The impact of multiple structural changes on mortality predictions

van Berkum, F.; Antonio, K.; Vellekoop, M.

DOI
10.1080/03461238.2014.987807
Publication date
2016
Document Version
Final published version
Published in
Scandinavian Actuarial Journal

Scandinavian Actuarial Journal, 2016
Vol. 2016, No. 7, 581–603, http://dx.doi.org/10.1080/03461238.2014.987807

The impact of multiple structural changes on mortality predictions

FRANK VAN BERKUM[a]*, KATRIEN ANTONIO[a,b] and MICHEL VELLEKOOP[a]

[a]Faculty of Economics and Business, University of Amsterdam, Amsterdam, The Netherlands
[b]Faculty of Economics and Business, KU Leuven, Leuven, Belgium

(*Accepted November 2014*)

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 therefore analyse the impact of allowing for multiple structural changes on a large collection of mortality models. We find that this may lead to more robust projections for the period effect but that there is only a limited effect on the ranking of the models based on backtesting criteria, since there is often not yet sufficient statistical evidence for structural changes. However, there are cases for which we do find improvements in estimates and we therefore conclude that one should not exclude on beforehand that structural changes may have occurred.

Keywords: stochastic mortality; structural changes; mortality forecasting; backtesting

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 & 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 (Currie 2006, Renshaw & Haberman 2006), functional forms for the age effects to limit the number of parameters (Cairns *et al*. 2006, 2009) and the introduction of age-group specific and quadratic effects (Plat 2009, O'Hare & 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,

*Corresponding author. E-mail: f.vanberkum@uva.nl

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 & Goderniaux (2005) use a frequentist 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 & Nunes (2011) introduce structural changes in trend and difference stationary processes.

In this paper we extend the approach of Coelho & 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 & Nunes (2011) by allowing for multiple structural changes in the period effects, we determine the structural changes in an objective manner (Bai & Perron 1998). The optimal number of structural changes is selected using the Bayesian Information Criterion. To evaluate the performance of this approach, we compare the projections using 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 & 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. 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.

2.1. Mortality model structures

Let the expected number of deaths during calendar year t aged x at death be $\hat{d}_{t,x}$, and the average population aged x during calendar year t (exposure) be $e_{t,x}$. The death rate, $m_{t,x}$, is defined by

$$m_{t,x} = \frac{\hat{d}_{t,x}}{e_{t,x}}. \tag{1}$$

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 $[t, t+1) \times [x, x+1)$, then the maximum likelihood estimate $\hat{\mu}_{t,x}$ of the force of mortality $\mu_{t,x}$ is given by (see Pitacco *et al.* (2009)):

$$\hat{\mu}_{t,x} = \frac{d_{t,x}}{e_{t,x}} = m^{\text{obs}}_{t,x}, \tag{2}$$

with $m^{\text{obs}}_{t,x}$ the observed death rate. Further, given the previous assumption, the mortality rate is linked to the force of mortality through the relationship:

$$q_{t,x} = 1 - e^{-\mu_{t,x}}. \tag{3}$$

We will estimate the force of mortality based on the observed death rates using age effects ($\beta^{(i)}_x$), period effects ($\kappa^{(i)}_t$) 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 in Table 1.

As in Brouhns *et al.* (2002), we assume a Poisson distribution for the number of deaths within a year, $D_{t,x} \sim \text{Poisson}(e_{t,x}\mu_{t,x})$. The various specifications for $\mu_{t,x}$ 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 the 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 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 1 gives an overview.

The models M5–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 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 and approaches used for forecasting mortality.

2.2.1. *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

[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.

Table 1. Model specifications used in this paper.

Model	Name	Formula	Original paper
M1	LC	$\ln \mu_{t,x} = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}$	Lee & Carter (1992)
M1A	LC2	$\ln \mu_{t,x} = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\kappa_t^{(3)}$	Renshaw & Haberman (2003)
M2	M	$\ln \mu_{t,x} = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x}$	Renshaw & Haberman (2006)
M2A	–	$\ln \mu_{t,x} = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\kappa_t^{(3)} + \gamma_{t-x}$	
M3	APC	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(2)} + \gamma_{t-x}$	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*	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + (\bar{x}-x)^+\kappa_t^{(3)} + \gamma_{t-x}$	Plat (2009)
M10	M5*	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + (\bar{x}-x)^+\kappa_t^{(3)}$	Haberman & Renshaw (2011)
M11	M7*	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + (\bar{x}-x)^+\kappa_t^{(3)} + b(x)\kappa_t^{(4)} + \gamma_{t-x}$	Haberman & Renshaw (2011)
M12	M8*	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + (\bar{x}-x)^+\kappa_t^{(3)} + (x_c - x)\gamma_{t-x}$	Haberman & Renshaw (2011)
M13	Expl.YM	$\ln \mu_{t,x} = \beta_x^{(1)} + \kappa_t^{(1)} + (\bar{x}-x)\kappa_t^{(2)} + c(x)\kappa_t^{(3)} + \gamma_{t-x}$	O'Hare & Li (2011)

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 and Wales (1961–2005) and the Netherlands (1951–2005). 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 & 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 & Renshaw (2011) argue that the extrapolation of the cohort effect 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.

[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).

Lovász (2011) considers several models for Finnish (1950–2009) and Swedish (1950–2008) data. He models the period effects as in Dowd *et al*. (2010) and Haberman & Renshaw (2011) by assuming a multivariate random walk with drift. For the cohort effects, he chooses the ARIMA(p, d, q) process 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 data-sets 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 & 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 & Goderniaux (2005). Furthermore, factors like medical advances (Bots & Grobbee 1996) and health system reforms (Moreno-Serra & Wagstaff 2010) 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, and several authors proposed different methods on how to deal with the sensitivity of the calibration period.

