KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

The impact of multiple structural changes on mortality predictions

Frank van Berkum, Katrien Antonio, Michel Vellekoop

AFI_1487

The impact of multiple structural changes on mortality predictions

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

January 2014

Abstract

Most mortality models proposed in recent literature rely on the standard ARIMA-framework (in particular: a random walk with drift) to project mortality rates. As a result the projections are highly sensitive to the calibration period. We apply a modelling strategy for the time-dependent parameters in a large collection of mortality models that allows for objective, statistical detection of one or more structural changes. By comparing projections based on different calibration periods and different time series specifications, we show that our proposed methodology leads to more robust mortality projections.

Key words: Stochastic mortality; structural changes; mortality forecasting; backtesting

*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 as discussed in, for example, Cairns et al. (2008) and Barrieu et al. (2012). Life insurance companies and pension funds need to monitor and predict mortality improvements for proper pricing and reserving. It is also important to quantify the uncertainty in future mortality rates, for regulatory purposes such as Solvency II.

Constructing mortality rate projections consists of two steps, namely (i) estimating a mortality model on historical data, and (ii) forecasting the time dependent parameters obtained in (i). The seminal paper by Lee and Carter (1992) introduces a stochastic mortality model that allows for mortality improvements. This is a single factor model with age and period effects. Several extensions have been proposed to the Lee-Carter model, such as the introduction of a cohort effect (Renshaw and Haberman (2006); Currie (2006)), functional forms for 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)).

The modelling of time-dependent effects in mortality models is underexposed in recent literature. The period and cohort effects are often projected using ARIMA-models. However, when structural changes are present, the time-dependent effects cannot always be captured by standard ARIMA-models. The resulting mortality forecasts are highly sensitive to the calibration period. Alternatives have been proposed to tackle this problem, e.g. Booth et al. (2002) and Denuit and Goderniaux (2005) use a deterministic approach and Li et al. (2013) a Bayesian approach to choose an optimal calibration period, Milidonis et al. (2011) introduce regime switching models to mortality modelling, and Li et al. (2011), Sweeting (2011) and Coelho and Nunes (2011) introduce structural changes in trend and difference stationary processes.

In this paper we extend the approach of Coelho and Nunes (2011). They allow for a single structural change in period effects. However, multiple structural changes may have occurred, as suggested for trend stationary processes by Sweeting (2011). We focus on the class of difference stationary processes as the majority of the above-mentioned literature does. When extending the approach of Coelho and Nunes (2011) by allowing for multiple structural changes in the period effects, we determine the structural changes in an objective manner (Bai and Perron (1998)). The optimal number of structural changes is selected using the Bayesian Information Criterion. As an example, we compare the projections from this approach to those obtained when no structural changes or a single structural change is allowed using the Dawid-Sebastiani scoring rule (Riebler et al. (2012)). Whereas the aforementioned papers often focus on a specific mortality model, we show results for Dutch and Belgian mortality data, calibrated to a wide variety of mortality models. We include both models with and without cohort effects since recent results by Coelho and Nunes (2013) show that evidence of structural changes in models without cohort effects may disappear once cohort effects have been included.

The remainder of this article is organised as follows. First, in Section 2 we introduce different mortality models and we review methods used for mortality forecasting. In Section 3 we present our approach for mortality forecasting when allowing for multiple structural changes within the period effects. We investigate the estimation and backtesting results in Section 4, and Section 5 concludes.

2 Literature review

We start with an overview of mortality models from recent literature. Then we review the literature on forecasting period and cohort effects when modelling mortality.

Model	Name	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.

2.1 Mortality model structures

Let the number of deaths during calendar year t aged x at death be $d_{t,x}$, and the average population aged x during calendar year t be $e_{t,x}$. The death rate, $m_{t,x}$, is defined by

$$m_{t,x} = \frac{d_{t,x}}{e_{t,x}}. \tag{1}$$

The probability that a person aged exactly x at the beginning of calendar year t dies within the next year is called the mortality rate $q_{t,x}$. The force of mortality $\mu_{t,x}$ is the instantaneous death rate at exact time t for individuals aged exactly x at time t. If we assume that $\mu_{t,x}$ is constant in the interval $[t, t+1) \times [x, x+1)$, the mortality rate is linked to the death rate by the approximation (see Cairns et al. (2009)):

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

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.

As in Brouhns et al. (2002), we assume a Poisson distribution for the number of deaths within a year, $D_{t,x} \sim \text{Poisson}(e_{t,x}m_{t,x})$. The various specifications for $m_{t,x}$ are listed in Table 1. Here, $b(x) = \left((x-\bar{x})^2 - \frac{1}{n}\sum_{x_i=x_1}^{x_n}(x_i-\bar{x})^2\right)$ where x_i are the ages included in the data set, $c(x) = (\bar{x}-x)^+ + [(\bar{x}-x)^+]^2$, $\bar{x}$ is the average of the ages, and x_c is a constant which can be chosen up front or can be estimated; in this paper we estimate this parameter[1]. For each of these models we specify the likelihood

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

and apply standard Newton-Raphson steps to maximise this likelihood. Since most models may involve identification issues, we apply the parameter constraints as proposed in the 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, whereas the other models are calibrated both for the ages 20-89 and the ages 60-89.

2.2 Forecasting mortality

We give an overview of standard ARIMA time series models, extensions to the standard ARIMA models, and other time series models used for forecasting mortality.

Standard ARIMA-models

Cairns et al. (2011) consider the models M1 to M5, M7 and M8. These models are fitted to England and Wales mortality data from the years 1961 to 2004. For the period effects they fit a (uni/multi)variate random walk with drift. For the cohort effects they estimate different ARIMA(p, d, q)-specifications. The specifications used in backtesting are selected based on biological reasonableness of the projections and the BIC. Second order differencing of the cohort effect ($d = 2$) leads to large confidence intervals which the authors find less plausible. For the data under consideration a mean reverting process (AR(1)) or an ARIMA$(1, 1, 0)$ process (both including a constant) is most appropriate 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, because he argues that this may lead to biologically unreasonable projections. Therefore, a mean reverting process is fitted with non-zero mean (AR(1) with a constant).

Plat (2009) considers two approaches for calibrating cohort effects: (i) estimate the cohort effect for all cohorts available, and (ii) 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 should therefore not be used to project mortality rates for the elderly[2]. The cohort effect is then projected using a mean reverting process with mean zero. As a result, there is no trend in the projected cohort effect.

Haberman and Renshaw (2011) consider the 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 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) by assuming a multivariate random walk with drift. For the cohort effects he chooses the ARIMA(p, d, q)-process

[2]In this paper, we set the cohort effects equal to zero for the models M9 and M13 when there are no observations available related to age 60 or higher, conform the idea in Plat (2009).

that is optimal in terms of BIC. He considers the combinations $d \in \{0, 1, 2\}$ and $(p, q) \in \{0, 1, 2\}$, and for those datasets 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 apply it to data from a range of developed countries from 1950 to 2006. The proposed model is a modification of Plat's model, and therefore 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 mentioned above all use a random walk with constant drift for the first period effect, and often also for the other period effects. However, different calibration periods are used 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). Furthermore, factors like medical advances (Bots and Grobbee (1996)) and health system reforms (Moreno-Serra and Wagstaff (2010)) can have an impact on the speed of the mortality improvements. Dropping the assumption of a random walk with a *constant* drift may therefore be a way to improve model performance.

