AR(1) with a constant). Finally, a mean reverting process with mean zero is used to project the cohort effect. As a result, there is no trend in the projected cohort effect.

Haberman and Renshaw (2011) consider almost all models listed in Table 1, except for M2A and M13, and they consider the Lee-Carter model extended with a cohort effect instead of our M3 specification. The models are fitted on England and Wales data from 1961 to 2007. To project mortality these authors fit a multivariate random walk with drift for all period effects, similar to the approach used in Dowd et al. (2010). Haberman and Renshaw (2011) argue that the extrapolation of the cohort effect for M2-M3 should be avoided, because there is no justification to treat the cohort effect and the period effect independently. Therefore, they do not project the cohort effect, and focus on modelling life expectancy and annuity values for *existing* cohorts.

Lovász (2011) considers several models for Finnish (1950 to 2009) and Swedish (1950 to 2008) data. He models the period effects as in Dowd et al. (2010) and Haberman and Renshaw (2011), namely by assuming a multivariate random walk with drift. For the cohort effects he decides which ARIMA(p,d,q)-process fits the cohort effect best, by considering the combinations of $d = 0,1,2$ and $p,q = 0,1,2$, and choosing the order which results in the highest BIC. The optimal ARIMA specifications are always integrated, possibly with a lag included (ARIMA$(p,1,0)$); two times differencing is never optimal.

Finally, O'Hare and Li (2011) introduce M13 and fit several models to data from a range of developed countries from 1950 to 2006. The proposed model is a modification of Plat's model, and hence, they use the same ARIMA-specifications as in Plat (2009). A random walk with drift is used for the main period effect, mean reverting processes with non-zero mean are used for the remaining period effects, and a mean-reverting process with mean zero is used for the cohort effect.

The papers above all use a random walk with constant drift for the first period effect, and often also for the other period effects. However, factors like medical advances (Bots and Grobbee (1996)) and reforms of health systems (Moreno-Serra and Wagstaff (2010)) can have an impact on the speed of the mortality improvements. Further, the papers above used different calibration periods, and projections based on a random walk with constant drift are potentially highly sensitive towards the calibration period, see e.g. Booth et al. (2002) and Denuit and Goderniaux (2005). The assumption of a *constant* drift may therefore be unrealistic. We propose to allow for structural changes to tackle this problem.

3.2 Structural changes and regime switching models: a literature overview

Regime switching models

Milidonis et al. (2011) calibrate the Lee-Carter model on US data. They propose a regime switching model with two regimes for the differenced series of κ_t. The two regimes are allowed to have both different mean as well as different variance, but from the estimation results it follows that especially the variance is different between the two regimes. They conclude that for the data set considered, the regime switching model performs better than a random walk with drift. Hainaut (2012) extends the regime switching model to model M1A and concludes that the increase in loglikelihood is significant.

Structural change in trend stationary models

Li et al. (2011) fit a broken-trend stationary model to the time series κ_t from the Lee-Carter model:

$$\kappa_t = \alpha_1 + \beta_1 t + (\alpha_2 - \alpha_1)\Gamma_t(T^*) + (\beta_2 - \beta_1)\Psi_t(T^*) + \varepsilon_t, \qquad \varepsilon_t \sim N(0,\sigma_\varepsilon^2) \tag{3}$$

where

$$\Gamma_t(T^*) = \begin{cases} 1 & \text{if } t > T^* \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \Psi_t(T^*) = \begin{cases} t & \text{if } t > T^* \\ 0 & \text{otherwise} \end{cases},$$

where T^* is to be estimated using a specific algorithm. This model implies

$$\mathrm{E}[\Delta\kappa_t] = \mathrm{E}[\kappa_t - \kappa_{t-1}] = \begin{cases} \beta_1 & t < T^* + 1 \\ (\alpha_2 - \alpha_1) + \beta_1 + (T^* + 1)(\beta_2 - \beta_1) & t = T^* + 1 \\ \beta_2 & t > T^* + 1. \end{cases} \tag{4}$$

The break comes with a jump if $\alpha_1 \neq \alpha_2$. Further, since this is a trend stationary process, predictions from this model do not depend on the previous observation, and for future unknown values the variance does not increase with time. The specification of the model above only allows for one break point to be detected, though it is possible to extend the model to allow for multiple break points.

Sweeting (2011) considers the original CBD-model (M5), and applies the following specification for the period effects:

$$\kappa_t = \begin{cases} \alpha_t + \beta_1 t & \text{if } t \leq b_1 \\ \alpha_2 + \beta_2 t & \text{if } b_1 < t \leq b_2 \\ \vdots & \\ \alpha_N + \beta_N t & \text{if } t > b_N \end{cases} \tag{5}$$

where b_j is the jth break point detected for κ_t, and N is the total number of break points detected. This model specification implies that piecewise linear lines are fitted, and restrictions are imposed such that the different lines connect. Sweeting allows for multiple structural changes to occur, and tests whether the inclusion of a new break point (due to a structural change) leads to a significant improvement in fit by using the Chow test (Chow (1960)). He also tests whether the change in parameter estimates from one period to another is significant. However, as Bai (1997) argue, the Chow test is not appropriate using critical values from the standard F-distribution if the break points are unknown *a priori*.

Structural change in difference stationary models

Coelho and Nunes (2011) consider the Lee-Carter model for a variety of countries, both for males and females. When modelling the period effect, they consider two types of models: a trend stationary model, and a difference stationary model (previously mentioned as the random walk model). The broken-trend stationary model is given by

$$\kappa_t = \alpha + \beta t + \gamma \mathrm{DT}_t(\tau) + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2), \qquad t = 1, \ldots, T \tag{6}$$

where the change dummy variable is defined as $\mathrm{DT}_t(\tau) = t - T_B$ if $t > T_B$ and $\mathrm{DT}_t(\tau) = 0$ if $t \leq T_B$, where $T_B = [\tau T]$ denotes the possible date of change in trend, with $\tau \in (0, 1)$. In contrast with Li et al. (2011), jumps will not occur in the specification used by Coelho and Nunes (2011), because of the different specification of the dummy variable and the fact that the constant is not time-varying. Their difference stationary model including a structural change, is given by

$$\Delta\kappa_t = \beta + \gamma \mathrm{DU}_t(\tau) + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \qquad t = 1, \ldots, T \tag{7}$$

where the dummy variable is given by $\mathrm{DU}_t(\tau) = 1$ if $t > T_B$ and $\mathrm{DU}_t(\tau) = 0$ if $t \leq T_B$.

Coelho and Nunes (2011) use tests suggested by Harvey et al. (2009) and Harris et al. (2009) to test whether the time series is trend stationary or difference stationary with possibly a structural change.

They perform this analysis for 18 countries both for males and females. From all these data sets, (6) is rejected 33 out of 36 times in favour of (7). Further, for 21 out of 36 data sets a structural change is detected. However, they note that a possible improvement would be to allow for multiple breaks, and to extend the approach to other mortality models.

Medical advances or health system reforms are permanent effects; we do not expect the advances or reforms to be reversed. Therefore, we believe that a trend or difference stationary process is more appropriate than a regime switching model. Following the findings from Coelho and Nunes (2011), we focus on the difference stationary process. However, we will extend the methodology such that multiple break points can be detected, as multiple events in the past may have affected the speed of mortality improvements.

3.3 Modelling strategy for period effects

Our starting point is the assumption that the period effects all follow a random walk with a (piecewise) constant drift. We follow the steps as outlined below to detect the prescence of structural changes, and to project the time dependent parameters.

1. Check for structural changes in each time series individually, and date the break points if any;
2. Fit a (uni/multi)variate random walk with piecewise constant drift, given the break points dated in step 1;
3. Simulate paths from the time series model, given the parameter estimates from step 2.

3.3.1 Dating the structural changes

To estimate the break dates, we follow the methodology as set out in Bai and Perron (2003) (see Zeileis et al. (2003) and Zeileis and Kleiber (2005) for a detailed description of an implementation). Suppose we have at our disposal a time series y_t $(t = 1, \ldots, T)$. We estimate a random walk with a piecewise constant drift for the time-dependent variables:

$$\Delta y_t = \begin{cases} \beta_1 + \varepsilon_t, & t \leq t_1 \\ \ldots & \\ \beta_i + \varepsilon_t & t_{i-1} < t \leq t_i \\ \ldots & \\ \beta_{m+1} + \varepsilon_t, & t_m < t \end{cases} \tag{8}$$

where $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. We estimate this model using OLS[7], hence, we minimise the sum of squared residuals (SSR):

$$\text{SSR}(t_1, \ldots, t_m) = \sum_{i=1}^{m+1} \sum_{t=t_{i-1}+1}^{t_i} [\Delta y_t - \beta_i]^2 \tag{9}$$

where $t_0 = 1$ and $t_{m+1} = T$. In the model specification above, we distinguish m break points, which divide the time series up into $m + 1$ regimes with different drifts[8]. Both the number of break points and the date of the break points are unknown. Bai and Perron (2003) describe an efficient algorithm to determine the sum of squared residuals as in (9) for each of the partitions, given m break points.

[7] Estimating a model using OLS is equivalent to maximising the likelihood assuming Gaussian errors

[8] In this general time series specification, we use the term drift. In the Lee-Carter model, which includes only one time-dependent variable, this drift term can be interpreted as the general 'mortality trend'. However, for the other models which include more time-dependent variables we can no longer call the drift term the mortality trend.

Following Bai and Perron (2003), let $\hat{\beta}(\{T_m\})$ denote the estimates $\{\hat{\beta}_1, \ldots, \hat{\beta}_{m+1}\}$ based on a given m-partition $(t_1, \ldots, t_m)$ denoted $\{T_m\}$. If we substitute these parameter estimates $\hat{\beta}(\{T_m\})$ into (9), then the estimated break points $(\hat{t}_1, \ldots, \hat{t}_m)$ are such that $(\hat{t}_1, \ldots, \hat{t}_m) = \text{argmin}_{t_1,\ldots,t_m} \text{SSR}(t_1, \ldots, t_m)$, where the minimisation is taken over all partitions $(t_1, \ldots, t_m)$ for which $t_i - t_{i-1} \geq h$. The parameter h corresponds to the minimum length of period that a regime should last, and is to be chosen upfront.