2.2.2. *Optimal calibration period*

Booth *et al*. (2002) note that a random walk with constant drift may not be appropriate over the whole period of available mortality data. For the Lee–Carter model, they propose to restrict the calibration period. The last year is determined by the most recent data available, and the first year is chosen by optimising the fit of the random walk with drift model relative to the fit of the Lee–Carter model. They note that age effects may change through time and that by optimising the calibration period, the age effects are chosen more appropriately for the purpose of projecting mortality rates.

Denuit & Goderniaux (2005) approximate the period effect κ_t in the Lee–Carter model by a straight line using OLS and choose the calibration period when the corresponding adjusted R^2 is optimal. Li *et al*. (2013) include the length of the calibration period in their parameter space in a Bayesian framework. Mortality rates are projected for three mortality models using different calibration periods where projections resulting from different calibration periods are weighted by their posterior distribution.

2.2.3. *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 reveals 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).

2.2.4. *Structural changes in trend stationary models*

Li *et al.* (2011) calibrate the Lee–Carter model on England and 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 that become wider over time.

Sweeting (2011) calibrates the original CBD-model (M5) on England and 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).

2.2.5. *Structural change in difference stationary models*

Coelho & 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.

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

O'Hare & Li (2014) 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, health care systems, etc. For example, in the Netherlands, changes in smoking habits have been an important driver of changes in mortality, which resulted in increasing (1950–1970) and decreasing (from 1970 onwards) mortality rates (Janssen *et al*. 2007). 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.

Optimisation of the calibration period as in Booth *et al*. (2002) and Denuit & Goderniaux (2005) has appealing characteristics. Since older data points are excluded, the age effects are based on more recent data and are therefore more appropriate for forecasting than when all data are included. However, this approach may lead to short calibration periods which may result in more volatile parameter estimates and projections. Further, by excluding data, the researcher implicitly chooses not to explain part of the available data. Finally, these methods have been applied to the Lee–Carter model, but they are not easily transferable to multi-factor models, since different factors may suggest different calibration periods.

Therefore, 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 & 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 & Nunes (2011) and the work of O'Hare & Li (2014) such that multiple structural changes can be detected, as multiple events in the past may have affected the speed of mortality improvements.

We assume a multivariate random walk with drift for the period effects. Each univariate series may experience multiple structural changes during the calibration period. We determine the optimal number of structural changes separately for each time series using an optimisation criterion. The period effects are then simulated using the latest drift parameters and the estimated covariance structure.

To determine the number of structural changes and their corresponding dates, we follow the methodology introduced in Bai & Perron (2003). Suppose we have at our disposal a period effect $\kappa_t^{(i)}$ $(t = 1, \ldots, T)$ and define the first-order differences by $\Delta\kappa_t^{(i)} = \kappa_t^{(i)} - \kappa_{t-1}^{(i)}$ for $t = 2, \ldots, T$. We estimate a random walk with a piecewise constant drift:

$$\Delta\kappa_t^{(i)} = \begin{cases} \beta_1 + \varepsilon_t, & t \le t_1 \\ \dots & \\ \beta_j + \varepsilon_t & t_{j-1} < t \le t_j \\ \dots & \\ \beta_{m+1} + \varepsilon_t, & t_m < t \end{cases} \tag{4}$$

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):

$$\mathrm{SSR}(t_1, \dots, t_m) = \sum_{j=1}^{m+1} \sum_{t=t_{j-1}+1}^{t_j} [\Delta\kappa_t^{(i)} - \beta_j]^2 \tag{5}$$

where $t_0 = 1$ and $t_{m+1} = T$. In the model specification above, we distinguish m break points that divide the time series into $m + 1$ periods with different drifts. Both the number of break points and the dates 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 (5), then the estimated break points $(\hat{t}_1, \dots, \hat{t}_m)$ are such that $(\hat{t}_1, \dots, \hat{t}_m) = \arg\min_{t_1,\dots,t_m} \mathrm{SSR}(t_1, \dots, t_m)$, where the minimisation is taken over all partitions $(t_1, \dots, t_m)$ for which $t_j - t_{j-1} \ge h$. The parameter h corresponds to the minimum period that a regime should last, and is to be chosen up front. Bai & 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 $m^* = \arg\max \mathrm{BIC}(m)$.

As in Bai & Perron (1998) and Bai & Perron (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:

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

$$F = \frac{(\text{RSSR} - \text{USSR})/(p_1 - p_0)}{\text{USSR}/(n - p_1)}, \tag{6}$$

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 the structural changes are unknown, we cannot use the standard critical values of the F-distribution 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 (6) 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:

$$\text{UDmax} = \max_k F(k) \tag{7}$$

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 & 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 applied to Dutch male mortality.

3.1.1. Illustration – the Lee–Carter model

We consider the period effect of the Lee–Carter model, estimated on Dutch male mortality data for the period 1960–2008, for the ages 60–89. We illustrate our method, but also show results of the optimal calibration period strategy in Denuit & Goderniaux (2005). 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 are not well connected with the observations. The blue lines correspond to projections when one structural change is allowed; the break point is dated at 1993. 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