Regime switching models

Milidonis et al. (2011) calibrate the Lee-Carter model on US data for the ages 0-100 in the years 1901-2005 (males and females combined). They propose a regime switching model with two regimes for the differenced series of κ_t. The two regimes are allowed to have a different mean as well as a different variance, and the estimation reveal that the variance differs substantially between the two regimes. Based on information criteria and a likelihood ratio test they conclude that for the data set considered the regime switching model outperforms the random walk with drift.

Hainaut (2012) extends the regime switching model to model M1A applied to French data for the ages 20-100 in the years 1946-2007 (males and females separately), and concludes that the improvement in loglikelihood is significant compared to the standard Lee-Carter model and the extension from Milidonis et al. (2011).

Structural changes in trend stationary models

Li et al. (2011) calibrate the Lee-Carter model on England & Wales and US data for the ages 0-99 in the years 1950-2006 (males and females combined). They perform a unit root test on the time series κ_t, which means that they test the null hypothesis of a random walk with constant drift versus the alternative hypothesis of a broken-trend stationary model. The broken-trend stationary model implies that the mortality trend κ_t fluctuates around a deterministic trend. The deterministic trend is piecewise linear and is estimated by regressing κ_t versus t and an intercept. Dummy variables are included in the regression such that the trend may change once in the data set, but the different trends do not have to be connected. For both data sets they conclude that a broken-trend stationary model is preferred over a random walk with constant drift model, and they use the latest trend for predictions. Since this is a trend stationary process, predictions from this model do not lead to confidence intervals which become wider over time.

Sweeting (2011) calibrates the original CBD-model (M5) on England & Wales data for the ages 60-89 in the years 1841-2005. He assumes a broken-trend stationary model as in Li et al. (2011), but he allows for multiple structural changes and imposes the different trends to connect. He then fits distributions to the frequency and the severity of the changes in the trend and uses these distributions for forecasting mortality. Structural changes are tested for significance using the Chow test (Chow (1960)).

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 for the ages 0-99 in the years after 1950[3]. They perform a unit root test as suggested by Harvey et al. (2009) and Harris et al. (2009) that allows for a single structural change both in the trend stationary and in the difference stationary model, where Li et al. (2011) only allow for a single structural change in the trend stationary model. They perform this analysis for 18 countries both for males and females. From all these data sets, the trend stationary model with possibly a structural change is rejected 33 out of 36 times in favour of a difference stationary model with possibly a structural change. Further, for 21 out of 36 data sets a structural change is detected.

O'Hare and Li (2012) investigate the impact of a single structural change on mortality models beyond Lee-Carter. They apply the methodology for difference stationary time series to the models M1 (Lee-Carter), M5 (CBD), M9 (Plat) and M13 (O'Hare and Li). They find that in mortality models other than the Lee-Carter model a structural change is often detected as well, and that allowing for a structural change can substantially improve the quality of forecasts, measured in Mean Absolute Error and Root Mean Squared Error.

3 Proposed forecasting method

3.1 Forecasting period effects

When regime switching models are applied to mortality models, it is known beforehand that mortality dynamics observed in the past will occur in the future. Changes in mortality dynamics may be a result of changes in lifestyle, healthcare systems, etc. For example, in the Netherlands changes in smoking habits have been an important driver in changes in mortality, which resulted in increasing (1950-1970) and decreasing (from 1970 onwards) mortality rates. Since we find it difficult to predict whether and how historical changes in mortality may occur again in the future, we will not use regime switching models.

Instead, we use recent information on mortality dynamics, but we use the entire calibration period to estimate the variability in the mortality dynamics. Following the findings from Coelho and Nunes (2011) and the fact that a random walk with drift is most prominent in the mortality literature, we focus on the difference stationary process. However, we extend the approach of Coelho and Nunes (2011) and the work of O'Hare and Li (2012) such that multiple structural changes can be detected, as multiple events in the past may have affected the speed of mortality improvements.

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 presence of structural changes, and to project the time dependent parameters.

1. Check for structural changes in each time series individually, and determine the break points if any;

2. Fit a (uni/multi)variate random walk with piecewise constant drift, given the break points determined in step 1;

3. Simulate paths from the time series model, given the parameter estimates from step 2.

To determine the break points, we follow the methodology introduced in Bai and Perron (2003). Suppose we have at our disposal a time series κ_t $(t = 1, \ldots, T)$ and define the first-order differences by $\Delta\kappa_t =$

[3]The data set depends on the data availability per country.

$\kappa_t - \kappa_{t-1}$. We estimate a random walk with a piecewise constant drift:

$$\Delta\kappa_t = \begin{cases} \beta_1 + \varepsilon_t, & t \leq t_1 \\ \dots & \\ \beta_i + \varepsilon_t & t_{i-1} < t \leq t_i \\ \dots & \\ \beta_{m+1} + \varepsilon_t, & t_m < t \end{cases} \tag{3}$$

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

$$\text{SSR}(t_1, \dots, t_m) = \sum_{i=1}^{m+1} \sum_{t=t_{i-1}+1}^{t_i} [\Delta\kappa_t - \beta_i]^2 \tag{4}$$

where $t_0 = 1$ and $t_{m+1} = T$. In the model specification above, we distinguish m break points, which divide the time series into $m+1$ periods with different drifts. Both the number of break points and the date of the break points are unknown.

Let $\beta(T_m)$ denote the estimates $\{\beta_1, \dots, \beta_{m+1}\}$ based on a given m-partition $(t_1, \dots, t_m)$ denoted T_m. If we substitute these parameter estimates $\beta(T_m)$ into (4), then the estimated break points $(\hat{t}_1, \dots, \hat{t}_m)$ are such that $(\hat{t}_1, \dots, \hat{t}_m) = \text{argmin}_{t_1,\dots,t_m} \text{SSR}(t_1, \dots, t_m)$, where the minimisation is taken over all partitions $(t_1, \dots, t_m)$ for which $t_i - t_{i-1} \geq h$. The parameter h corresponds to the minimum period that a regime should last, and is to be chosen up front. Bai and Perron (2003) describe an efficient algorithm to determine the optimal break points for a given m.

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.

Given the method described above, we can determine the optimal break points $(t_1, \dots, t_m)$ 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) using an information criterion like the BIC, and (ii) performing F-tests to test the significance of the improvement in fit when adding one or multiple break points.

If the information criterion is used, then one determines the BIC for all $m \in \{0, \dots, 5\}$[4], see Zeileis et al. (2003). Denote BIC(m) as the BIC corresponding to the optimal break points for a given m. The optimal number of break points is then defined by $m^* = \arg\max \text{BIC}(m)$.

As in Bai and Perron (1998, 2003), we may 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 hypothesis of $m = 1$ break points. If the null hypothesis of no break points is rejected, then one continues testing for the significance of two break points versus the null hypothesis of one break point, and so on. The F-statistic is a function of the restricted sum of squared residuals (RSSR) and the unrestricted sum of squared residuals (USSR), the null and alternative hypothesis, respectively:

$$F = \frac{(\text{RSSR} - \text{USSR})/(p_1 - p_0)}{\text{USSR}/(n - p_1)}, \tag{5}$$

where 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

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

the structural changes are unknown, we cannot use the standard critical values of the F-distribtion as used in Sweeting (2011), but critical values have to be obtained through simulation (see Andrews (1992)). 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 than one break point. Therefore, we shall not use the sequential F-test.