If we set h too low it is possible that spurious effects are picked up, which is undesirable. On the other hand, if we set h too high, then it is possible that we miss break points because they are not allowed. We take $h = 5$ which is in line with Zeileis et al. (2003) and Harvey et al. (2009), who suggest to set h equal to 10% of the sample.

3.3.2 Determining the number of break points

Given the method described above, we can determine the optimal break dates $(t_1, \ldots, t_m)$ for an *a priori* given number of break points m. We then have to determine what the optimal number of break points, say m^*, is. In general there are two ways to choose the optimal number of break points: (i) performing F-tests to test the significance of the increase in fit, and (ii) using an information criterion like the BIC.

If the information criterion is used, then one determines the BIC for all $m \in \{0, \ldots, 5\}$[9]. Denote BIC(m) as the BIC corresponding to the optimal break dates for a given number of break points m. The optimal number of break points is then defined by $m^* = \arg\max \text{BIC}(m)$, see Zeileis et al. (2003) and Zeileis and Kleiber (2005) for an implementation.

Bai and Perron (1998, 2003) consider two F-tests. The first is the sequential test of $m = l$ versus $m = l + 1$ break points. This is a sequential procedure: one starts with the null hypothesis of $m = 0$ versus the alternative of $m = 1$ break points. The F-statistic is a function of the restricted and the unrestricted sum of squared residuals (the null and alternative hypothesis, respectively):

$$F = \frac{(\text{RSSR} - \text{USSR})/(p_1 - p_0)}{\text{USSR}/(n - p_1)}, \tag{10}$$

where RSSR is the restricted sum of squared residuals (SSR), USSR is the unrestricted SSR, p_0 is the number of parameters in the model under the null hypothesis, p_1 the number of parameters in the model under the alternative hypothesis, and n is the number of observations. Since the dates of the structural changes are unknown, the F-statistic does not follow the standard F-distribution and critical values have to be obtained through simulation. If the break point is significant, then this break point is fixed and one searches for a new break point. The old break point is not allowed to move, which may be suboptimal when searching for more break points. Suppose there are two break points but we only allow for one, then the estimated break point will most likely be dated between the two actual break points. Using the sequential method it is then unlikely that the real break points are estimated. Therefore, we shall not use the sequential F-test.

The second F-test from Bai and Perron (1998, 2003) is based on the null hypothesis of no break point ($m = 0$) versus the alternative hypothesis of $m = k$ break points. To determine the optimal number of break points, they determine the F-statistic as defined in (10) for all $k \in \{1, \ldots, 5\}$ which leads to $F(k)$. They then define the UDmax test statistic as the maximum value of those F-statistics:

$$\text{UDmax} = \max_k F(k) \tag{11}$$

Since the number and dates of the break points are unknown, critical values have to be obtained through simulation. If the observed UDmax test statistic is larger than the critical value, then the number of

[9] We consider at most five structural changes. In the analysis performed there was no reason to allow for more structural changes.

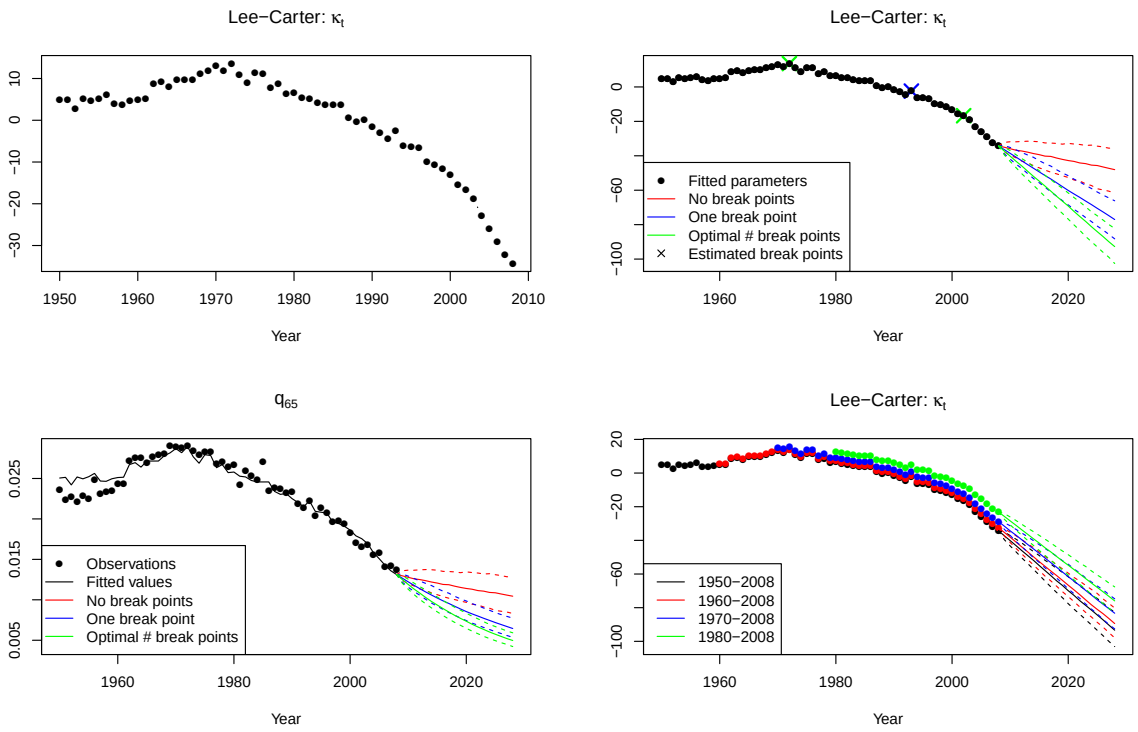


Figure 2: Top left: parameter estimates of κ_t in the Lee-Carter model, calibrated on data from Dutch males aged 20-89 in the period 1950-2008. Top right: projections for the period effect using different projection methods. Bottom left: mortality projections for $x = 65$ using different projection methods for the period effect. Bottom right: projections of the period effect for different calibration periods while allowing for multiple structural changes.

break points is equal to the number of break points on which the UDmax test is based. If the test statistic is smaller than the critical value, then there is no proof of a structural change.

The latter F-test is close to using the BIC. The essential difference is in using either an information criterion or a statistical test. Yao (1988) shows that the number of break points that follows from optimising the BIC is a consistent estimator of the true number of break points, though he does note that this estimation method may overestimate the number of break points in case of data with a fat tail (compared with the normal distribution). We will use the BIC to choose the number of break points. In the following paragraph we show the results of using the BIC and the UDmax test to determine the number of break points.

Case study - the Lee-Carter model

We consider the period effect of the Lee-Carter model, estimated on Dutch male mortality data for the period 1950 to 2008, for the ages 20 to 89. The top left graph in Figure 2 shows the parameter estimates for κ_t. A random walk with constant drift does not seem appropriate based on visual inspection.

Figure 3 shows the corresponding break points as obtained with the package `strucchange` in `R`. From the upper right graph we observe that the first break point is accurately estimated, as the confidence interval[10] (shown by the red line) is small. The lower left graph in Figure 3 shows the confidence intervals for the case of two break points. The second break point (around the year 2002) is estimated accurately,

[10] See Bai and Perron (1998) for a description how these confidence intervals are derived.

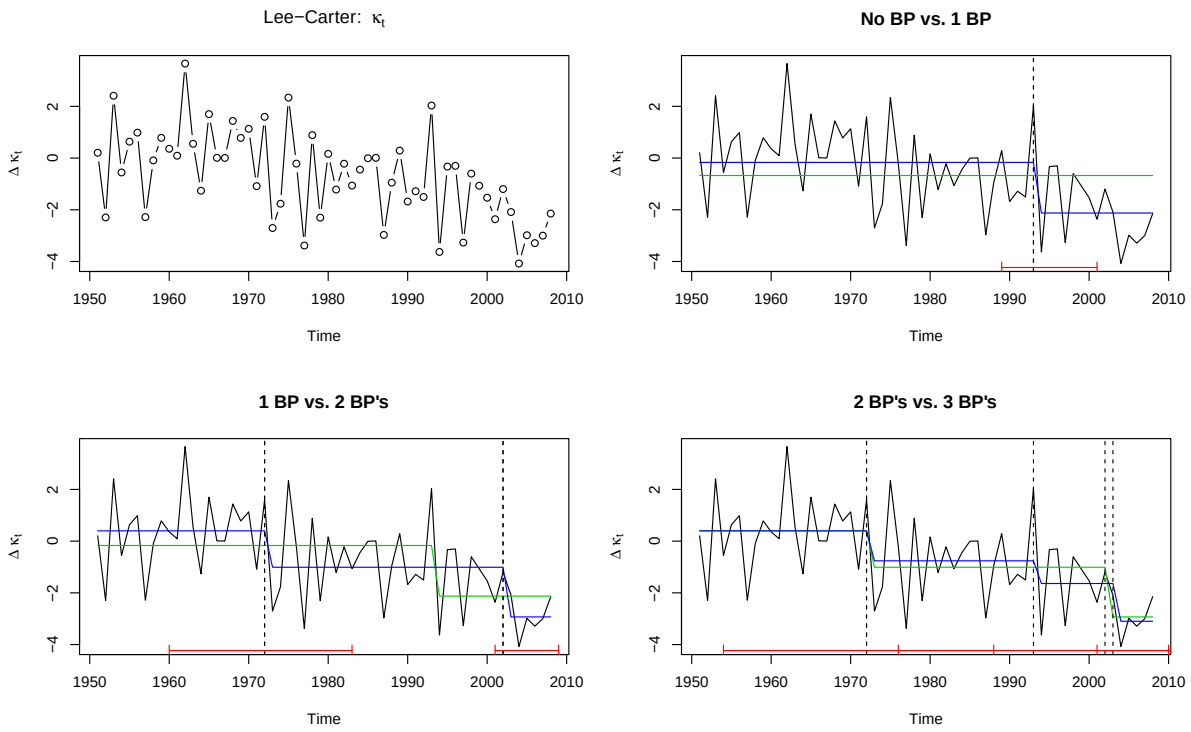


Figure 3: Confidence intervals for estimated break points for κ_t^2 in the Lee-Carter model, calibrated on Dutch males aged 20-89. In the plots (i) BP's vs. $(i+1)$ BP's the green lines represent the mean of $\Delta\kappa_t$ for the different periods when (i) BP's are allowed, and the blue lines represent the mean of $\Delta\kappa_t$ when $(i+1)$ BP's are allowed. The red lines show represent the confidence intervals corresponding to the break points.

but the confidence interval corresponding with the first break point is wide. This can be explained by the outliers before and after the year 1972. However, allowing for the second break point leads to an increase in fit over the whole observation period. This is illustrated by the differences between the green and blue lines in Figure 3.