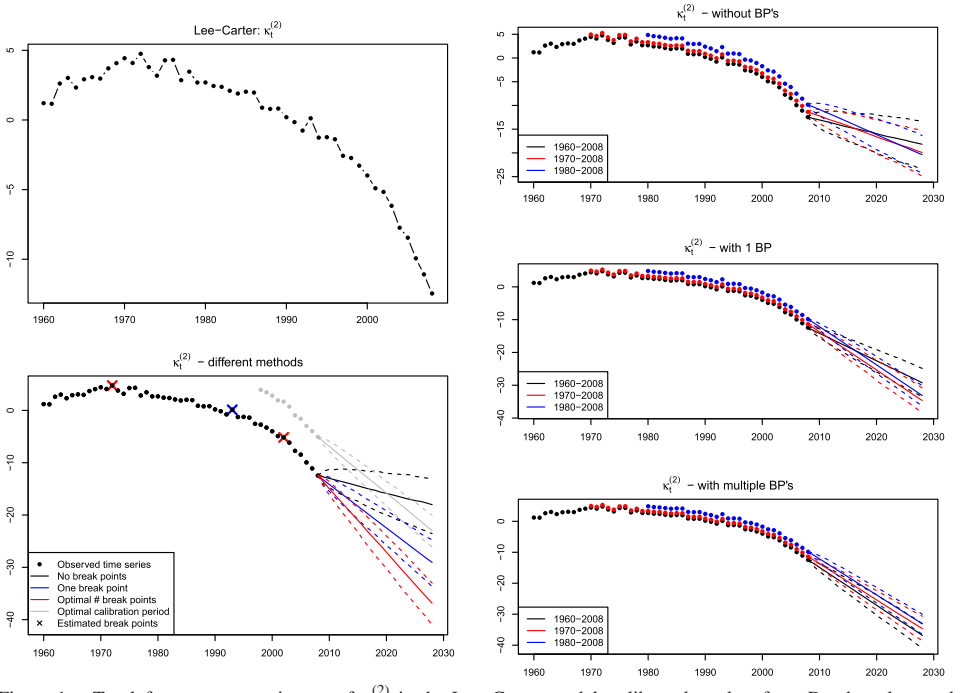

Figure 1. Top left: parameter estimates of $\kappa_t^{(2)}$ in the Lee–Carter model, calibrated on data from Dutch males aged 60–89 in the period 1960–2008. Bottom left: projections for the period effect using different projection methods. Top right through bottom right: projections of the period effect 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 dashed lines are the 5th and 95th percentiles of the projections. (Coloured versions of the figures can be found online.)

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 effects from the Lee–Carter model calibrated on different 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 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.

[5]See Bai & Perron 1998 for a description how these confidence intervals are derived. We used the R-package `strucchange` (Zeileis *et al*. 2002) to detect structural changes.

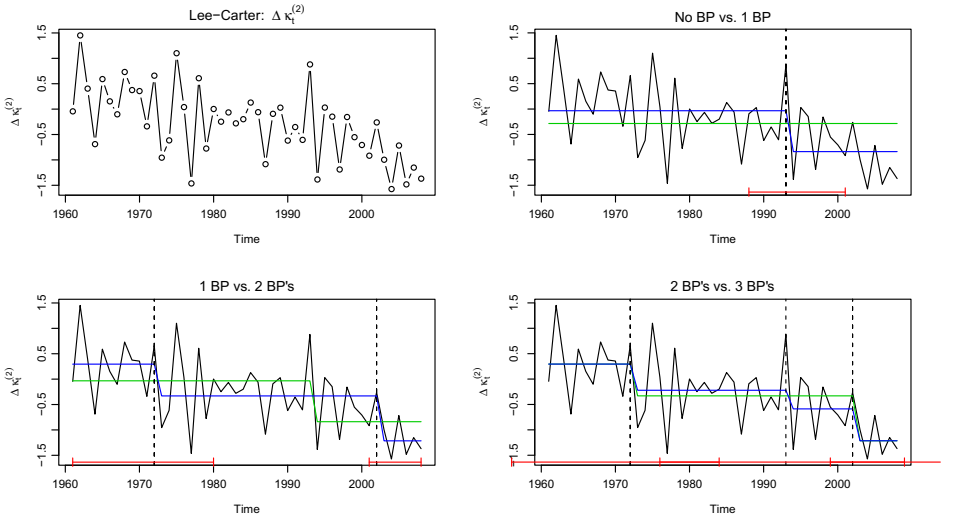

Figure 2. Confidence intervals for estimated break points for $\kappa_t^{(2)}$ in the Lee–Carter model, calibrated on Dutch males aged 60–89 in the years 1960–2008. In the plots (i) BP's vs. $(i+1)$ BP's the green lines represent the mean of $\Delta\kappa_t^{(2)}$ for the different periods when (i) BP's 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. (Coloured versions of the figures can be found online.)

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.

Figure 1 (bottom left) also shows the estimated period effect if the calibration period is chosen according to the procedure proposed in Denuit & Goderniaux (2005). We calibrate the Lee–Carter model to the entire calibration period, and then estimate OLS on different subsets of the period effect while keeping the end date fixed. The optimal calibration period is chosen where the adjusted R^2 is maximal. Using that calibration period, we recalibrate the Lee–Carter model, and the result is plotted here in grey[6]. In line with Denuit & Goderniaux (2005), we enforce that the calibration period must be larger than ten years, and in this example the optimised calibration period turns out to be of minimal length, in contrast to the findings of Denuit & Goderniaux (2005) for Belgian data.

In this recent calibration period, the period effect shows little variability which is translated into narrower confidence intervals than when we would have required the model to explain the entire data-set. Figure 3 shows the parameter estimates of the Lee–Carter model based on the entire and the optimal calibration period. Given the parameter restrictions, $\beta_x^{(1)}$ is the mean mortality rate, which explains the downward shift. The estimates for $\beta_x^{(2)}$ differ substantially and those for the optimal (and shorter) calibration period are more volatile.

[6]The estimated period effect on the optimal calibration period is shifted upwards due to the parameter restrictions.