The second F-test 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, we determine the F-statistic as defined in (5) for all $k \in \{1, \ldots, 5\}$ which we denote by $F(k)$. We then define the UDmax test statistic as the maximum value of those F-statistics:

$$\text{UDmax} = \max_k F(k) \tag{6}$$

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 break points is equal to $k^* = \arg\max F(k)$. If the test statistic is smaller than the critical value, then there is insufficient proof for a structural change.

The latter F-test is close to using the BIC, because an optimal model is chosen while considering all model specifications. 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, and Bai and Perron (2003) note that the BIC performs well in the absence of serial correlation. We will therefore use the BIC to choose the number of break points. In the following paragraph we illustrate the method to Dutch male mortality.

Illustration - the Lee-Carter model

We consider the period effect of the Lee-Carter model, estimated on Dutch male mortality data for the period 1960 to 2008, for the ages 60 to 89. The top left graph in Figure 1 shows the parameter estimates for $\kappa_t^{(2)}$. A random walk with constant drift does not seem appropriate, because of apparent structural changes around 1972 and 2000. This is confirmed in the bottom left graph. The black lines correspond to projections from a random walk with constant drift and these projections suggest that $\kappa_t^{(2)}$ had values above its expected value in all years 1960-2008. The blue lines correspond to projections when one structural change is allowed; the break point is dated at 1993, between the two dates expected by visual inspection. These projections are not unreasonable, but the drift of the period effect does not appear to be piecewise constant before and after the break point. If we allow for multiple structural changes, then we obtain the projections represented by the red lines. The break points are estimated at 1972 and 2002. The projections look reasonable, because the drift of the period effect is piecewise constant between the different break points, and the lines connecting the break points are not always below or above the observed values.

The graphs on the right-hand side of Figure 1 show the projections for the period effect using different calibration periods. We compare scenarios without structural changes, with a single structural change and with multiple structural changes. Allowing for a single structural change leads to more robust projections with respect to the calibration period, and if we allow for multiple structural changes, projections become even more robust.

Figure 2 shows the first order differences of the estimated period effect from Figure 1 (top left). From the upper right graph we observe that the first break point is accurately estimated, since the confidence interval[5] (shown by the red line) is narrow. The lower left graph in Figure 2 shows the confidence intervals for the case of two break points. The second break point (around the year 2002) is estimated accurately, but the confidence interval corresponding to the first break point is wide. This can

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

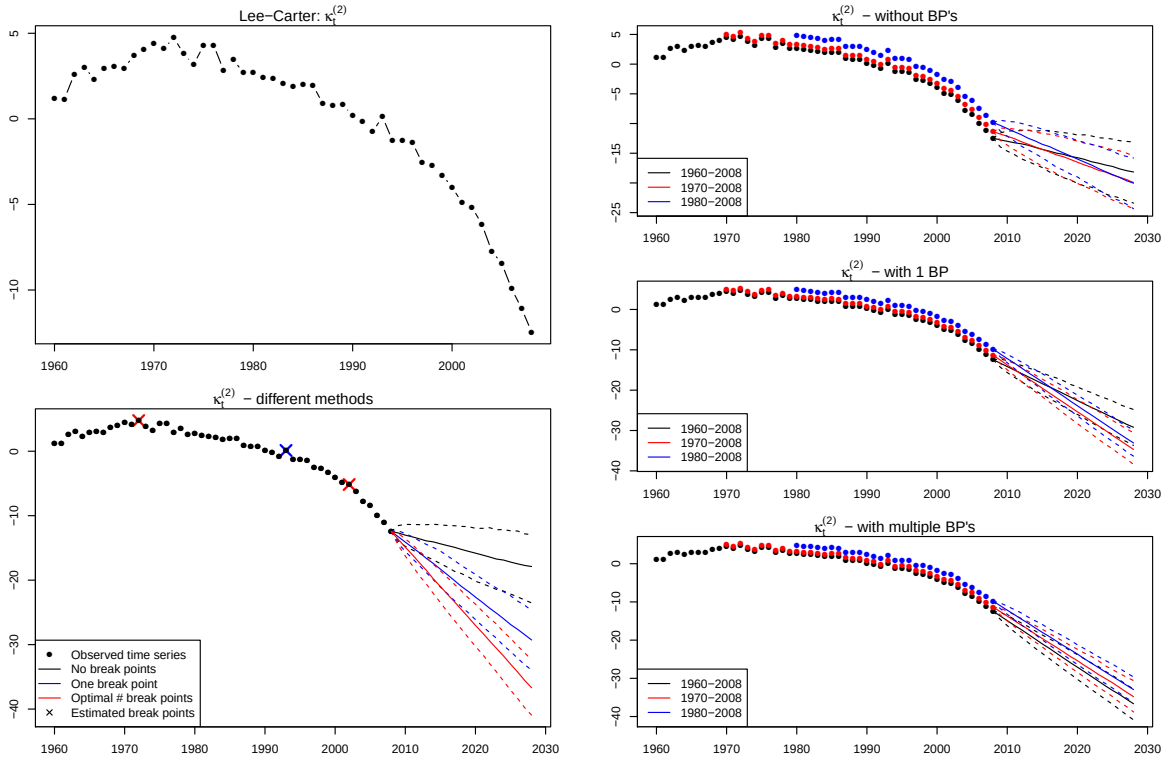

Figure 1: Top left: parameter estimates of $\kappa_t^{(2)}$ in the Lee-Carter model, calibrated on data from Dutch males aged 60-89 in the period 1960-2008. Bottom left: projections for the period effect using different projection methods. Top right through bottom right: projections of the period effect for different calibration periods without allowing for structural changes, allowing for one structural change and allowing for multiple structural changes. Dots are estimated parameters, solid lines are the 50th percentile and dotted lines are the 5th and 95th percentiles of the projections.

be explained by the outliers before and after the year 1972. However, allowing for the second break point leads to an improvement in fit over the whole observation period. This is illustrated by the differences between the green and blue lines in Figure 2. 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.

3.2 Forecasting cohort effects

Section 2.2 contains an overview of different approaches to project the cohort effect. Imposing an ARIMA-specification up front can lead to biologically unreasonable forecasts. Therefore, we use the BIC, but we only consider ARIMA(p, d, q)-specifications for $d \in \{0, 1\}$ and $(p, q) \in \{0, 1, 2\}$. We do not consider the case $d = 2$, because from Cairns et al. (2011) we conclude that using a second order differencing model leads to implausibly large confidence intervals.

4 Results

In this section we calibrate the mortality models from Table 1 to Dutch and Belgian mortality data. Then we perform an out-of-sample backtest to investigate the predictive properties of the models while allowing for no, a single or multiple structural change(s).