The bottom right graph shows the confidence intervals for the case of three break points. The confidence intervals overlap and they are much larger than for the case of two break points. Further, the impact on the fit is limited, which is illustrated by the overlap of the green and blue line for the largest part of the observation period.

The estimation results for different numbers of break points are presented in Table 5. As we would expect, the SSR decreases as the number of break points increases[11]. Since the BIC is minimal for $m = 2$, using the BIC as a criterion suggests using two break points. The F-statistic for $m = 0$ versus the alternative of $m = k$ break points is largest for $m = 1$. The critical value for the UDmax-test is 10.17 (see Table 1 in Bai and Perron (1998)). Since the F-statistic for $m = 1$ is larger than the critical value, using the UDmax-test suggests using one break point.

The break points estimated for $m = 2$ correspond to well known break points for the Dutch situation. Around the year 1970 people started smoking less in the Netherlands, which increased health of the population. Further, around the year 2001 the Dutch government decided the waiting lists for angioplasty should be reduced and spent money to achieve this. The lower left graph in Figure 5 clearly reveals the

[11] It may be that the SSR increases if the number of break points is increased. This is than due to the minimum length that a trend should last, which may cause break points to be dated suboptimally.

No. BP's	SSR	BIC	F-stat
0	164.50	233.18	-
1	121.90	223.92	**19.92**
2	105.28	**223.54**	15.75
3	100.35	228.87	11.72
4	96.92	234.98	9.41
5	95.73	242.38	7.61

Table 5: Estimation results for different numbers of break points. The boldface numbers show the optimal values for the BIC and for the F-statistic.

increase in mortality improvements for the elderly.

In the top row in Figure 2, the red lines show predictions when structural changes are not allowed for and corresponds to the method most commonly used for projecting period effects. The blue lines show predictions when one structural change is allowed for, which corresponds to the outcome of the UDmax-test and with the single break point strategy in Coelho and Nunes (2011). Finally, the green lines show predictions when multiple structural changes are allowed for as determined by the BIC.

It is clear from Figure 2 that allowing for multiple break points leads to more appropriate projections of the period effects, and possibly also of mortality rates. The lower right graph shows the projections of the period effect using the optimal BIC strategy for different calibration periods, and the slope of the projections using different calibration periods is similar[12]. This implies that the projection of the rate of mortality improvements is robust with respect to the calibration period when multiple break points are allowed for.

3.4 Modelling strategy for cohort effect

Figure 4 shows the estimated cohort effects for Dutch males, calibrated on the ages 60-89 and the years 1950-2008. The shapes of the cohort effects are partially due to the parameter restrictions, and these shapes are hard to capture within an ARIMA-specification. However, for consistency with recent mortality modelling literature, we shall restrict ourselves to the ARIMA models. In contrast with Plat (2009), we shall not impose any kind of model specification like a mean reverting process; we only impose that the time series model belongs to the ARIMA-framework.

In preliminary analysis we based the choice for an ARIMA-specification on the BIC of the model, like e.g. Lovász (2011). However, using only the BIC as a criterion may lead to unrealistic and biologically implausible mortality forecasts, see Cairns et al. (2011). Therefore, we project the cohort effect using ARIMA(p, d, q)-specifications for $\{p, d, q\} \in \{0, 1, 2\}$ and choose one that yields reasonable projections. The choice for the specification is based on (i) the reasonability of the best estimate projection compared to the historical observations, and (ii) an assessment of the width of the 95% confidence interval for projections in the future: the confidence interval should not be excessively wide compared to the range of the historical observations.

[12] In the Lee-Carter model, the slope of the period effect corresponds to the rate of the mortality improvements.

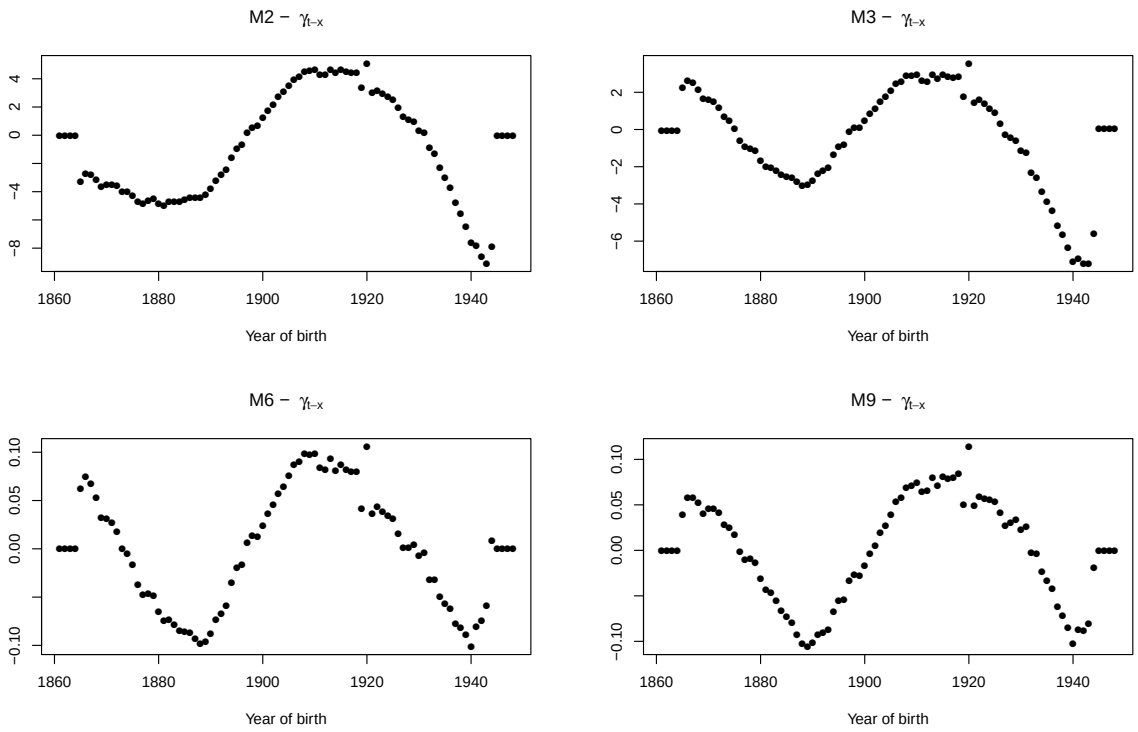


Figure 4: Parameter estimates of the cohort effect for different models, calibrated on data of Dutch males, ages 60-89, years 1950-2008.

Case study - M3 γ_{t-x}

Figure 5 shows in the upperright graph projections for the cohort effect based on different strategies. The blue line shows the projections obtained with a mean reverting process (as suggested by Plat (2009)), the red line shows the projections obtained with a BIC optimal ARIMA process, and the green line shows the projections resulting from our choice, based on a visual inspection of the projections for different ARIMA specifications.

The median of the projections of the cohort effect based on the optimal BIC ARIMA model connects well with historical observations. However, the projections based on the optimal BIC lead to excessively wide confidence intervals in this particular case. The projections for the AR(1) model lead to a confidence interal that is much smaller, but the projections do not connect well with the historical observations. Our visual inspection leads to a random walk with drift in this example, as this leads to more realistic projections: the projections connect better with the historical observations and the confidence interval is more reasonable.

The bottom row in Figure 5 shows the projected mortality rates resulting from the different cohort projections (left for $x = 65$, right for $x = 75$). From the estimated model we have obtained cohort effects for the cohorts up to 1935[13]. Therefore, uncertainty in the cohort effect will start from the cohort 1936. This corresponds to the year 2001 for $x = 65$ and to the year 2011 for $x = 75$. This is why the projections for the different cohort projection methods move away from each other from the year 2000 onwards for $x = 65$. As cohort parameters are available up to the year 2010 for $x = 75$, the mortality

[13]The calibration sample consists of the years 1950-1999 and the ages 60-89. Hence, the first cohort in the sample is $1950 - 89 = 1861$ and the last cohort in the sample is $1999 - 60 = 1939$. However, the first four and last four cohorts are excluded from the sample to avoid spurious effects to be picked up, since those cohort effects will then be estimated on four or less observations. The cohort effect is estimated for the cohorts 1865 to 1935.