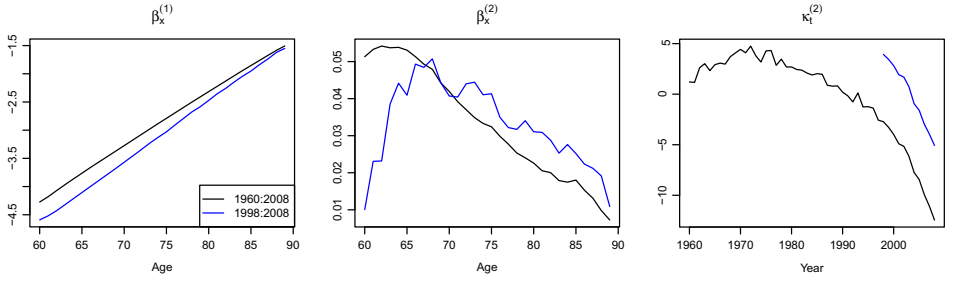

Figure 3. Parameter estimates of the Lee–Carter model, calibrated on data from Dutch males aged 60–89 using the large calibration period 1960–2008, and the optimal calibration period 1998–2008 according to the method of Denuit & Goderniaux (2005). (Coloured versions of the figures can be found online.)

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 to select the optimal specification, 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).

4.1. *Model fit*

We calibrate the models on male mortality data[7] from the Netherlands and Belgium for the years 1950–2008. Earlier data are 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 & Goderniaux (2005).

We present the estimation results for Dutch and Belgian males[8] for ages 20–89 in Table 2 and for ages 60–89 in Table 3. These tables contain the effective[9] number of parameters that

[7]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.
[8]Similar results for Dutch and Belgian females are available upon request from the authors.
[9]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.

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	Parameters	AIC	BIC	AIC	BIC
M1	197	−22, 000 (10)	−22, 623 (10)	−22, 332 (10)	−22, 955 (10)
M1A	324	−19, 535 (8)	−20, 559 (8)	−20, 122 (8)	−21, 146 (8)
M2	385	−18, 425 (5)	−19, 642 (5)	−19, 129 (6)	−20, 345 (6)
M2A	513	−18, 438 (6)	−20, 060 (7)	−18, 994 (5)	−20, 616 (7)
M3	246	−18, 947 (7)	−19, 724 (6)	−19, 538 (7)	−20, 315 (5)
M9	327	−18, 359 (4)	−19, 392 (2)	−18, 885 (4)	−19, 919 (2)
M10	244	−19, 905 (9)	−20, 676 (9)	−21, 419 (9)	−22, 190 (9)
M11	422	−18, 258 (1)	−19, 591 (4)	−18, 810 (1)	−20, 144 (4)
M12	364	−18, 289 (2)	−19, 439 (3)	−18, 840 (2)	−19, 990 (3)
M13	327	−18, 358 (3)	−19, 392 (1)	−18, 873 (3)	−19, 907 (1)

Note: The numbers in brackets represent the ranking of the models for a specific dataset.

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

		The Netherlands		Belgium	
Model	Parameters	AIC	BIC	AIC	BIC
M1	117	−11, 035 (14)	−11, 355 (14)	−10, 421 (14)	−10, 741 (14)
M1A	204	−9, 204 (12)	−9, 762 (13)	−9, 665 (12)	−10, 223 (12)
M2	225	−8, 797 (8)	−9, 412 (4)	−8, 991 (6)	−9, 606 (4)
M2A	313	−8, 820 (9)	−9, 675 (11)	−8, 995 (7)	−9, 850 (9)
M3	166	−8, 941 (11)	−9, 395 (3)	−9, 101 (10)	−9, 555 (2)
M5	118	−9, 345 (13)	−9, 668 (10)	−9, 912 (13)	−10, 235 (13)
M6	196	−8, 732 (1)	−9, 268 (1)	−8, 935 (1)	−9, 471 (1)
M7	254	−8, 735 (2)	−9, 429 (5)	−8, 938 (2)	−9, 632 (5)
M8	198	−8, 792 (7)	−9, 333 (2)	−9, 031 (9)	−9, 572 (3)
M9	284	−8, 752 (4)	−9, 528 (8)	−8, 942 (4)	−9, 719 (7)
M10	204	−8, 908 (10)	−9, 465 (6)	−9, 347 (11)	−9, 905 (11)
M11	342	−8, 771 (5)	−9, 706 (12)	−8, 965 (5)	−9, 900 (10)
M12	284	−8, 783 (6)	−9, 560 (9)	−9, 002 (8)	−9, 778 (8)
M13	284	−8, 748 (3)	−9, 524 (7)	−8, 939 (3)	−9, 716 (6)

Note: The numbers in brackets represent the ranking of the models for a specific dataset.

is estimated in each of the models, and the corresponding AIC and BIC that we define as $\text{AIC} = \log L - k$ and $\text{BIC} = \log L - \frac{1}{2}k \cdot \log n$, where $\log L$ is the loglikelihood, n is the number of observations and k is the effective number of parameters. A higher AIC or BIC means that the model is better able to explain the data. The difference between the AIC and the BIC is that the BIC imposes a higher penalty for the number of parameters used. Mortality models contain many parameters and we therefore believe the BIC to be a more appropriate information criterion. However, the ranking based on fit on historical data does not predict whether a model will produce good mortality projections.

For the age range 20–89, the models with a cohort effect and interaction between age and period effects have the highest AIC and BIC. As expected, models that score well on AIC but which have many parameters, score worse on BIC; M11 is the clearest example of this. 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 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.