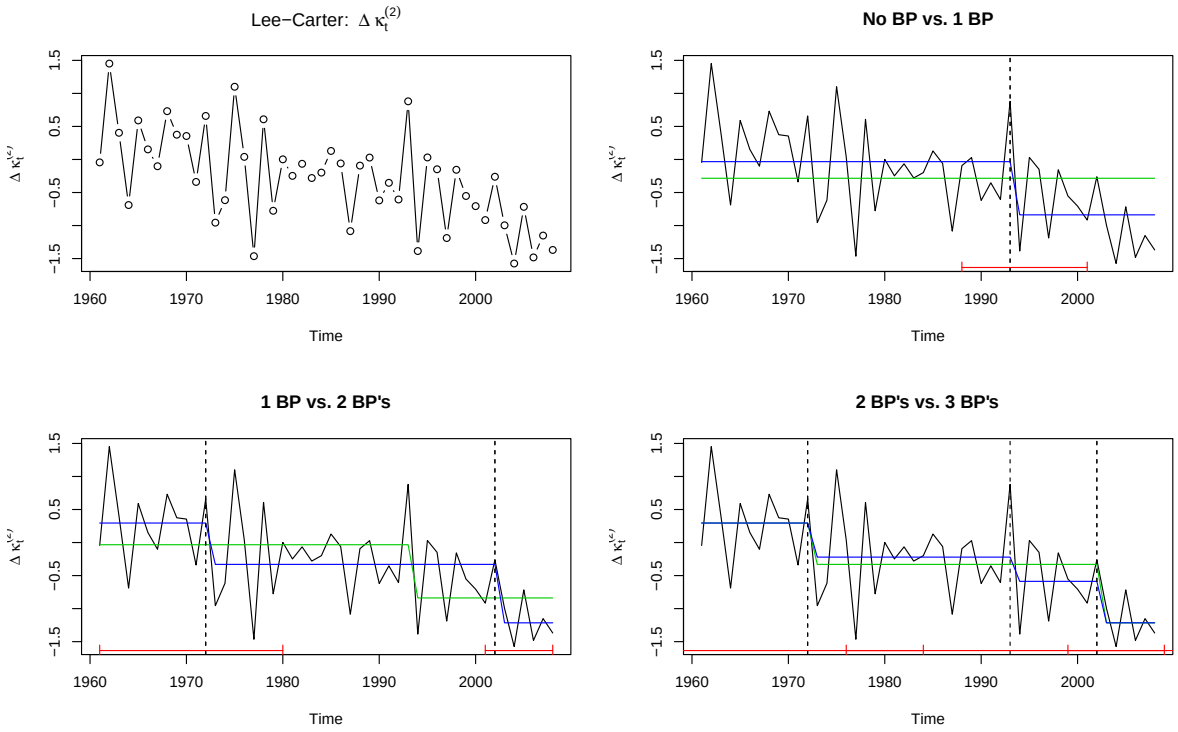

Figure 2: Confidence intervals for estimated break points for $\kappa_t^{(2)}$ in the Lee-Carter model, calibrated on Dutch males aged 60-89 in the years 1960-2008. In the plots (i) BP's vs. $(i+1)$ BP's the green lines represent the mean of $\Delta\kappa_t^{(2)}$ for the different periods when (i) BP's are allowed, and the blue lines represent the mean of $\Delta\kappa_t^{(2)}$ when $(i+1)$ BP's are allowed. The red lines represent the confidence intervals corresponding to the break points.

4.1 Model fit

We calibrate the models on male mortality data[6] from the Netherlands and Belgium for the years 1950 to 2008. Earlier data is excluded such that there are no world wars in the data set. We consider the ages 20-89, because mortality rates for younger ages are not relevant for insurers and pension funds, and mortality rates for ages above 89 are less reliable and are therefore excluded. If mortality rates are needed for higher ages, multiple techniques are available to close mortality tables; see e.g. Vaupel (1990), Lindbergson (2001) and Denuit and Goderniaux (2005).

We present the estimation results for Dutch and Belgian males[7] for ages 20-89 in Table 2 and for ages 60-89 in Table 3. These tables contain the effective[8] number of parameters that is estimated in each of the models, and the corresponding BIC that we define as $\text{BIC} = \log L - \frac{1}{2}k \cdot \log n$, where $\log L$ is the loglikelihood, n is the number of observations, and k is the effective number of parameters. Higher BIC's mean better performance of the models.

For the age range 20-89, the models with a cohort effect and interaction between age and period effects have the highest BIC. The ranking of the models for Dutch males is similar to the ranking for Belgian males. However, some models that score well on the age range 20-89 score worse for the age

[6] Human Mortality Database is a joined project of the University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany). Data are available at http://www.mortality.org.

[7] Similar results for Dutch and Belgian females are available upon request from the authors.

[8] 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.

	The Netherlands			Belgium		
Model	Number of parameters	BIC	Rank	Number of parameters	BIC	Rank
M1	197	-22,623.00	10	197	-22,955.00	10
M1A	324	-20,559.09	8	324	-21,146.23	8
M2	385	-19,641.72	5	385	-20,345.36	6
M2A	513	-20,059.54	7	513	-20,615.64	7
M3	246	-19,724.25	6	246	-20,315.39	5
M9	327	-19,392.03	2	327	-19,918.98	2
M10	244	-20,676.05	9	244	-22,189.76	9
M11	422	-19,591.33	4	422	-20,143.62	4
M12	364	-19,439.22	3	364	-19,990.09	3
M13	327	-19,391.52	1	327	-19,906.79	1

Table 2: 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	-11,355.25	14	117	-10,741.05	14
M1A	204	-9,761.62	13	204	-10,222.80	12
M2	225	-9,411.94	4	225	-9,605.78	4
M2A	313	-9,675.49	11	313	-9,850.17	9
M3	166	-9,395.01	3	166	-9,555.03	2
M5	118	-9,667.92	10	118	-10,234.93	13
M6	196	-9,268.00	1	196	-9,470.91	1
M7	254	-9,428.90	5	254	-9,632.33	5
M8	198	-9,333.30	2	198	-9,572.43	3
M9	284	-9,527.90	8	284	-9,718.64	7
M10	204	-9,465.34	6	204	-9,905.13	11
M11	342	-9,705.58	12	342	-9,899.65	10
M12	284	-9,559.58	9	284	-9,777.92	8
M13	284	-9,524.37	7	284	-9,715.78	6

Table 3: Estimation results for Dutch and Belgian male mortality rates, estimated on the age range 60 to 89 and calibration period 1950-2008.

		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 4: 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.

range 60-89 (M9, M11, M12 and M13) and vice versa (M2 and M3). The ranking of the models for the age range 60-89 is again similar for the Dutch and Belgian males.

Table 4 shows the results for optimisation over the parameter x_c. Given the model specification of M8 and M12, the importance of the cohort effect for different ages depends on the parameter x_c. We conclude that 4 out of 12 times the cohort effect mainly affects younger ages as it is estimated optimally at the upper bound ($x = 89$), and 8 out of 12 times the cohort effect mainly affects the elderly.

For illustration purposes we present the parameter estimates for M2 estimated on Dutch male mortality data in Figure 3 since this model fits the data reasonably well for both age ranges. The parameter estimates for the two age ranges are similar and the fitted mortality rates differ only marginally. In order to project mortality, the parameter $\kappa_t^{(2)}$ needs to be projected into the future, and for new cohorts we also have to project the cohort effect γ_{t-x}. As the time-dependent parameters are different, it is possible that mortality projections resulting from the two different age ranges are different, regardless of the similar in-sample fit.