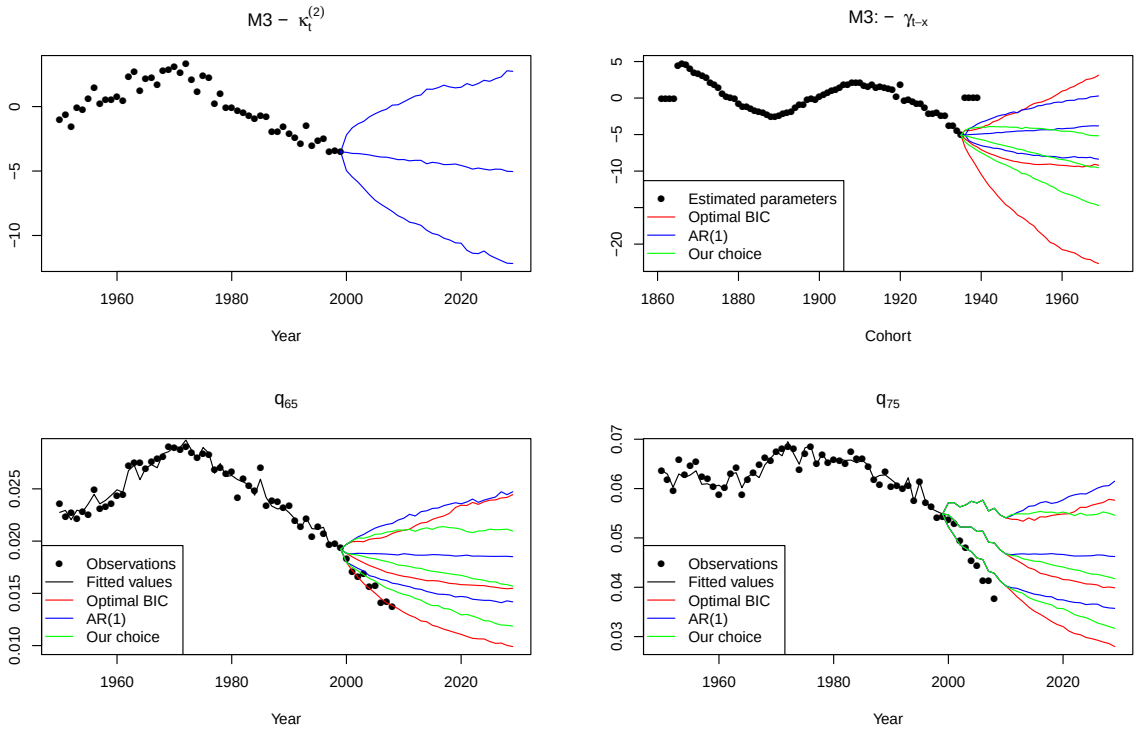


Figure 5: Parameter estimations and back tests for M3, calibrated on Dutch male mortality on the ages 60 to 89 and the years 1950 to 1999. The upper left graph shows the prediction interval of the period effect, based on the projection method as described in Section 3.3. The upper right graph shows predictions of the cohort effect based on optimal BIC, AR(1), and our choice based on visual assesment. The two lower graphs show back tests for M3 on q_{65} and q_{75} by comparing the projections from the year 2000 onwards with the corresponding realisations, using different projection methods for the cohort effect.

projections are the same up to the year 2010, and diverge from 2011 onwards because then the different cohort projections are used.

4 Back testing the mortality models

4.1 Mortality forecasts

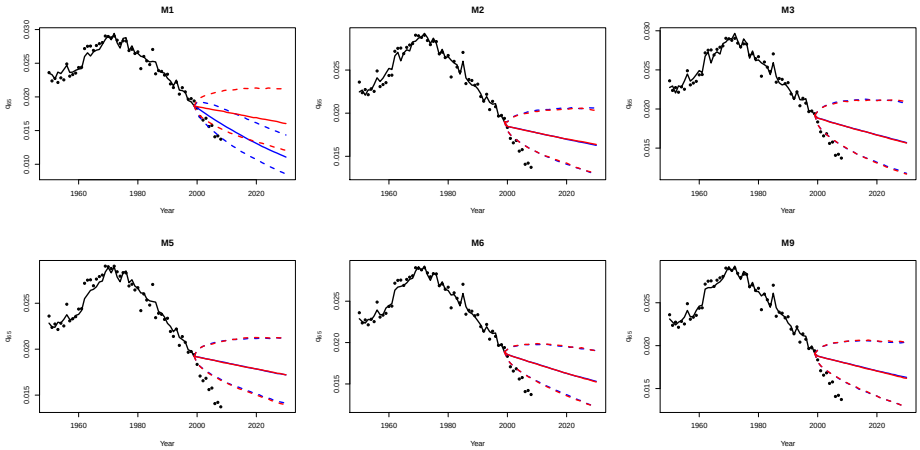
Ages 60 to 89

Figure 6 shows mortality forecasts for Dutch males for the models M1, M2, M3, M5, M6 and M9 for $x = \{65, 75, 85\}$, calibrated on the period 1950-1999 and the ages 60-89. The red lines show projections when break points are not allowed for, the blue lines represent projections when break points are allowed for[14].

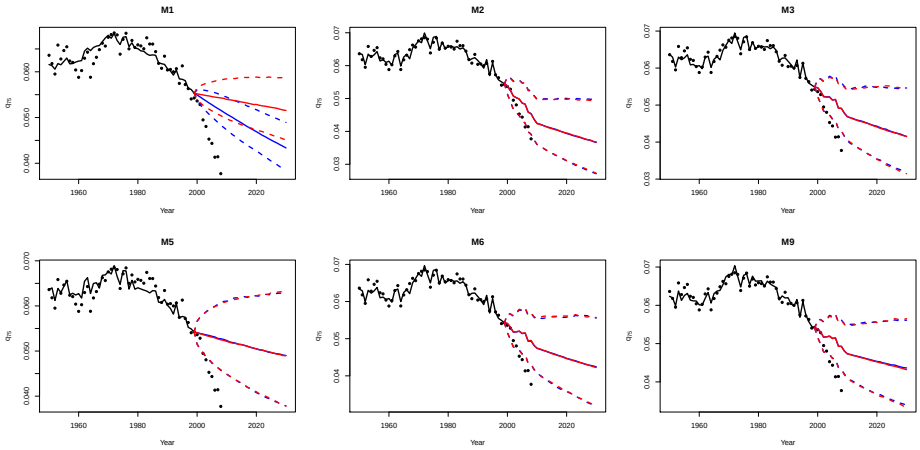
The Lee-Carter model is not able to capture the dynamics at all ages appropriately, which can be clearly seen from the graph for $x = 85$. The other models include more time-dependent parameters and the fitted mortality rates from those models are closer to the realised mortality rates.

For models including a cohort effect, the estimated cohort effects for the most recent cohorts can be

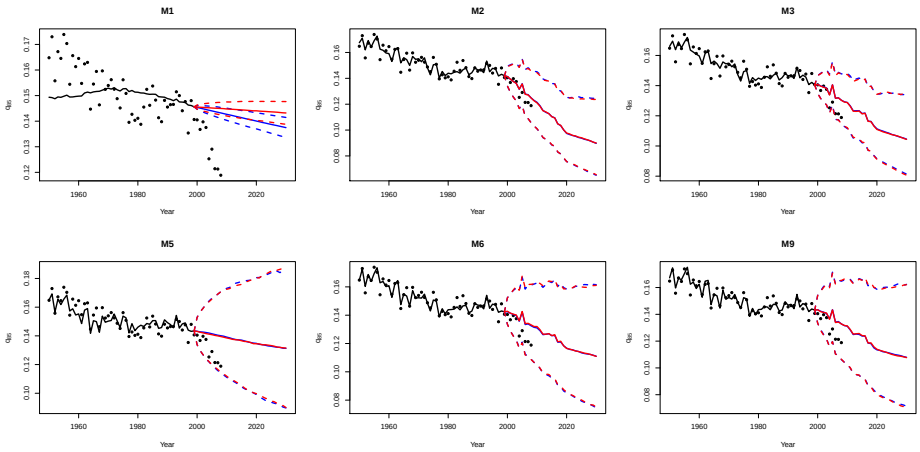
[14]Similar graphs are available for Dutch females and Belgian males and females upon request from the authors.



(a) 65 year old



(b) 75 year old



(c) 85 year old

Figure 6: Mortality rate projections for Dutch males, calibrated on the years 1950-1999 and the ages 60-89. The red lines correspond to projections when break points are not allowed for, the blue lines correspond to projections when break points are allowed for.

used for mortality projections. The projections using an estimated cohort effect are closer to the out-of-sample mortality rates compared to projections that do not include an estimated cohort. Hence, M2, M3, M6 and M9 perform better in the near future when estimated cohort parameters are available[15]. However, it is difficult to create appropriate cohort projections, which becomes apparent in the projections for $x = 65$ from the year 2000 onwards, for $x = 75$ from the year 2010 onwards, and for $x = 85$ from the year 2020 onwards. Further, for other data sets M6 sometimes results in implausible forecasts due to the estimated cohort parameters.

None of the models lead to projections that follow the out-of-sample mortality rates closely. This can be explained by the structural change that was identified around the year 2002 in Section 3.3.2 (which is out-of-sample in order to perform the back test). However, we do observe that allowing for structural changes leads to more appropriate projections for the Lee-Carter model. For the period effects of the other models we do not find statistical evidence for the presence of structural changes. A possible explanation for this is that those models include more time-dependent effects. The time-dependent effects are then more volatile, which makes it more difficult to find statistical evidence for structural changes.

Ages 20 to 89

Figure 7 shows mortality forecasts for Dutch males for the models M1, M2, M3 and M9 for $x = \{25, 45, 65, 85\}$, calibrated on the period 1950-1999 and the ages 20-89. For this age range we find statistical evidence for structural changes in M1 and M9, and as a result the projections from those models are more accurate than when structural changes are not allowed for. However, the projections from the Lee-Carter model do not follow the mortality developments closely for all ages.