For the models M8 and M12, the impact of the cohort effect on the mortality rates for age x depends on the parameter x_c. The cohort effect γ_{t-x} is multiplied with $(x_c - x)$, so it has a larger impact on mortality rates for ages farther away from x_c. From Table 4 we conclude that, for the data-sets considered, four out of 12 times the cohort effect mainly affects younger ages ($x_c = 89$), and eight 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 4 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 forces of mortality for the years 2001–2008. This leads to a predictive distribution for the forces of mortality $\Pi_{t,x}$ for $x = x_1, \ldots, x_n$ and $t = T + 1, \ldots, T + s$. As in Riebler *et al.* (2012), we obtain the mean $\mathbb{E}(\Pi_{t,x})$ and variance $\mathrm{Var}(\Pi_{t,x})$ of future forces of mortality from the simulated predictive distribution. With $D_{t,x} \sim \mathrm{Poisson}(e_{t,x}\Pi_{t,x})$ and using the law of total expectation it follows that for $t > T$ the expected death counts are

$$\widehat{d}_{t,x} = \mathbb{E}(D_{t,x}) = e_{t,x}\mathbb{E}(\Pi_{t,x}) \tag{8}$$

and the variance of the death counts is

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

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.

		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

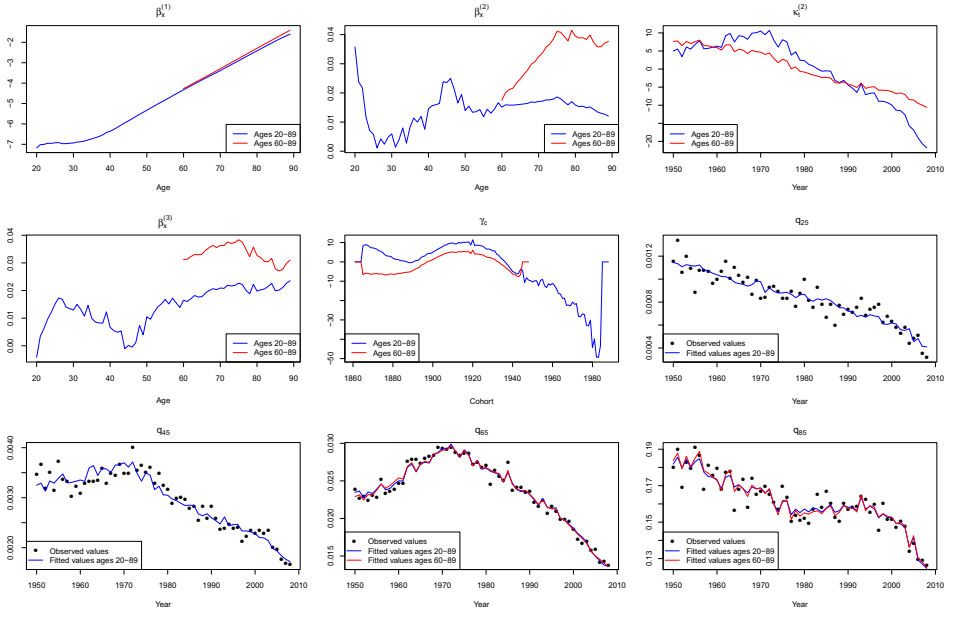

Figure 4. The first five panels show the parameter estimates for M2 calibrated on Dutch male mortality in the years 1950–2008 on the ages 20–89 and 60–89. The last four panels show realised mortality rates (dots) and fitted mortality rates for $x = \{25, 45, 65, 85\}$ (calibrated on ages 20–89 and ages 60–89). (Coloured versions of the figures can be found online.)

since we assume the population size $e_{t,x}$ given[10]. In the evaluation of the out-of-sample performance Gneiting & Raftery (2007) 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 and without the possibility of structural changes:

$$\text{RMSE} = \sqrt{\frac{1}{n \cdot s} \sum_{t,x} \left(d_{t,x} - \widehat{d}_{t,x}\right)^2}. \tag{9}$$

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 under consideration differ in scale for different ages and years due to different forces of mortality and exposures. The Dawid-Sebastiani scoring rule (DSS) introduced by Gneiting & 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 all forecasted death counts into a single statistic:

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

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.

	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	67.7	67.7	67.7	7.79	7.79	7.79	56.1	56.1	56.1	7.36	7.36	7.36
M1A	75.9	75.9	75.9	7.76	7.76	7.76	57.7	57.7	57.7	7.46	7.46	7.46
M2	71.7	71.7	71.7	8.73	8.73	8.73	–	–	–	–	–	–
M2A	79.7	*82.1*	*82.1*	7.62	*7.69*	*7.69*	39.0	**36.8**	**36.8**	7.24	*7.31*	*7.31*
M3	118.1	118.1	118.1	8.69	8.69	8.69	82.2	82.2	82.2	7.72	7.72	7.72
M9	81.9	*166.9*	*166.9*	8.50	*9.34*	*9.34*	61.4	*81.0*	*81.0*	8.31	*8.62*	*8.62*
M10	92.5	92.5	92.5	8.71	8.71	8.71	86.9	86.9	86.9	9.62	9.62	9.62
M11	64.6	64.6	64.6	8.19	8.19	8.19	38.0	38.0	38.0	7.05	7.05	7.05
M12	121.9	**76.2**	**76.2**	9.06	**8.21**	**8.21**	72.9	72.9	72.9	7.32	7.32	7.32
M13	91.7	91.7	91.7	8.54	8.54	8.54	60.8	*61.2*	*61.2*	8.04	*8.16*	*8.16*

Notes: Bold numbers indicate improved backtesting results compared with no structural changes; italic numbers indicate worsened results compared with no structural changes.

$$\overline{\text{DSS}} = \frac{1}{n \cdot s} \sum_{t,x} \left[\left(\frac{d_{t,x} - \widehat{d}_{t,x}}{\sigma_{t,x}} \right)^2 + \log \sigma_{t,x}^2 \right]. \tag{10}$$

Tables 5 and 6 show the backtesting results for Dutch and Belgian females for the ages 20–89 and 60–89, respectively[11], and Tables 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 italic 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 on historical data in Tables 2 and 3.