4.2 Out-of-sample performance

We now evaluate the predictive power of the models under consideration. We calibrate the models using data from 1950 to 2000 and then simulate mortality rates for the years 2001 to 2008. This leads to a predictive distribution for the stochastic mortality rates $M_{t,x}$ for $x = x_1, \ldots, x_n$ and $t = T+1, \ldots, T+s$. As in Riebler et al. (2012), we obtain the mean $\mathbb{E}(M_{t,x})$ and variance $\mathrm{Var}(M_{t,x})$ of future mortality rates from the simulated predictive distribution. With $D_{t,x} \sim \mathrm{Poisson}(e_{t,x}M_{t,x})$ and using the law of total expectation it follows that for $t > T$ the expected death counts are

$$\widehat{d}_{t,x} = \mathbb{E}(D_{t,x}) = e_{t,x}\mathbb{E}(M_{t,x})$$

and the variance of the death counts is

$$\begin{aligned}\sigma^2_{t,x} = \mathrm{Var}(D_{t,x}) &= \mathbb{E}(\mathrm{Var}(D_{t,x}|M_{t,x})) + \mathrm{Var}(\mathbb{E}(D_{t,x}|M_{t,x})) \\ &= \mathbb{E}(e_{t,x}M_{t,x}) + \mathrm{Var}(e_{t,x}M_{t,x}) \\ &= e_{t,x}\mathbb{E}(M_{t,x}) + e^2_{t,x}\mathrm{Var}(M_{t,x}),\end{aligned}$$

since we assume the population size $e_{t,x}$ given[9]. In our evaluation of the out-of-sample performance we consider the differences between observations and projections (hereafter: calibration of the projections), and the width of the confidence intervals of the projections (hereafter: sharpness of the projections). We compare the calibration of the mortality models using the root mean squared error (RMSE), both with

[9] We shall not simulate the population size, because then assumptions needs to be made on immigration and emigration.

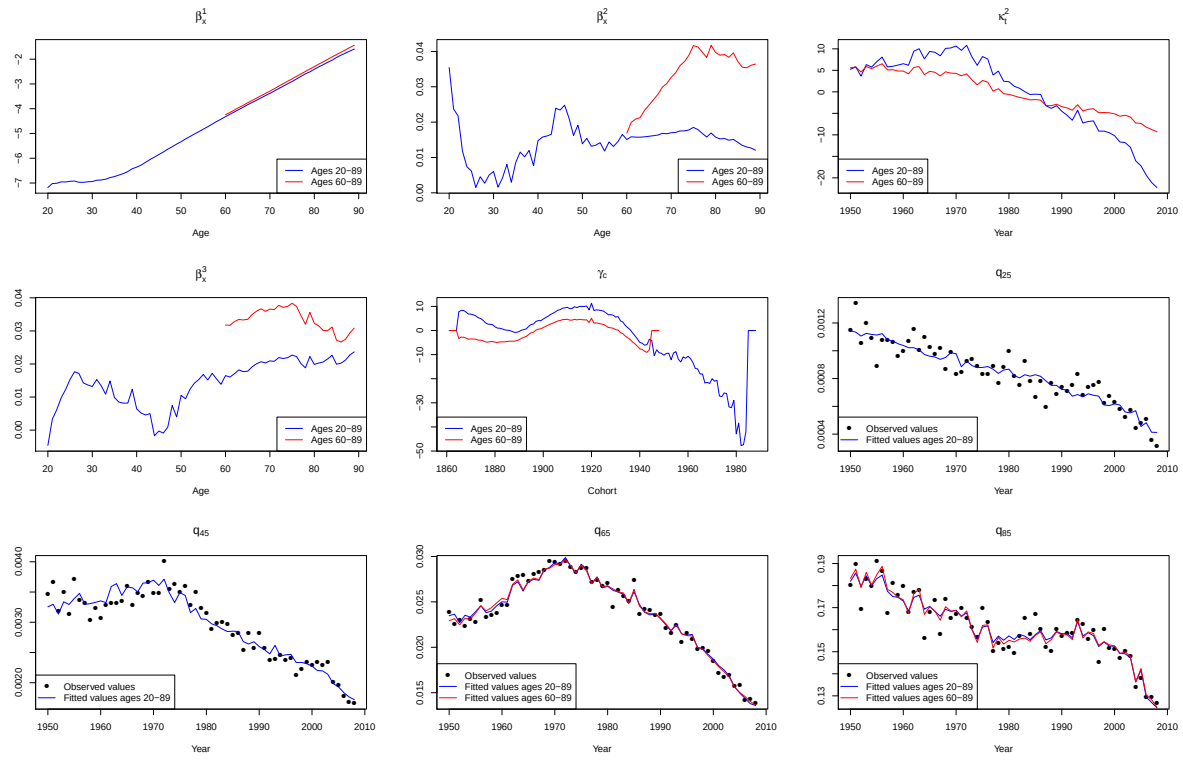

Figure 3: The first five panels show the parameter estimates for M2 calibrated on Dutch male mortality in the years 1950 to 2008 on the ages 20-89 and 60-89. The last four panels show realised mortality rates (dots) and fitted mortality rates for $x = \{25, 45, 65, 85\}$ (calibrated on ages 20-89 and ages 60-89)

and without the possibility of structural changes:

$$\text{RMSE} = \sqrt{\frac{1}{n \cdot s} \sum_{t,x} \left(d_{t,x} - \widehat{d}_{t,x} \right)^2}. \tag{7}$$

The RMSE only accounts for differences between observations and predictions, but not for differences in scale. A typical problem for mortality data is to summarise the quality of the forecasts for different ages and years in a single statistic. The death counts that we consider differ in scale for different ages and years due to different mortality rates and population sizes. The Dawid-Sebastiani scoring rule (DSS) introduced by Gneiting and Raftery (2007) is a statistic that evaluates the calibration and the sharpness of the projections, and also takes the scale of the observations into account. We compute the average DSS ($\overline{\text{DSS}}$) as introduced by Riebler et al. (2012), which allows us to summarise the quality of the forecasts into one statistic for all forecasted death counts:

$$\overline{\text{DSS}} = \frac{1}{n \cdot s} \sum_{t,x} \left(\frac{d_{t,x} - \widehat{d}_{t,x}}{\sigma_{t,x}} \right)^2 + \log \sigma_{t,x}^2. \tag{8}$$

Table 5 and 6 show the backtesting results for Dutch and Belgian females for the ages 20-89 and 60-89 respectively, and Table 7 and 8 show similar results for Dutch and Belgian males. For some models the statistics are lower when structural changes are incorporated (the bold figures in the tables), which means that allowing for structural changes has improved the quality of the mortality forecasts; especially the decrease in RMSE can be large. For other models however, the statistics are higher (the red figures), which means that the quality of the forecasts has worsened. Allowing for structural changes has little effect on the ranking of the models based on RMSE or $\overline{\text{DSS}}$, but the ranking of the models based on the backtest is markedly different from the ranking based on the fit of historical data in Table 2 and 3.

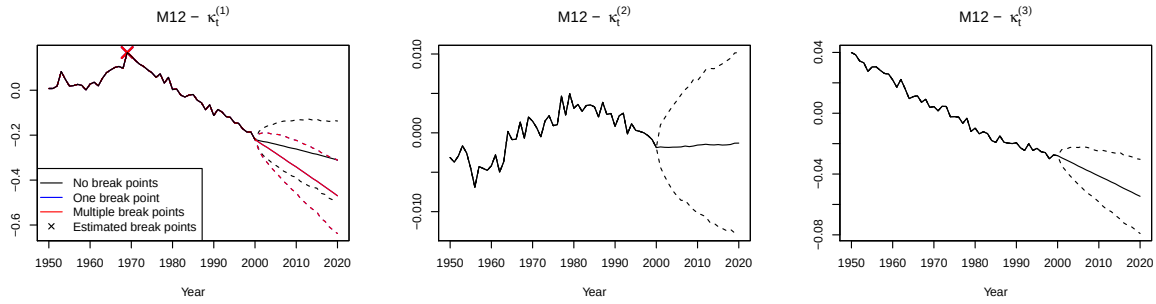

Figure 4: Projections for the period effects of M12 applied to Dutch females aged 20-89 in the period 1950-2000. The structural change for $\kappa_t^{(1)}$ is identified both if we allow for one and if we allow for multiple structural changes.

