MDPI

Article

Projecting mortality rates using a Markov chain

Jaap Spreeuw [1], Iqbal Owadally [1] and Muhammad Kashif [2,*]

[1] Bayes Business School, City, University of London, 106 Bunhill Row, London EC1Y 8TZ, United Kingdom; j.spreeuw@city.ac.uk (J.S.), m.i.owadally@city.ac.uk (I.O.)
[2] School of Business and Economics, Universidad de las Americas Puebla, Sta. Catarina Mártir, Cholula, 72810 Puebla, México; muhammad.kashif@udlap.mx
* Correspondence: muhammad.kashif@udlap.mx

Abstract: We present a mortality model where future stochastic changes in population-wide mortality are driven by a finite-state hierarchical Markov chain. A baseline mortality in an initial 'Alive' state is calculated as the average logarithm of observed mortality rates. There are several more 'Alive' states and a jump to the next 'Alive' state leads to a change (typically, an improvement) in mortality. In order to estimate the model parameters, we minimize a weighted average quadratic distance between observed mortality rates and expected mortality rates. A two-step estimation procedure is used, and a closed-form solution for the optimal estimates of model parameters is derived in the first step, which means that the model can be parameterized very fast and efficiently. The model is then extended to allow for age effects whereby stochastic mortality improvements also depend on age. Forecasting relies on state space augmentation and an innovations state space time series model. We show that, in terms of forecasting, our model outperforms a naïve model of static mortality within a few years. The Markov approach also permits an exact computation of mortality indices like the complete expectation of life and annuity present values which are key in the life insurance and pensions industry.

Keywords: mortality forecasting; Markov chain; model calibration; life insurance; pensions

Citation: Spreeuw, J.; Owadally, I.; Kashif, M. Projecting mortality rates using a Markov chain. *Mathematics* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

Mathematical modeling of mortality trends is becoming a central concern for researchers and practitioners due to its importance for public health planning, social insurance, private life insurance, and pension systems. Accurate mortality forecasts are critical to allocate resources in a timely manner for forward planning. In this context, prolonged life expectancy, also known as longevity risk, poses challenges for the pricing, advance funding and reserving of life insurance and pension schemes, which may require forecasts up to 50 years ahead. Generally, it is difficult to measure and hedge the effects of mortality improvement on retirement planning. Ideally, the difference between observed mortality and mortality estimates should be negligible. However, over the last several decades, old-age mortality projections have underestimated mortality improvement. The aim of this study is to introduce a new method for population-wide mortality modeling that is driven by a finite-state hierarchical Markov chain.

The literature on stochastic mortality models has been developing rapidly over the past 25 to 30 years. Mortality forecasting approaches can be classified into three main categories: extrapolative, explanatory, expectation. The extrapolative approach applies simple extrapolation to measures like life expectancy considering that observed age patterns and trends exhibit regularity over time. One example of this approach is the seminal Lee-Carter model [1], which is a discrete-time model that is driven by a time series component. The explanatory approach makes use of epidemiological or structural models to forecast mortality by cause of death, where exogenous variables are measurable and known. The expectation-based models set the parameters of mortality by fitting deterministic functions

to recent trends or by consultation with demographers and other experts. Such models are only useful for short-term forecasting, as they do not capture the stochastic features of mortality

Recent reviews of mainstream mortality forecasting models can be found in [2–9] and the references therein. The vast majority of mortality models are extrapolative and are easier to apply and are more accurate than the other approaches [10]. Over the last few decades, the most prevalent mortality forecasting method has been the Lee-Carter model and its variants. The Lee-Carter model decomposes age-specific mortality into an overall time trend over a certain time period. The model extrapolates the overall time trend using past time series to forecast the underlying factors of force of mortality [1]. The main advantage of the Lee-Carter model is its robustness in the case of linear past trend and a simple stochastic model is able to forecast age pattern of mortality with one time-varying parameter.

More recently, various modifications in estimation methods have been made to the original Lee-Carter model by Lee and Miller [11], Booth *et al.* [12] and Brouhns *et al.* [13] and Hatzopoulos and Haberman [14]. In some other extensions, nonparametric smoothing, Kalman Filtering and multiple principal components are included: see Hyndman and Ullah [15] and De Jong and Tickle [16]. Booth *et al.* [17] compare the performance of five different extensions of the Lee-Carter model. The Cairns-Blake-Dowd (CBD) model introduces the use of logit of the death probabilities as a linear or quadratic function of age to better capture mortality at older age [2,18,19]. Furthermore, for non-linear trends, the cohort parameter was included to improve mortality prediction by Renshaw and Haberman [20], Plat [21], Cairns *et al.* [22] and Reither *et al.* [23]. Other related studies include the application of machine learning in standard stochastic mortality models to identify the patterns and calibrate parameters to improve the goodness of fit [24]. Atance *et al.* [25] compare the Lee–Carter model and its extended two-factor version to predict dynamic life tables and conclude that the Lee-Carter model projects mortality better than other versions.