Figure 5 shows projections of the period effects for M12 applied to Dutch females aged 20–89 and Figure 6 shows resulting mortality projections. The non-monotone behaviour observed in the red and grey projections is due to 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 5, we observe that the projections of $\kappa_t^{(1)}$ are more convincing if we allow for structural changes, and in Figure 6 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 Figures 7 and 8 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[12],

[11] In Table 5, the results for M2 applied to Belgian females are implausible due to unrealistic cohort projections and are therefore not included in the table. Using a different time series model for the cohort effect might lead to better results.

[12] 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 $2000 - 65 = 1935$.

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.

	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	128.2	128.2	128.2	10.38	10.38	10.38	78.2	78.2	78.2	9.36	9.36	9.36
M1A	112.8	112.8	112.8	9.71	9.71	9.71	87.8	87.8	87.8	9.58	9.58	9.58
M2	102.5	102.5	102.5	10.03	10.03	10.03	61.8	61.8	61.8	9.49	9.49	9.49
M2A	201.8	**124.7**	**124.7**	10.06	**9.83**	**9.83**	154.7	**84.4**	**84.4**	9.95	**9.68**	**9.68**
M3	160.7	160.7	160.7	11.32	11.32	11.32	111.4	111.4	111.4	9.97	9.97	9.97
M5	134.1	134.1	134.1	12.75	12.75	12.75	101.9	101.9	101.9	13.28	13.28	13.28
M6	339.5	*412.0*	*412.0*	15.22	*15.36*	*15.36*	177.8	177.8	177.8	10.92	10.92	10.92
M7	517.3	*719.3*	**421.9**	19.88	*23.39*	**16.77**	399.8	*500.7*	*470.5*	15.33	*15.54*	**14.17**
M8	141.2	**88.6**	**88.6**	10.03	*10.38*	*10.38*	149.1	149.1	149.1	10.49	10.49	10.49
M9	114.4	114.4	114.4	10.08	10.08	10.08	86.6	86.6	86.6	9.55	9.55	9.55
M10	113.1	113.1	113.1	9.93	9.93	9.93	86.0	86.0	86.0	9.56	9.56	9.56
M11	137.0	137.0	137.0	10.06	10.06	10.06	83.3	83.3	83.3	9.40	9.40	9.40
M12	151.5	151.5	151.5	10.51	10.51	10.51	98.5	98.5	98.5	9.66	9.66	9.66
M13	135.5	*218.6*	*218.6*	10.16	*11.66*	*11.66*	87.8	87.8	87.8	9.40	9.40	9.40

Note: see Table 5.

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.

	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	266.4	**243.4**	**243.4**	22.01	**21.43**	**21.43**	147.2	147.2	147.2	10.56	10.56	10.56
M1A	222.9	222.9	222.9	14.35	14.35	14.35	124.1	**113.3**	**113.3**	9.52	**9.42**	**9.42**
M2	105.2	105.2	105.2	9.41	9.41	9.41	84.7	84.7	84.7	8.79	8.79	8.79
M2A	164.9	164.9	164.9	10.76	10.76	10.76	87.3	**61.7**	**61.7**	8.46	**8.25**	**8.25**
M3	145.4	145.4	145.4	10.06	10.06	10.06	71.8	71.8	71.8	8.77	8.77	8.77
M9	176.7	**120.3**	**120.3**	9.71	**8.99**	**8.99**	79.7	79.7	79.7	8.61	8.61	8.61
M10	193.7	**159.4**	**159.4**	10.55	**9.87**	**9.87**	117.1	**83.4**	**83.4**	9.32	*9.45*	*9.45*
M11	187.2	187.2	187.2	10.25	10.25	10.25	93.3	93.3	93.3	8.38	8.38	8.38
M12	178.1	**118.2**	**118.2**	12.69	**11.31**	**11.31**	45.3	45.3	45.3	8.58	8.58	8.58
M13	164.4	**111.8**	**111.8**	9.59	**8.96**	**8.96**	72.3	72.3	72.3	8.67	8.67	8.67

Note: see Table 5.

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[13], 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 8. The projections for q_{80} including the structural change in $\kappa_t^{(2)}$ do not follow the realised mortality improvements, while the projections without structural changes do follow

[13]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.

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.

	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	296.9	296.9	296.9	16.30	16.30	16.30	160.6	160.6	160.6	10.99	10.99	10.99
M1A	297.0	297.0	297.0	15.53	15.53	15.53	173.3	173.3	173.3	11.19	11.19	11.19
M2	120.2	120.2	120.2	10.58	10.58	10.58	77.3	77.3	77.3	10.11	10.11	10.11
M2A	166.9	166.9	166.9	10.97	10.97	10.97	112.1	**80.5**	**80.5**	10.17	**10.15**	**10.15**
M3	200.2	200.2	200.2	11.63	11.63	11.63	91.7	91.7	91.7	9.81	9.81	9.81
M5	286.5	286.5	286.5	13.86	13.86	13.86	166.6	166.6	166.6	10.68	10.68	10.68
M6	232.3	232.3	232.3	13.59	13.59	13.59	163.1	163.1	163.1	10.63	10.63	10.63
M7	202.4	202.4	202.4	12.53	12.53	12.53	132.3	132.3	132.3	10.23	10.23	10.23
M8	386.6	**284.7**	**284.7**	15.47	**14.14**	**14.14**	209.4	209.4	209.4	11.35	11.35	11.35
M9	207.9	207.9	207.9	12.00	12.00	12.00	148.3	148.3	148.3	10.41	10.41	10.41
M10	283.4	283.4	283.4	13.61	13.61	13.61	161.5	161.5	161.5	10.62	10.62	10.62
M11	283.2	283.2	283.2	13.42	13.42	13.42	174.5	174.5	174.5	10.99	10.99	10.99
M12	343.7	**227.5**	**227.5**	14.09	**12.41**	**12.41**	154.6	154.6	154.6	10.51	10.51	10.51
M13	233.1	233.1	233.1	12.56	12.56	12.56	198.8	198.8	198.8	11.65	11.65	11.65

Note: see Table 5.