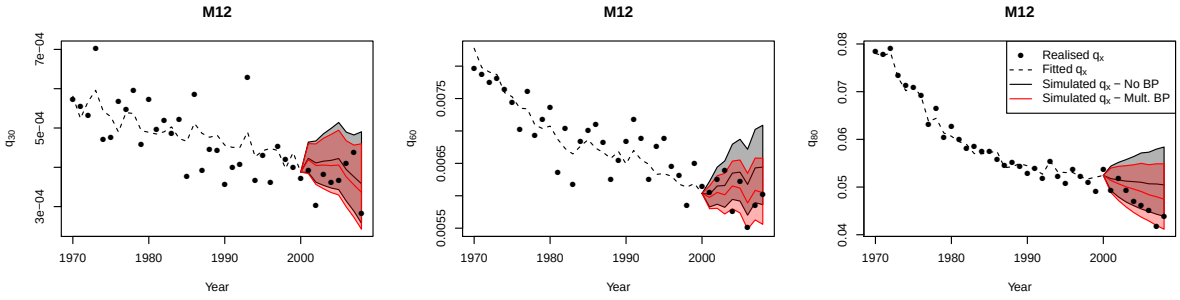

Figure 5: Mortality projections from M12 for $x = \{30, 60, 80\}$ calibrated on Dutch females aged 20-89 in the period 1950-2000. The black and red lines represent projections without and with multiple structural changes, respectively, at the 5th, 50th and 95th percentile.

Figure 4 shows projections of the period effects for M12 applied to Dutch females aged 20-89 and Figure 5 shows resulting mortality projections. The simultaneous jumps in the grey and red surfaces for q_{30} and q_{60} are caused by the estimated cohort effect. This effect is not visible for q_{80} because for Dutch females aged 20-89 we found $x_c = 89$, which implies that the cohort effect hardly affects the highest ages. From Figure 4 we observe that the projections of $\kappa_t^{(1)}$ are more convincing if we allow for structural changes, and in Figure 5 the mortality projections with structural changes are more convincing as well. This is confirmed in Table 5 as both the RMSE and the $\overline{\text{DSS}}$ have improved substantially.

Similar results are shown in Figure 6 and 7 for model M9 applied to Dutch females aged 20-89. The projections for $\kappa_t^{(2)}$ are more plausible when structural changes are allowed, but the projections for $\kappa_t^{(3)}$ are still implausible. The last fitted cohort effect is the cohort 1935[10], and later cohort effects are projected using an appropriate ARIMA-process. The cohort effect needed for projections for $x = 30$ are therefore projected over 35 years into the future[11], while for $x = 60$ the cohort effect is projected only few years into the future and for $x = 80$ it is available from the model calibration. This explains the relatively large confidence interval for q_{30} in Figure 7. The projections for q_{80} including the structural change in $\kappa_t^{(2)}$ do not capture the realised mortality improvements, while the projections without structural changes do follow the realised mortality rates closely. Hence, even though the projected period effect

[10] For M9 and M13 the cohort effect is set equal to zero if there are no observations related to the age 60 or higher. For the age range 20-89 and the calibration period 1950-2000 this means that the last estimated cohort is $2000 - 65 = 1935$.

[11] The cohort effect needed in 2001 for $x = 30$ is for the cohort 1971. The last estimated cohort effect is for the cohort 1935. Hence, the cohort effect for the cohort 1971 is projected 36 years from the last estimated cohort effect.

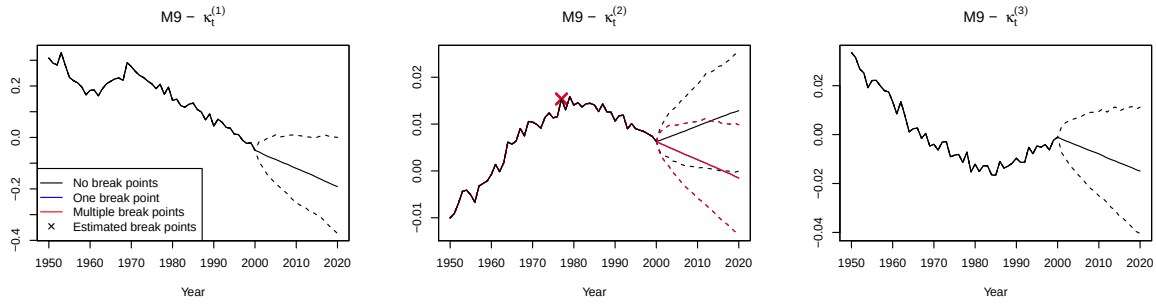

Figure 6: Projections for the period effects of M9 applied to Dutch females aged 20-89 in the period 1950-2000. The structural change for $\kappa_t^{(2)}$ is identified both if we allow for one and if we allow for multiple structural changes.

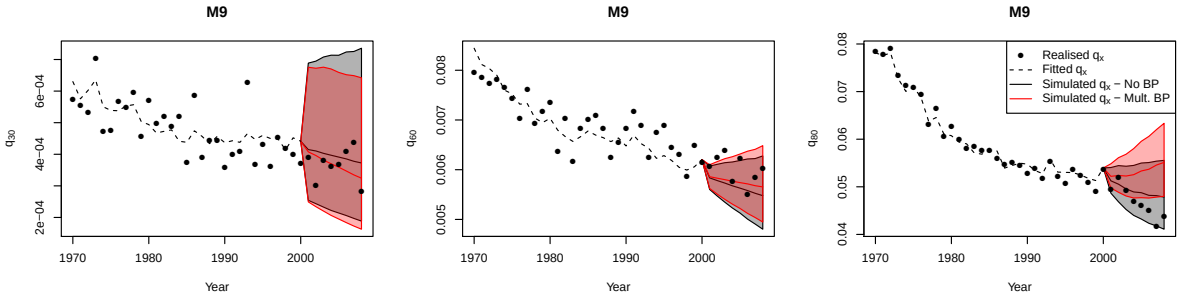

Figure 7: Mortality projections from M9 for $x = \{30, 60, 80\}$ calibrated on Dutch females aged 20-89 in the period 1950-2000. The black and red lines represent projections without and with multiple structural changes, respectively, at the 5th, 50th and 95th percentile.

is more plausible when structural changes are accounted for, the resulting mortality projections can be implausible for certain ages leading to worse backtesting results in Table 5.

The most interesting example is M7 applied to Dutch females aged 60-89. In Table 6 we see that both the RMSE and $\overline{\text{DSS}}$ worsen if we allow for a single structural change, but the statistics improve if we allow for multiple structural changes. Figure 8 shows the projections for the period effects while allowing for no, one or multiple structural changes. The projections for $\kappa_t^{(1)}$ with a single structural change are less convincing than when no structural changes are allowed, because the last structural change has not been identified. When we allow for multiple structural changes we are able to detect both structural changes, and the projections for the period effects are more convincing. The projections for $\kappa_t^{(2)}$ are also most convincing if we allow for multiple structural changes. This example clearly illustrates the potential added value from allowing for multiple structural changes.