This paper introduces a new model of stochastic mortality based on a time-homogeneous continuous-time Markov chain of mortality changes. The model allows for age effects whereby mortality improvements differentially impact individuals of different ages. To forecast mortality rates, states are added to the Markov chain and an innovations state space time series model is employed. To the best of our knowledge, such models have not been employed in the mortality forecasting literature. Our model is inspired by the one discussed by Norberg [26] except that, rather than involving specific causes of death which may diminish over time, we look at mortality in aggregate terms. Markov models have also been applied to human mortality by Lin and Liu [27] and Liu and Lin [28], although in a different way from what we propose in this article. Both of these papers employ a finite-state Markov model to capture the human ageing process. In Liu and Lin [28], the Markov model is subordinated by a gamma process to allow for stochastic mortality. By contrast, in our paper, the Markov model itself drives the stochastic mortality.

Our model has three major advantages. First, it is flexible: one can create as many states in the Markov model as one sees fit. Even with 200 states, say, the computations are still fast, and transparency is not compromised. Second, the calibration part is easy to implement and can even be performed on a spreadsheet. Finally, once the forecasting part has been completed, i.e. projections have been made of future changes in mortality, it is straightforward to calculate the exact distributions of key quantities like (future) expectancies of life and fixed-rate annuity present values. These can be obtained by solving Thiele's differential equation (Dickson *et al.* [29, p. 211]).

The setup of this paper is as follows. Section 2 introduces the model while section 3 describes the mortality data, from the well-known Human Mortality Database, which is used to calibrate the model. The calibration procedure and its results are described in section 4. In section 5, the innovations state space time series model as in Hyndman *et al.* [30] is introduced and applied to forecasting mortality rates. The forecasting power of the model, as compared with a naïve model of static mortality, is also discussed. Applications

in life insurance and pensions are briefly presented in section 6, and section 7 concludes. The technical details of the innovations state space model are left to Appendix A.

2. The model

Markov processes have found their applications in life insurance mathematics for more than 50 years. Amsler [31] gave a seminal lecture at the 18th International Congress of Actuaries, while a publication by Hoem [32] introduced the Markov model in the actuarial literature. Of the many publications about Markov processes, the textbooks by Haberman and Pitacco [33] (in particular the applications in disability insurance), Wolthuis [34] and Dickson *et al.* [29, ch. 8] are noteworthy.

A basic survival model that is used in elementary life insurance mathematics is the Markov process which involves only two states: an initial 'Alive' state and a terminal 'Dead' state, denoted by a and d respectively. The instantaneous transition intensity from 'Alive' to 'Dead' is then an instantaneous mortality rate which depends on age. For an x-year old individual, the instantaneous mortality rate is denoted by $\mu(x)$. Let $p_{jk}(s,t)$, with $j,k \in \{a,d\}$ and for $0 \leq s \leq t$, be the transition probability of a life being in state k at time t given that the life is in state j at time s. Since the 'dead' state is an absorbing state, it follows that

$$\begin{aligned} p_{dd}(s,t) &= 1 - p_{da}(s,t) = 1, \\ p_{aa}(s,t) &= 1 - p_{ad}(s,t) = e^{-\int_s^t \mu_{x+u} du}. \end{aligned}$$

Our model extends the basic 2-state model above by augmenting the state space to allow for mortality improvements. We consider a continuous-time, time-homogeneous hierarchical Markov process consisting of $N+2$ states, i.e. $N+1$ strongly transient 'Alive' states and one 'Dead' state, the latter denoted by D. This is depicted in Figure 1. Each individual aged x starts in 'Alive' state 0 with age-dependent instantaneous rate of mortality $\mu_x^{(0)}$ specified as $\mu_x^{(0)} = \mu_x^s \exp(\gamma(0))$, where μ_x^s denotes the standard (or benchmark or baseline) rate of mortality, while $\exp(\gamma(0))$ reflects the relative difference between the initial mortality and the standard mortality. From 'Alive' state $i \in [0, N)$, an individual can only make a transition either to (a) the next 'Alive' state $i+1$ with intensity $\lambda(i)$, which is independent of time but allowed to depend on the state of sojourn, or to (b) the 'Dead' state with rate of mortality $\mu_x^{(i)}$. Thus, $\mu_x^{(i)}$ denotes the instantaneous rate of mortality for an x-year old in state i, with $i \in \{0, .., N-1\}$. In the last 'Alive' state N, only a transition to the 'Dead' state, with mortality rate $\mu_x^{(N)}$ is possible. This implies that any life can experience at most N age-independent changes in mortality.