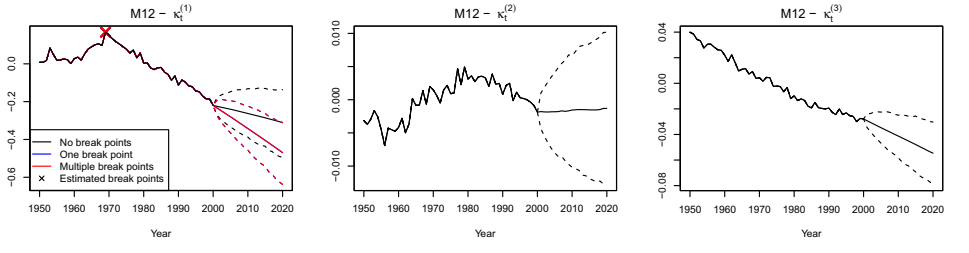

Figure 5. Projections for the period effects of M12 applied to Dutch females aged 20–89 in the period 1950–2000. The structural change for $\kappa_t^{(1)}$ is identified both if we allow for one and if we allow for multiple structural changes. (Coloured versions of the figures can be found online.)

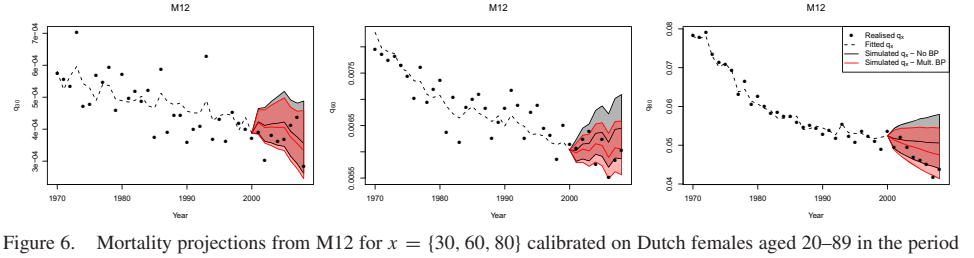

Figure 6. Mortality projections from M12 for $x = \{30, 60, 80\}$ calibrated on Dutch females aged 20–89 in the period 1950–2000. The black and red lines represent projections without and with multiple structural changes, respectively, at the 5th, 50th and 95th percentile.

the realised mortality rates closely. Hence, even though the projected period effect 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.

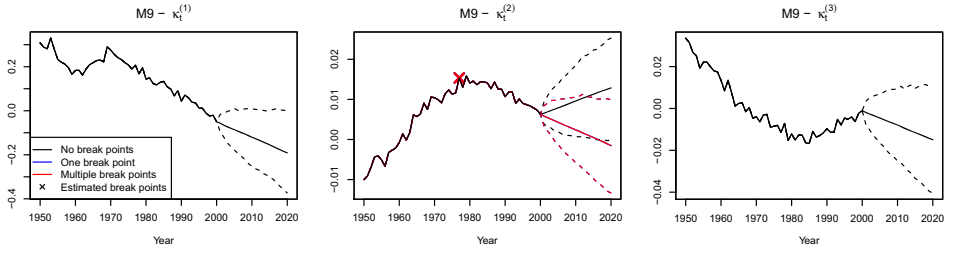

Figure 7. Projections for the period effects of M9 applied to Dutch females aged 20–89 in the period 1950–2000. The structural change for $\kappa_t^{(2)}$ is identified both if we allow for one and if we allow for multiple structural changes.

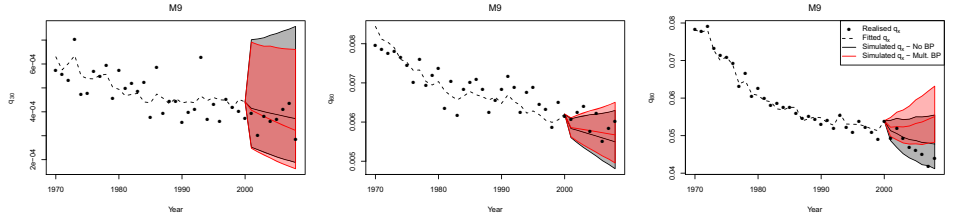

Figure 8. Mortality projections from M9 for $x = \{30, 60, 80\}$ calibrated on Dutch females aged 20–89 in the period 1950–2000. The black and red lines represent projections without and with multiple structural changes, respectively, at the 5th, 50th and 95th percentile.

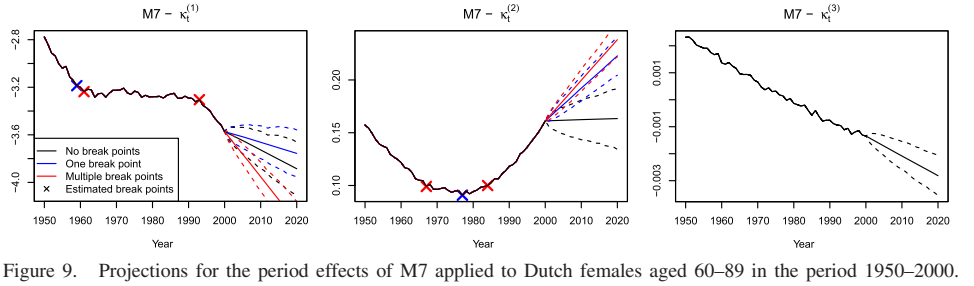

Figure 9. Projections for the period effects of M7 applied to Dutch females aged 60–89 in the period 1950–2000.