The projections from M2, M3 and M9 for $x = 65$ and $x = 85$ include estimated cohort effects, which causes the projections to follow the out-of-sample mortality rates closely. For $x = 45$ the projections from M9 are closer to the out-of-sample mortality rates than the projections from M2 and M3. However, for $x = 25$ all models are unable to provide accurate mortality projections, because there is a clear break in the mortality rates around the year 2000, which has not been captured in a period or cohort effect. We expect this effect to be detected, if persistent, in a model with structural changes once more data have become available. As such it reinforces our central message that incorporating multiple structural changes may lead to significant improvements in dynamic models for mortality.

4.2 Expanding calibration period back test

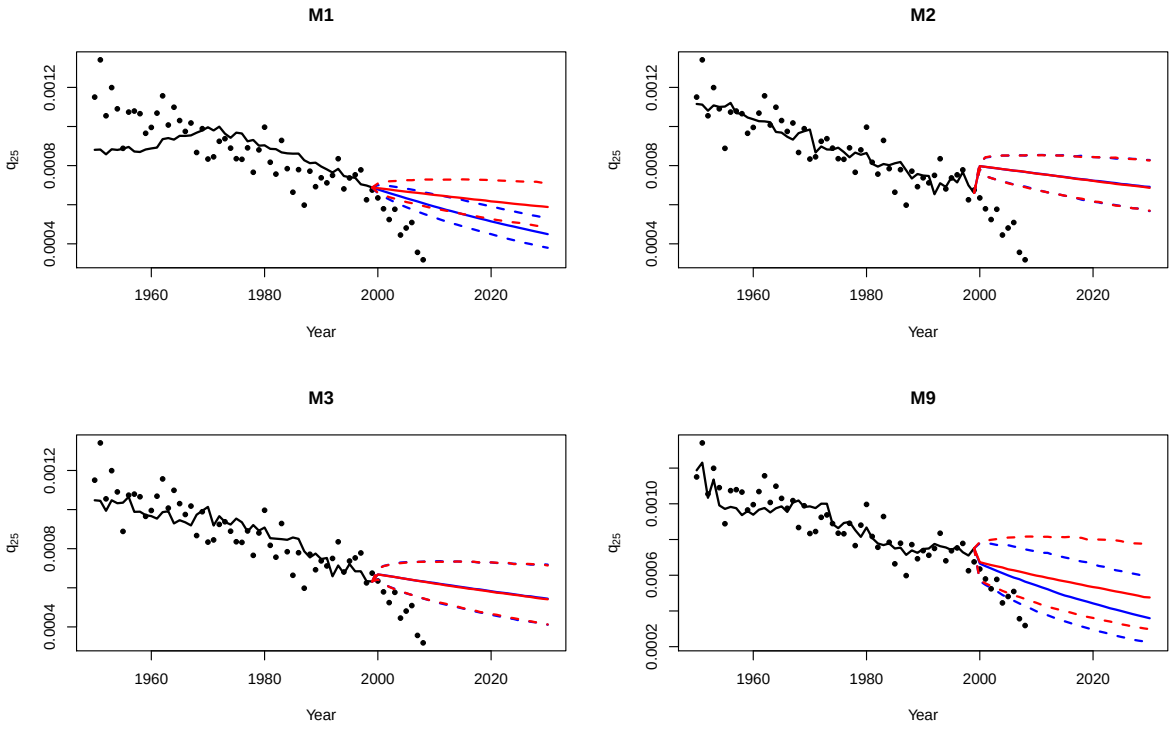
To assess the predictive power of the mortality models, we use an expanding calibration period back test[16]. We calibrate the models on the period 1950-1980 and project mortality rates 5, 10 and 20 years ahead in the future where we choose the optimal number of structural changes based on the BIC. We repeatedly do this by expanding the calibation period one year at a time, up to the calibration period 1950-2003. We shall refer to the collection of calibration periods as 1950-1980(1)2003. From these projections we obtain the 90% confidence intervals for the mortality rates.

Ages 60 to 89

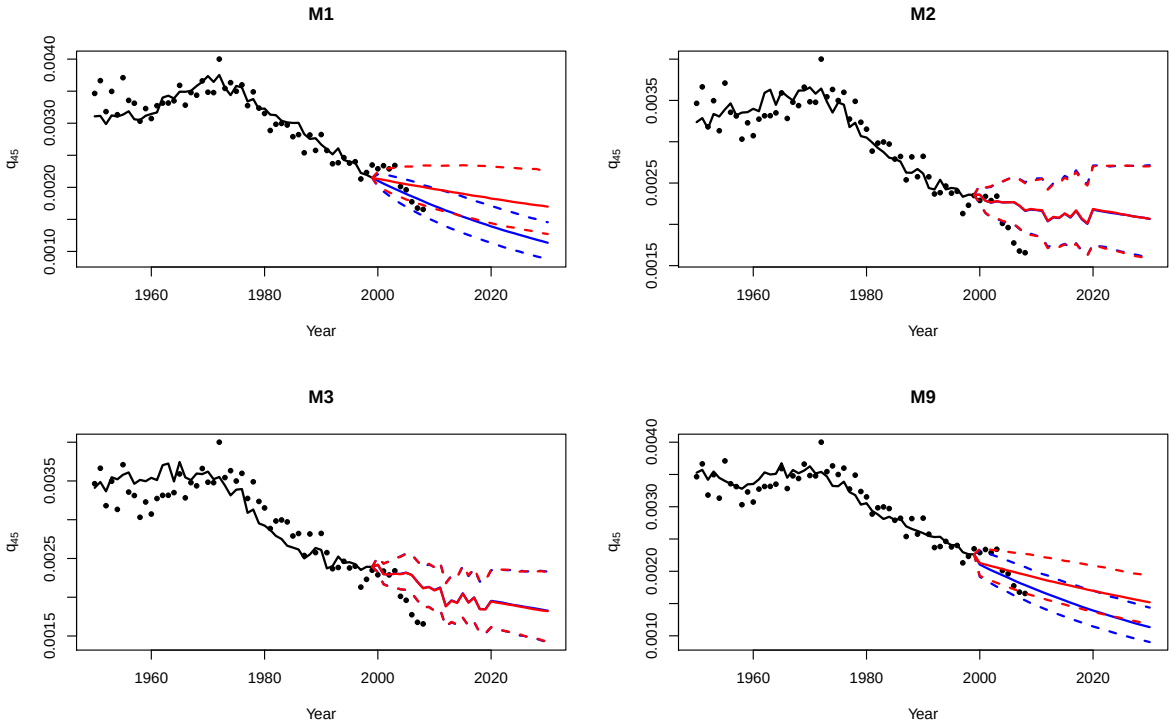
Figure 8 shows the back test for Dutch males for the models M1, M2, M3, M5, M6 and M9 for $x = \{65, 75, 85\}$, calibrated on the period 1950-1980(1)2003 and the ages 60-89. The red areas correspond to

[15] For $x = 85$ there are more estimated cohort parameters available then for $x = 65$. Therefore, projections for q_{85} are accurate futher into the future than for q_{65}.

[16] This back test is closely related to the rolling fixed-length horizon back test as in Dowd et al. (2010). However, in Dowd et al. (2010) the calibration period is fixed at twenty years, whereas we expand the calibration period. Also, we do not look at the p-value, but at the confidence intervals for the mortality rates.