5 Conclusion

In this paper we calibrate a selection of stochastic mortality models on historical mortality data from the Netherlands and Belgium. To create mortality projections, we project the period and the cohort effects. The cohort effects are projected using an ARIMA(p, d, q)-specification, where (p, d, q) are chosen such that the BIC is optimal. The period effect is projected using a modelling strategy that allows for objective detection of multiple structural changes in the difference stationary process.

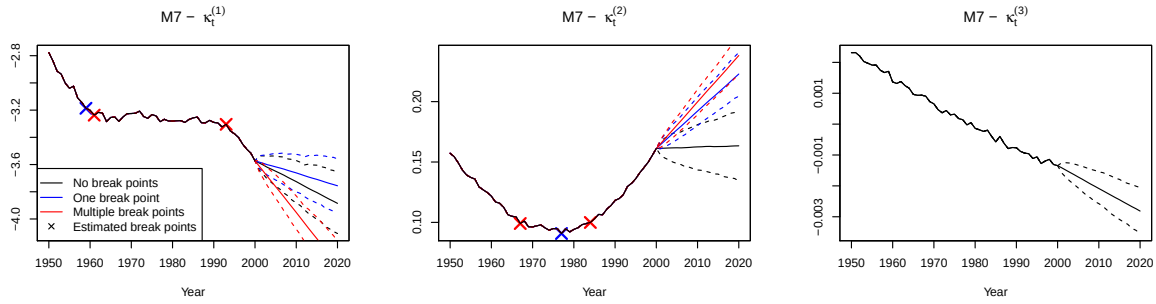

Figure 8: Projections for the period effects of M7 applied to Dutch females aged 60-89 in the period 1950-2000.

We compare the impact of not allowing for structural changes with allowing for a single or multiple structural changes. We observe that mortality projections are most robust with respect to the calibration period if we allow for multiple structural changes. Further, we find that allowing for structural changes can lead to improved backtesting results. Often one structural change, and sometimes even multiple structural changes are estimated. Allowing for structural changes does not always lead to improved backtesting results, because apparent structural changes are not identified. For these apparent structural changes the improvement in fit from including a structural change is not sufficient yet to to overcome the penalty in BIC caused by the extra parameter.

In this paper, the mortality model and the time series models are estimated seperately. Ideally, all sources of randomness should be addressed at once, which means that the Poisson likelihood and the likelihood of the time series should be optimised simultaneously. However, this raises severe computational challenges since the conveniently simple structure of the logarithmic likelihood can no longer be exploited in the same way as in the standard approach. This is therefore left as a subject for future research.

Acknowledgements

Frank van Berkum would like to thank Anja De Waegenaere, Bertrand Melenberg, and Steven Haberman and 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.

	The Netherlands, females 20-89						Belgium, females 20-89					
	RMSE			$\overline{\text{DSS}}$			RMSE			$\overline{\text{DSS}}$		
Model	0	1	>1	0	1	>1	0	1	>1	0	1	>1
M1	63.0	63.0	63.0	7.92	7.92	7.92	63.4	63.4	63.4	7.55	7.55	7.55
M1A	64.7	64.7	64.7	7.78	7.78	7.78	58.5	58.5	58.5	7.50	7.50	7.50
M2	85.8	85.8	85.8	8.77	8.77	8.77	2069.8	2069.8	2069.8	8.36	8.36	8.36
M2A	65.9	**64.7**	**64.7**	7.61	**7.67**	**7.67**	53.9	**53.8**	**53.8**	7.33	**7.42**	**7.42**
M3	129.5	129.5	129.5	9.04	9.04	9.04	101.1	101.1	101.1	8.08	8.08	8.08
M9	100.5	**141.0**	**141.0**	8.57	**9.17**	**9.17**	84.7	**74.2**	**74.2**	8.43	**8.59**	**8.59**
M10	102.2	102.2	102.2	8.75	8.75	8.75	94.8	94.8	94.8	9.71	9.71	9.71
M11	98.6	98.6	98.6	8.30	8.30	8.30	59.7	59.7	59.7	7.10	7.10	7.10
M12	100.5	**75.4**	**75.4**	8.93	**8.18**	**8.18**	96.8	96.8	96.8	7.48	7.48	7.48
M13	106.9	106.9	106.9	8.61	8.61	8.61	82.8	**83.1**	**83.1**	8.16	**8.28**	**8.28**

Table 5: Results for Dutch and Belgian female mortality rates for the ages 20-89 calibrated on the years 1950-2000. Mortality forecasts are backtested for the years 2001-2008 using different forecasting methods for the period effects. "0", "1" or ">1" means we allow for no, a single or multiple structural changes, respectively. Bold numbers indicate improved backtesting results compared with no structural changes; red numbers indicate worsened results compared with no structural changes.

	The Netherlands, females 60-89						Belgium, females 60-89					
	RMSE			$\overline{\text{DSS}}$			RMSE			$\overline{\text{DSS}}$		
Model	0	1	>1	0	1	>1	0	1	>1	0	1	>1
M1	88.3	88.3	88.3	9.77	9.77	9.77	79.6	79.6	79.6	9.51	9.51	9.51
M1A	95.3	95.3	95.3	9.73	9.73	9.73	84.3	84.3	84.3	9.63	9.63	9.63
M2	112.7	112.7	112.7	10.09	10.09	10.09	84.1	84.1	84.1	9.66	9.66	9.66
M2A	112.6	**90.7**	**90.7**	9.75	**9.74**	**9.74**	90.4	**65.5**	**65.5**	9.80	**9.71**	**9.71**
M3	183.1	183.1	183.1	11.99	11.99	11.99	144.4	144.4	144.4	10.66	10.66	10.66
M5	135.9	135.9	135.9	12.79	12.79	12.79	106.2	106.2	106.2	13.32	13.32	13.32
M6	278.5	**331.2**	**331.2**	14.26	**14.22**	**14.22**	149.7	149.7	149.7	10.68	10.68	10.68
M7	471.4	**590.1**	**327.2**	18.95	**21.38**	**15.35**	376.4	**434.8**	**385.3**	15.11	**14.93**	**13.46**
M8	198.9	**109.4**	**109.4**	10.67	**10.45**	**10.45**	182.3	182.3	182.3	10.97	10.97	10.97
M9	110.1	110.1	110.1	10.03	10.03	10.03	83.7	83.7	83.7	9.51	9.51	9.51
M10	97.6	97.6	97.6	9.88	9.88	9.88	83.7	83.7	83.7	9.53	9.53	9.53
M11	102.8	102.8	102.8	9.96	9.96	9.96	68.2	68.2	68.2	9.35	9.35	9.35
M12	122.2	122.2	122.2	10.36	10.36	10.36	87.1	87.1	87.1	9.56	9.56	9.56
M13	121.2	**171.2**	**171.2**	10.07	**11.35**	**11.35**	84.1	84.1	84.1	9.45	9.45	9.45

Table 6: Results for Dutch and Belgian female mortality rates for the ages 60-89 calibrated on the years 1950-2000, backtested on the years 2001-2008. Notes: see Table 5.