(Coloured versions of the figures can be found online.)

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 9 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 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. We observe that projections of the period effects are most robust with respect to the calibration period if we allow for multiple structural changes.

We compare the impact on mortality projections of not allowing for structural changes with allowing for a single or multiple structural changes. We find evidence for one structural change, and sometimes even multiple structural changes are estimated. We also find that allowing for structural changes can lead to improved backtesting results. Allowing for structural changes does not always lead to improved backtesting results, because apparent structural changes may not be identified until sufficient evidence for their existence has accumulated, i.e. the improvement in fit from including a structural change is not sufficient yet to overcome the penalty in BIC caused by the extra parameter. Another explanation why backtesting results may not have improved is because changes in age effects have not been accounted for Zhao & Sweeting (2012) propose a method to account for this, but further investigation is needed.

The model we propose relaxes the assumption that all parameter values remain constant over the considered time period. We check for different mortality trends in the period effects and use the latest trend for projecting mortality. In that sense it resembles methods in which the calibration period is restricted to a particular subset of recent data points which is chosen to provide the best model fit. Such alternative methods also allow that other, age-dependent, parameters are only fitted for this restricted period and this may improve fit for the most recent observations.

Our approach has the advantage that it can still be used when one requires that a model structure describes the entire collection of data points. This would for example be the case if we want to compare the performance of different model structures for a given dataset. If such structures involve more than one stochastic factor, we do not have to exclude the possibility that one of the multiple time series undergoes a structural change while the others remain the same as before and we do not need to adjust the overall calibration period as a result of such a change.

Each approach therefore has its advantages and disadvantages, but it is reassuring that our numerical example for a single factor model suggests that estimates generated by the two methods will not differ substantially in their fit over the most recent years.

In this paper, the mortality model and the time series models are estimated separately. 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

The authors would like to thank the anonymous referees who provided helpful suggestions to improve an earlier draft of this paper. 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.

References

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

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

Bai, J. & 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. & Salhi, Y. (2012). Understanding, modelling and managing longevity risk: key issues and main challenges. *Scandinavian Actuarial Journal* **3**, 203–231.

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

Bots, M. & 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. & 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. & 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. & Dowd, K. (2008). Modelling and management of mortality risk: A review. *Scandinavian Actuarial Journal* **2–3**, 79–113.

Cairns, A., Blake, D., Dowd, K., Coughlan, G., Epstein, D. & 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. & 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. & Nunes, L. (2011). Forecasting mortality in the event of a structural change. *Journal of the Royal Statistical Society* **174** (3), 713–736.

Coelho, E. & Nunes, L., (2013). Cohort effects and structural changes in the mortality trend. Working paper. Available online at: http://www.unece.org/leadmin/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. Available online at: http://www.ma.hw.ac.uk/~iain/research/talks.html.

Denuit, M. & Goderniaux, A., (2005). Closing and projecting lifetables using log-linear models. *Bulletin of the Swiss Association of Actuaries* **1**, 29–49.

Dowd, K., Cairns, A., Blake, D., Coughlan, G., Epstein, D. & 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. & Raftery, A. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102** (477), 359–378.
Haberman, S. & 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. & 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. & Taylor, A. (2009). Simple, robust and powerful tests of changing trend hypothesis. *Econometric Theory* **25**, 995–1029.
Janssen, F., Kunst, A. & Mackenbach, J. (2007). Variations in the pace of old-age mortality decline in seven European countries, 1950–1999: The role of smoking and other factors earlier in life. *European Journal of Population* **23** (2), 171–188.
Lee, R. & Carter, L. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association* **87** (419), 659–671.
Li, H., Waegenaere, A.D. & 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. & 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. & Cox, S. (2011). Mortality regimes and pricing. *North American Actuarial Journal* **15** (2), 266–289.
Moreno-Serra, R. & 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. & Li, Y. (2011). Explaining young mortality. *Insurance: Mathematics and Economics* **50** (1), 12–25.
O'Hare, C. & Li, Y., (2014). Identifying structural breaks in stochastic mortality models. *Journal of Risk and Uncertainty in Engineering part B*. Available online at SSRN: http://ssrn.com/abstract=2192208.
Pitacco, E., Denuit, M., Haberman, S. & Olivieri, A. (2009). *Modelling longevity dynamics for pensions and annuity business*. New York: Oxford University Press.
Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics* **45** (3), 393–404.
Renshaw, A. & Haberman, S. (2003). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics* **33** (2), 255–272.
Renshaw, A. & 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. & 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. & Hornik, K. (2003). Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis* **44** (12), 109–123.
Zeileis, A., Leisch, F., Hornik, K. & Kleiber, C. (2002). Strucchange: An *R* package for testing for structural change in linear regression models. *Journal of Statistical Software* **7** (2), 1–38.
Zhao, Y. & Sweeting, P. (2012). Modelling the cohort effect in CBD models using a piecewise linear approach. Discussion paper. Pensions Institute.

Appendix 1. Parameter constraints

Some of the mortality models experience identifiability issues. Therefore, we impose parameter constraints. Table A1 provides an overview of the parameter constraints that are imposed on the models.

Table A1. Overview of the parameter constraints 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	$\sum_c \gamma_c = 0$	$\sum_c c\gamma_c = 0$	$\sum_t \kappa_t^{(3)} = 0$		