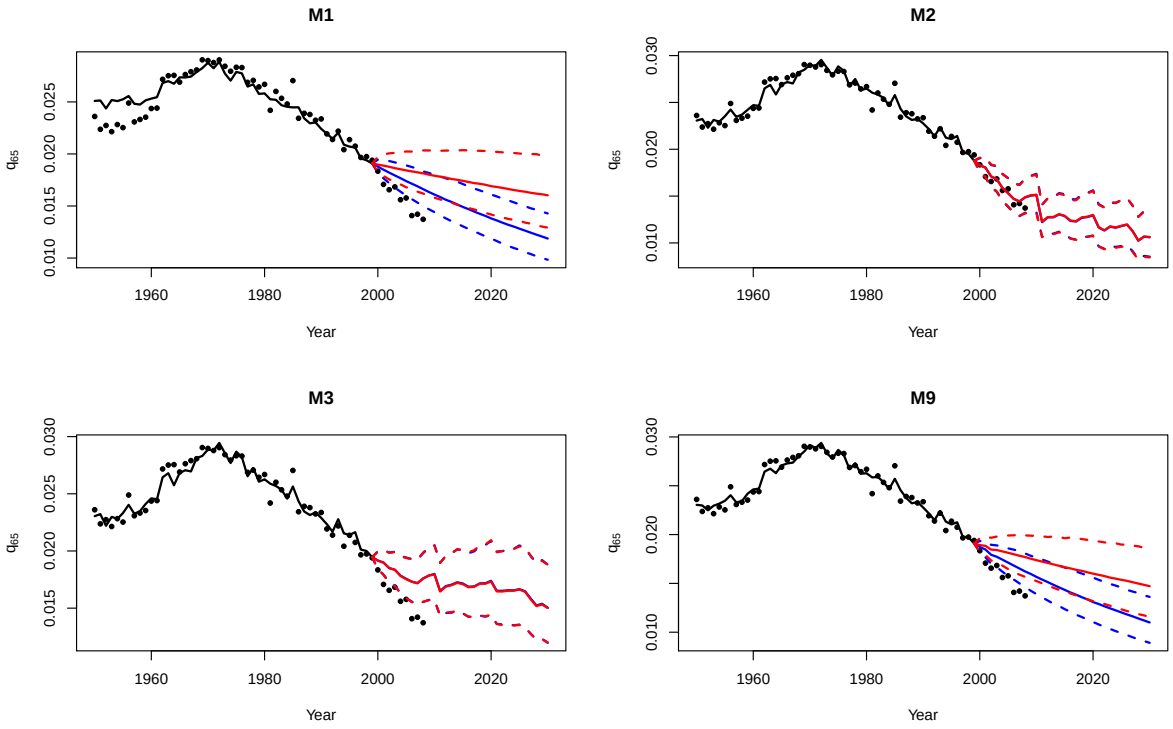


(a) 25 year old

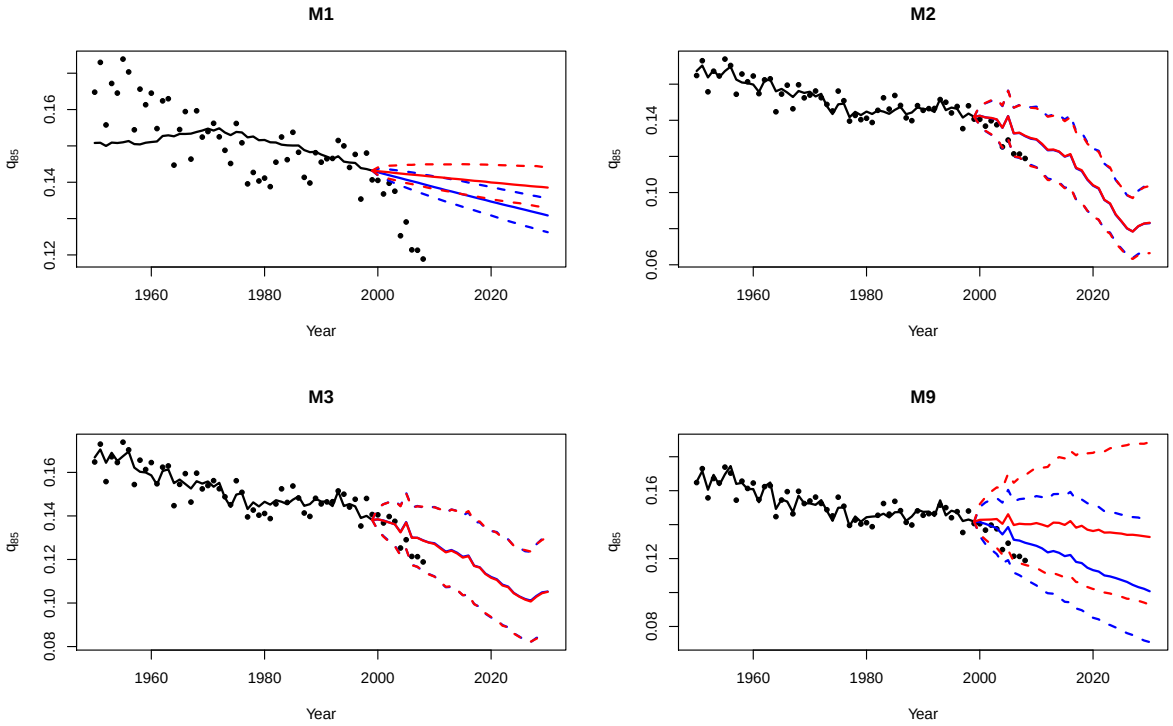


(b) 45 year old

Figure 7: Mortality rate projections for Dutch males, calibrated on the years 1950-1999 and the ages 20-89. The red lines correspond to projections when break points are not allowed for, the blue lines correspond to projections when break points are allowed for.



(c) 65 year old



(d) 85 year old

Figure 7: Mortality rate projections for Dutch males, calibrated on the years 1950-1999 and the ages 20-89. The red lines correspond to projections when break points are not allowed for, the blue lines correspond to projections when break points are allowed for.

90% confidence intervals for mortality projections five years ahead, the blue areas correspond to ten year ahead projections, and yellow areas correspond to twenty years ahead projections. The x-axis represents the years for which the projections are made.

The Lee-Carter model has the smallest confidence intervals, but they do not often capture the realised mortality rates. The other models perform better, though all of them have difficulties creating accurate mortality forecasts twenty years into the future. At $x = 65$, an important age for pension funds, this problem is most prominent for M5 and M9, and for $x = 75$ this is most prominent for M2. We conclude that for this specific data set, M3 and M6 produce accurate mortality predictions for all ages, even at longer projection horizons.

Ages 20 to 89

Figure 9 shows the back test for Dutch males for the models M1, M2, M3 and M9 for $x = \{35, 45, 65, 85\}$, calibrated on the period 1950-1980(1)2003 and the ages 20-89. Again, the Lee-Carter model has the smallest confidence intervals and the corresponding confidence intervals do not often capture the realised mortality rates. The projections from M2 are very volatile, even when the cohort effect does not need to be projected (because estimated cohort effects can be used in the projection). Further, similar graphs without allowing for break points show similar patterns. Hence, the unrealistic projections from M2 are a result of non robust parameter estimates. Similar findings result when we apply the same back test to England and Welsh men.

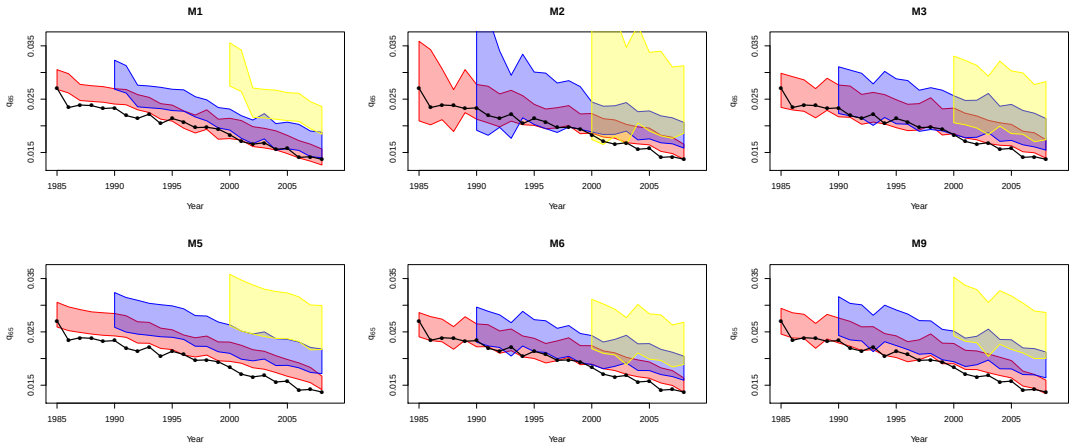
The projections from M9 are more stable than those from M2, and the confidence intervals from those projections capture the realised mortality rates. The projections from M3 are even more robust, but for $x = 25$ the confidence interval for twenty year ahead projections does not capture the realised mortality rates. Hence, we conclude that both M3 and M9 perform well on the age range 20-89.

5 Conclusions

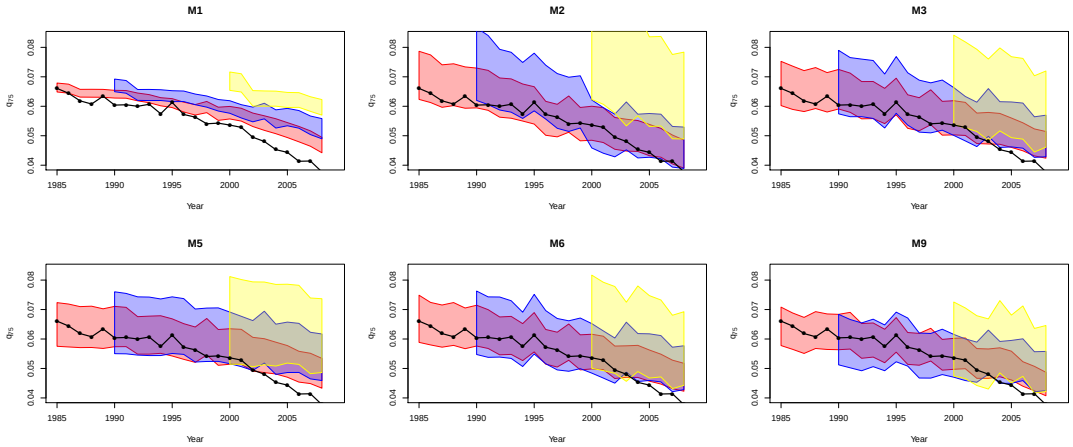
In this paper we estimate a selection of stochastic mortality models to historical mortality rates from the Netherlands and Belgium. We project mortality rates using a modelling strategy that allows to detect (objectively) multiple structural changes in the time series of time dependent parameters. As a result, the mortality rate projections are less sensitive to the calibration period, and the projections for the period effects connect better with the observed trends.

However, an expanding calibration period back test shows that none of the models considered is able to fully capture the mortality improvements for all ages simultaneously, not on the short horizon nor on the longer horizon. This result is also confirmed by Dowd et al. (2010) for England and Wales males. For Dutch males the Lee-Carter model is unable to create accurate mortality forecasts for all ages. The Renshaw-Haberman model (M2: Renshaw and Haberman (2006)) leads to non-robust mortality projections when applied to the ages 20-89. For this age range the APC-model (M3: Currie (2006)) and Plat's model (M9: Plat (2009)) perform well, and for the age range 60-89 the APC-model and the CBD-model (M6: Cairns et al. (2006)) result in most accurate mortality projections.