	The Netherlands, males 20-89						Belgium, males 20-89					
	RMSE			$\overline{\text{DSS}}$			RMSE			$\overline{\text{DSS}}$		
Model	0	1	>1	0	1	>1	0	1	>1	0	1	>1
M1	240.9	**220.9**	**220.9**	20.65	**20.30**	**20.30**	126.4	126.4	126.4	10.36	10.36	10.36
M1A	186.0	186.0	186.0	12.79	12.79	12.79	96.9	**86.6**	**86.6**	9.08	**9.02**	**9.02**
M2	84.8	84.8	84.8	9.35	9.35	9.35	65.5	65.5	65.5	8.71	8.71	8.71
M2A	131.7	131.7	131.7	10.39	10.39	10.39	68.3	**55.1**	**55.1**	8.35	**8.27**	**8.27**
M3	131.2	131.2	131.2	10.13	10.13	10.13	72.8	72.8	72.8	9.01	9.01	9.01
M9	148.1	**104.6**	**104.6**	9.52	**8.93**	**8.93**	68.8	68.8	68.8	8.60	8.60	8.60
M10	165.7	**132.4**	**132.4**	10.34	**9.67**	**9.67**	100.7	**78.8**	**78.8**	9.25	**9.50**	**9.50**
M11	154.0	154.0	154.0	9.89	9.89	9.89	71.8	71.8	71.8	8.22	8.22	8.22
M12	150.7	**105.2**	**105.2**	12.43	**11.21**	**11.21**	50.0	50.0	50.0	8.64	8.64	8.64
M13	139.9	**100.4**	**100.4**	9.43	**8.94**	**8.94**	64.3	64.3	64.3	8.68	8.68	8.68

Table 7: Results for Dutch and Belgian male mortality rates for the ages 20-89 calibrated on the years 1950-2000, backtested on the years 2001-2008. Notes: see Table 5.

	The Netherlands, males 60-89						Belgium, males 60-89					
	RMSE			$\overline{\text{DSS}}$			RMSE			$\overline{\text{DSS}}$		
Model	0	1	>1	0	1	>1	0	1	>1	0	1	>1
M1	265.2	265.2	265.2	15.50	15.50	15.50	138.5	138.5	138.5	11.15	11.15	11.15
M1A	248.0	248.0	248.0	14.33	14.33	14.33	140.3	140.3	140.3	10.81	10.81	10.81
M2	121.2	121.2	121.2	10.80	10.80	10.80	91.4	91.4	91.4	10.38	10.38	10.38
M2A	121.3	121.3	121.3	10.62	10.62	10.62	70.2	**119.9**	**119.9**	9.89	**10.44**	**10.44**
M3	172.3	172.3	172.3	11.67	11.67	11.67	88.2	88.2	88.2	10.22	10.22	10.22
M5	239.4	239.4	239.4	13.16	13.16	13.16	133.0	133.0	133.0	10.40	10.40	10.40
M6	198.1	198.1	198.1	13.13	13.13	13.13	129.1	129.1	129.1	10.36	10.36	10.36
M7	141.8	141.8	141.8	10.85	10.85	10.85	111.8	111.8	111.8	10.09	10.09	10.09
M8	328.1	**242.5**	**242.5**	14.53	**13.45**	**13.45**	170.0	170.0	170.0	10.95	10.95	10.95
M9	175.6	175.6	175.6	11.58	11.58	11.58	120.9	120.9	120.9	10.20	10.20	10.20
M10	238.7	238.7	238.7	13.00	13.00	13.00	130.2	130.2	130.2	10.34	10.34	10.34
M11	243.7	243.7	243.7	12.89	12.89	12.89	144.2	144.2	144.2	10.75	10.75	10.75
M12	283.6	**186.4**	**186.4**	13.31	**11.89**	**11.89**	123.2	123.2	123.2	10.25	10.25	10.25
M13	199.1	199.1	199.1	12.19	12.19	12.19	166.9	166.9	166.9	11.29	11.29	11.29

Table 8: Results for Dutch and Belgian male mortality rates for the ages 60-89 calibrated on the years 1950-2000, backtested on the years 2001-2008. Notes: see Table 5.

References

Andrews, D. (1992), 'Tests for parameter instability and structural change with unknown change point', *Econometrica* **61**(4), 821–856.

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

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

Barrieu, P., Bensusan, H., Karoui, N. E., Hillairet, C., Loisel, S., Ravanelli, C. and Salhi, Y. (2012), 'Understanding, modelling and managing longevity risk: key issues and main challenges', *Scandinavian Actuarial Journal* **3**, 203 – 231.

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

Bots, M. and Grobbee, D. (1996), '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.

Brouhns, N., Denuit, M. and Vermunt, J. (2002), 'A poisson log-bilinear regression approach to the construction of projected lifetables', *Insurance: Mathematics and Economics* **31**(3), 373 – 393.

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

Cairns, A., Blake, D. and Dowd, K. (2008), 'Modelling and management of mortality risk: a review', *Scandinavian Actuarial Journal* **2008**(2-3), 79–113.

Cairns, A., Blake, D., Dowd, K., Coughlan, G., Epstein, D. and Khalaf-Allah, M. (2011), 'Mortality density forecasts: An analysis of six stochastic mortality models', *Insurance: Mathematics and Economics* **48**(3), 355 – 367.

Cairns, A., Blake, D., Dowd, K., Coughlan, G., Epstein, D., Ong, A. and Balevich, I. (2009), '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.

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

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

Coelho, E. and Nunes, L. (2013), Cohort effects and structural changes in the mortality trend, Working paper.
URL: *http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.11/2013/WP_5.1.pdf*

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

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

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

Gneiting, T. and Raftery, A. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359 – 378.

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

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

Harris, D., Harvey, D., Leybourne, S. and Taylor, A. (2009), 'Testing for a unit-root in the presence of a possible break in trend', *Econometric Theory* **25**, 1545 – 1588.

Harvey, D., Leybourne, S. and Taylor, A. (2009), 'Simple, robust and powerful tests of changing trend hypothesis', *Econometric Theory* **25**, 995 – 1029.

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

Li, H., Waegenaere, A. D. and Melenberg, B. (2013), The choice of sample size for mortality forecasting: a Bayesian learning approach, Working paper, Tilburg University.

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

Lindbergson, M. (2001), 'Mortality among the elderly in Sweden 1988−1997', *Scandinavian Actuarial Journal* (3), 79 – 94.

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

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

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

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

O'Hare, C. and Li, Y. (2012), Identifying structural breaks in stochastic mortality models, Working paper.
URL: *http://ssrn.com/abstract=2192208*

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

Renshaw, A. and Haberman, S. (2003), 'Lee-Carter mortality forecasting with age-specific enhancement', *Insurance: Mathematics and Economics* **33**(2), 255 – 272.

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

Riebler, A., Held, L. and Rue, H. (2012), 'Estimation and extrapolation of time trends registry data - borrowing strength from related populations', *The Annals of Applied Statistics* **6**(1), 304 – 333.

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

Vaupel, J. (1990), 'Relatives' risks: frailty models of life history data', *Theoretical population biology* **37**(1), 220 – 234.

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

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

A Parameter constraints

Some of the mortality models experience identifiability issues. Therefore, we impose parameter constraints. Table 9 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	$\sum_t \kappa_t^{(1)} = 0$	$\sum_t \kappa_t^{(2)} = 0$	$\sum_t \kappa_t^{(3)} = 0$		
M11	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$	$\sum_c c^2\gamma_c = 0$	$\sum_t \kappa_t^{(3)} = 0$	
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 9: 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