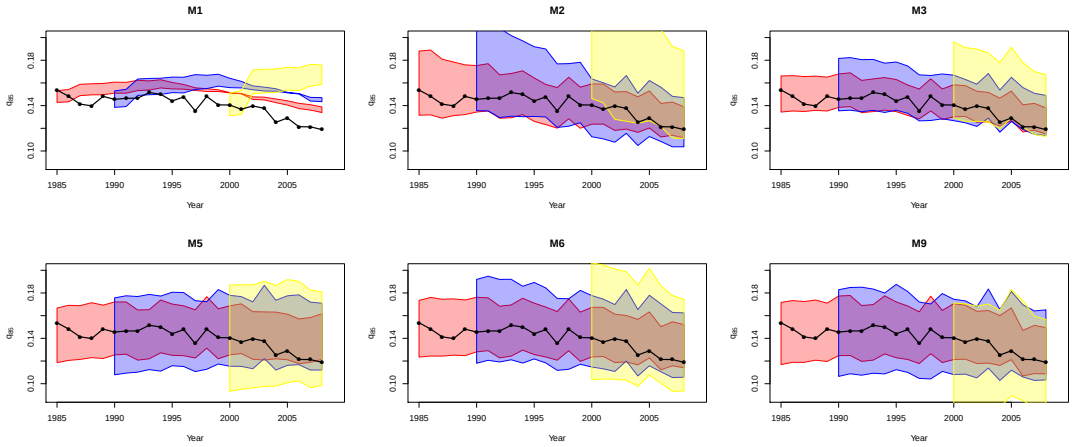
Our break point detection methodology leads to more appropriate projections of the period effects. However, the projection of the cohort effect remains difficult. Improvement in the projection of the cohort effect is left for future research. Furthermore, we compare the out-of-sample performance using an expanding calibration period back test, but it remains difficult to determine which model performs best over all ages. It is desirable to summarise information for different ages, for different projection horizons in one statistic. The development of such a statistic is also left for future research.



(a) 65 year old

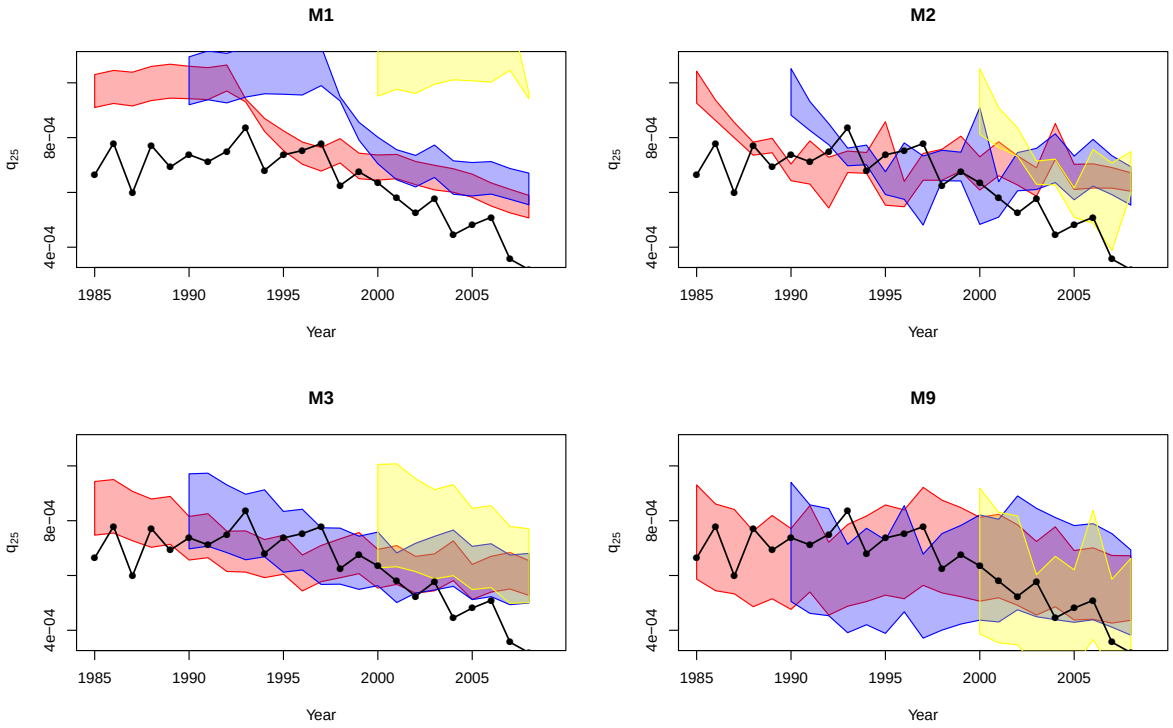


(b) 75 year old

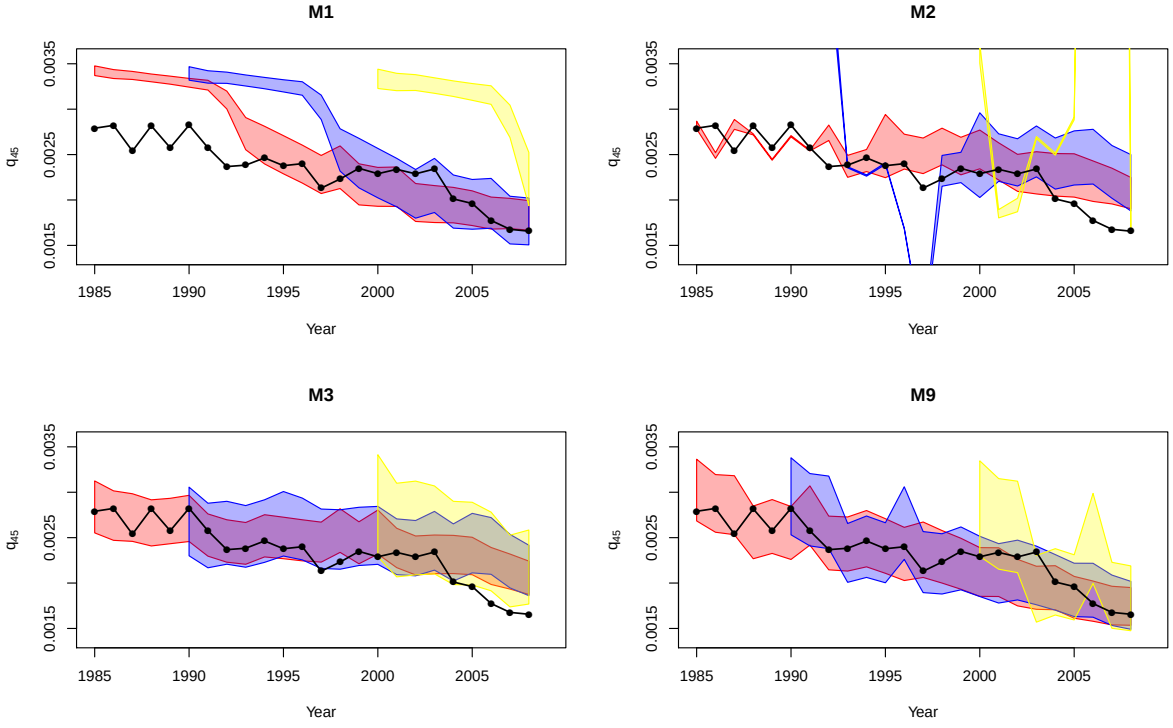


(c) 85 year old

Figure 8: Expanding calibration period back test for Dutch males, calibrated on the years 1950-1980(1)2003 and the ages 60-89. The red areas correspond to 90% confidence intervals for five year ahead mortality projections, the blue areas correspond to ten year ahead projections, and the yellow areas correspond to twenty year ahead projections. The black lines represent the realised mortality rates.

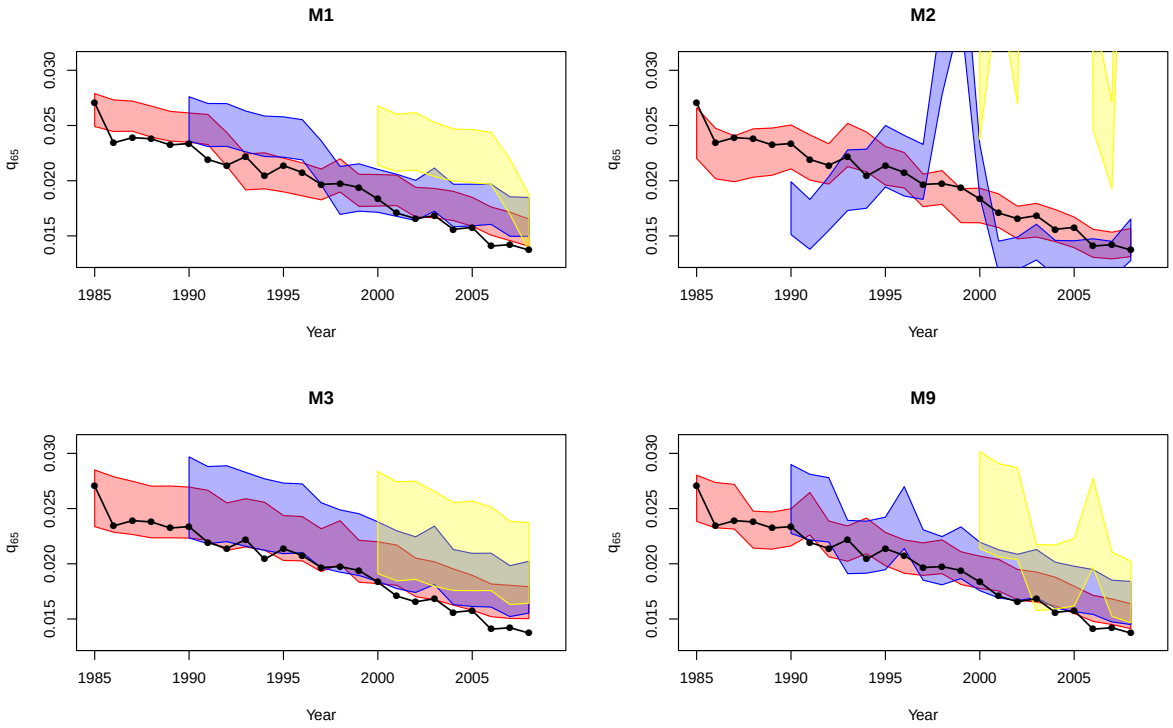


(a) 25 year old

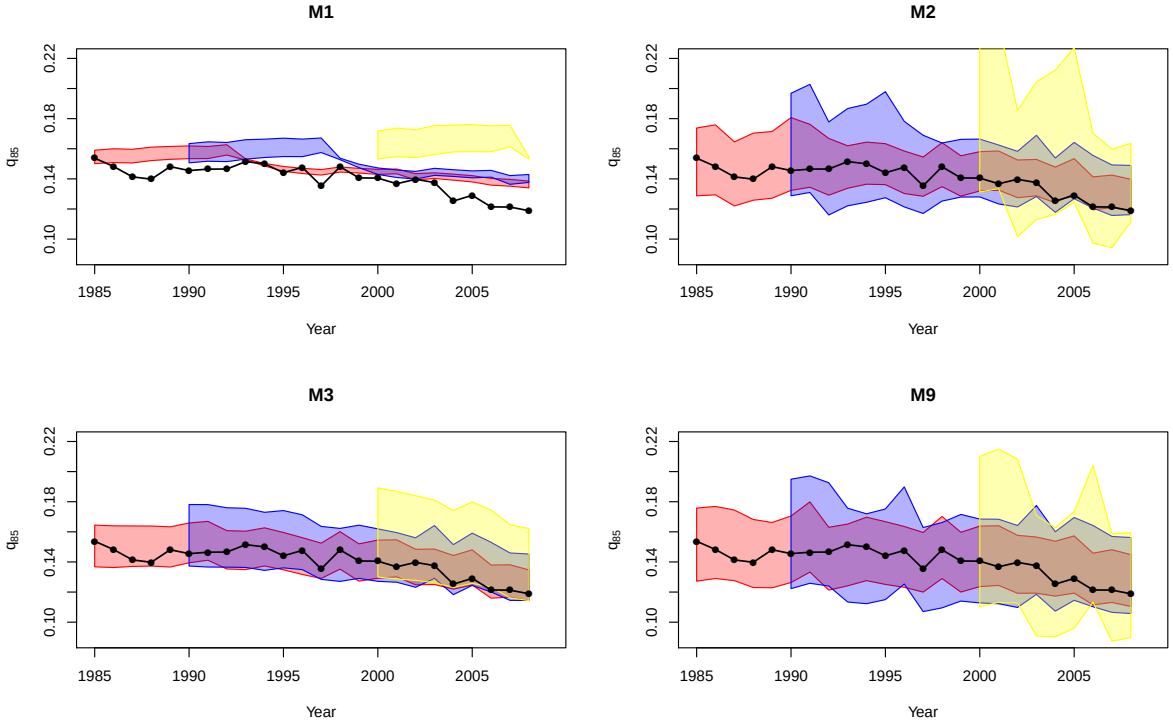


(b) 45 year old

Figure 9: Expanding calibration period back test for Dutch males, calibrated on the years 1950-1980(1)2003 and the ages 20-89. The red areas correspond to 90% confidence intervals for five year ahead mortality projections, the blue areas correspond to ten year ahead projections, and the yellow areas correspond to twenty year ahead projections. The black lines represent the realised mortality rates.



(c) 65 year old



(d) 85 year old

Figure 9: Expanding calibration period back test for Dutch males, calibrated on the years 1950-1980(1)2003 and the ages 20-89. The red areas correspond to 90% confidence intervals for five year ahead mortality projections, the blue areas correspond to ten year ahead projections, and the yellow areas correspond to twenty year ahead projections. The black lines represent the realised mortality rates.

Ackownledgements

Frank van Berkum would like to thank Anja de Waegenaere, Bertrand Melenberg, and Steven Haberman and other participants at the PARTY2013 workshop in Ascona (Switzerland) for their fruitful comments and suggestions. Katrien Antonio acknowledges financial support from NWO through a Veni 2009 grant and from AG Insurance through the AG Insurance Research Chair at KU Leuven. Frank van Berkum and Michel Vellekoop acknowledge financial support from Netspar.

References

J. Bai. Estimating multiple breaks one at a time. *Econometric Theory*, 13(03):315–352, 1997.

J. Bai and P. Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

J. Bai and P. Perron. Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22, 2003.

H. Booth, J. Maindonald, and L. Smith. Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56(3):325 – 336, 2002.

M.L. Bots and D.E. Grobbee. Decline of coronary heart disease mortality in the Netherlands from 1978 to 1985: contribution of medical care and changes over time in presence of major cardiovascular risk factors. *Journal of cardiovascular risk*, 3(3):271–276, 1996.

N. Brouhns, M. Denuit, and J.K. Vermunt. A poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31:373 – 393, 2002.

A.J.G. Cairns, D. Blake, , and K. Dowd. A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718, 2006.

A.J.G. Cairns, D. Blake, K. Dowd, G.D. Coughlan, D. Epstein, A. Ong, and I. Balevich. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1 – 35, 2009.

A.J.G. Cairns, D. Blake, K. Dowd, G.D. Coughlan, D. Epstein, and M. Khalaf-Allah. Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, 48: 355 – 367, 2011.

G.C. Chow. Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3): 591–605, 1960.

E. Coelho and L.C. Nunes. Forecasting mortality in the event of a structural change. *Journal of the Riyal Statistical Society*, 174(3):713 – 736, 2011.

I.D. Currie. Smoothing and forecasting mortality rates with P-splines. Talk given at the Institute of Actuaries, 2006. URL http://www.ma.hw.ac.uk/~iain/research/talks.html.

M. Denuit and A.C. Goderniaux. Closing and projecting lifetables using log-linear models. *Bulletin of the Swiss Association of Actuaries*, pages 29 – 49, 2005.

K. Dowd, A.J.G. Cairns, D. Blake, G.D. Coughlan, D. Epstein, and M. Khalaf-Allah. Back testing stochastic mortality models: an ex post evaluation of multiperiod-ahead density forecasts. *North American Actuarial Journal*, 14(3):281 – 298, 2010.

S. Haberman and A. Renshaw. A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48(1):35 – 55, 2011.

D. Hainaut. Multi dimensional Lee-Carter model with switching mortality processes. *Insurance: Mathematics and Economics*, 50(2):236 – 246, 2012.

D. Harris, D.I. Harvey, S.J. Leybourne, and A.M.R. Taylor. Testing for a unit-root in the presence of a possible break in trend. *Econometric Theory*, 25:1545 – 1588, 2009.

D.I. Harvey, S.J. Leybourne, and A.M.R. Taylor. Simple, robust and powerful tests of changing trend hypothesis. *Econometric Theory*, 25:995 – 1029, 2009.

R.D. Lee and L.R. Carter. Modeling and forecasting U. S. mortality. *Journal of the American Statistical Association*, 87(419):659–671, 1992.

J.S.-H. Li, W.-S. Chan, and S.-H. Cheung. Structural changes in the Lee-Carter mortality indexes: detection and implications. *North American Actuarial Journal*, 15(1):13 – 31, 2011.

E. Lovász. Analysis of Finnish and Swedish mortality data with stochastic mortality models. *European Actuarial Journal*, 1:259–289, 2011.

A. Milidonis, Y. Lin, and S.H. Cox. Mortality regimes and pricing. *North American Actuarial Journal*, 15(2):266 – 289, 2011.

R. Moreno-Serra and A. Wagstaff. System-wide impacts of hospital payment reforms: evidence from Central and Eastern Europe and Central Asia. *Journal of Health Economics*, 29(4):585 – 602, 2010.

C. O'Hare and Y. Li. Explaining young mortality. *Insurance: Mathematics and Economics*, 50(1):12 – 25, 2011.

R. Plat. On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393 – 404, 2009.

A.E. Renshaw and S. Haberman. Lee-Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33:255 – 272, 2003.

A.E. Renshaw and S. Haberman. A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, 38(3):556 – 570, 2006.

P.J. Sweeting. A trend-change extension of the Cairns-Blake-Dowd Model. *Annals of Actuarial Science*, 5(2):143 – 162, 2011.

Y.-C. Yao. Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, 6(3):181 – 189, 1988.

A. Zeileis and C. Kleiber. Validating multiple structural change models - a case study. *Journal of Applied Econometrics*, 20(5):685–690, 2005.

A. Zeileis, C. Kleiber, W. Krämer, and K. Hornik. Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, 44(12):109 – 123, 2003.

A Parameter constraints

Some of the mortality models experience identifiability issues. Therefore, we impose parameter constraints. Table 6 provides an overview of the parameter constraints that are imposed on the models.

Model	Constraints				
M1	$\sum_x \beta_x^{(2)} = 1$	$\sum_t \kappa_t^{(2)} = 0$			
M1A	$\sum_x \beta_x^{(2)} = 1$	$\sum_t \kappa_t^{(2)} = 0$	$\sum_x \beta_x^{(3)} = 1$	$\sum_t \kappa_t^{(3)} = 0$	
M2	$\sum_x \beta_x^{(2)} = 1$	$\sum_t \kappa_t^{(2)} = 0$	$\sum_x \beta_x^{(3)} = 1$	$\sum_{t,x} \gamma_{t-x} = 0$	
M2A	$\sum_x \beta_x^{(2)} = 1$	$\sum_t \kappa_t^{(2)} = 0$	$\sum_x \beta_x^{(3)} = 1$	$\sum_t \kappa_t^{(3)} = 0$	$\sum_{t,x} \gamma_{t-x} = 0$
M3	$\sum_t \kappa_t^{(2)} = 0$	$\sum_{t,x} \gamma_{t-x} = 0$			
M5	-				
M6	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$			
M7	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$	$\sum_c c^2\gamma_c = 0$		
M8	$\sum_{t,x} \gamma_{t-x} = 0$				
M9	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$	$\sum_t \kappa_t^{(3)} = 0$		
M10	-				
M11	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$	$\sum_c c^2\gamma_c = 0$	$\sum_t \kappa_t^{(4)} = 0$	
M12	$\sum_{t,x} \gamma_{t-x} = 0$				
M13	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$	$\sum_t \kappa_t^{(3)} = 0$		

Table 6: Overview of the parameter constraints imposed on the models.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be