We first set up a preliminary model where improvements in population-wide mortality proceed as follows. A transition from state i to state $i+1$ entails a relative improvement in mortality of $100(1 - \exp(\gamma(i+1)))\%$ at all ages, so $\mu_x^{(i+1)} = \exp(\gamma(i+1))\mu_x^{(i)}$ for all age x. Thus, $\gamma(i+1)$ is the log-change in instantaneous mortality rate from state i to state $i+1$. It is also useful to define $\Gamma(i) = \sum_{j=0}^{i} \gamma(j)$ as the cumulative log-change in mortality rate by state i.

Notice that the transition intensities do not depend on time. The transition intensity $\lambda(i)$ from one 'Alive' state i to the next 'Alive' state $i+1$ depends only on state i. The mortality rate $\mu_x^{(i)}$ depends on the age x of an individual, but not on clock time t for the population as a whole. The Markov chain is therefore a time-homogeneous process capturing population-wide mortality improvements.

For this hierarchical Markov chain, using the same definition as above for the transition probability $p_{jk}(s,t)$, but now with $j,k \in \{0,1,...,N,D\}$, we have the following expressions:

$$\begin{aligned} p_{ik}(s,t) &= 0, \quad \text{for } i \in \{1,\ldots,N\} \text{ and } k < i, \\ p_{ii}(s,t) &= e^{-\int_s^t \left(\lambda(i) + \mu_{x+u}^{(i)}\right) du}, \quad \text{for } i \in \{0,\ldots,N-1\}, \end{aligned}$$

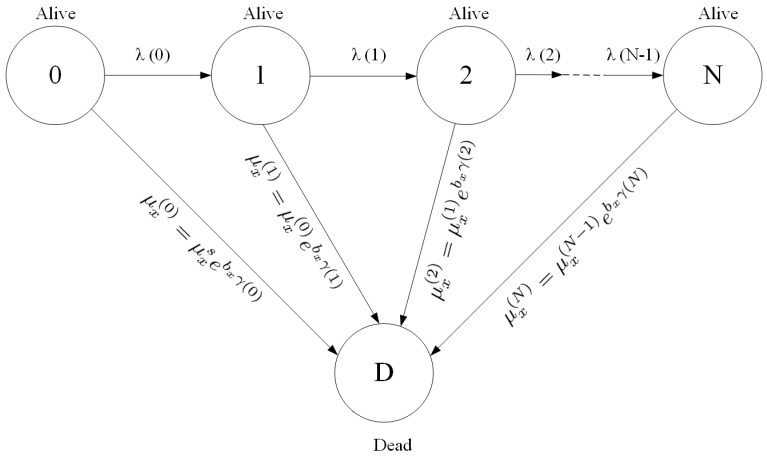

Figure 1. Transition diagram for the Markov chain model of population-wide mortality changes. Instantaneous transition intensities are shown for each allowable transition. There are $N+1$ "Alive" states and one "Dead" state. Age is denoted by x and, in the preliminary model, there are no age effects ($b_x = 1$).

$$p_{NN}(s,t) \;=\; e^{-\int_s^t \mu_{x+u}^{(N)} du}.$$

In reality, when population-wide mortality improvements occur, they will be of a different magnitude at different ages. Our full model therefore extends the preliminary model by allowing for age effects, which are captured by the factor b_x for an individual aged x. In the 'Alive' state 0, the mortality rate for an individual aged x is $\mu_x^{(0)} = \mu_x^s \exp(b_x\gamma(0))$. For any x-year old, a transition from state i to state $i+1$ entails a relative improvement in mortality of $100(1-\exp(b_x\gamma(i+1)))\%$, so that $\mu_x^{(i+1)} = \mu_x^{(i)} \exp(b_x\gamma(i+1))$.

3. Mortality data

We consider a life insurer, at the end of year 2000, which has mortality data for several years until the year 2000. The data pertains to female policyholders of ages varying from 20 to a limiting age, which is assumed to be 105. We choose female mortality data purely for illustrative purposes, and could equally have chosen male mortality data. (Most mortality forecasting studies use either male or female data, unless a gender-comparative analysis is being undertaken, e.g. Chiou and Müller [35] use female data whereas Cairns *et al.* [2] use male data.) The life insurer is in charge of devising a sound model of stochastic mortality, using the available mortality statistics which we assume are drawn from the Human Mortality Database (2020) [36].

At present, the Human Mortality Database (HMD) [36] contains detailed population and mortality data for 41 countries or areas. This includes input data like death counts, census counts, birth counts and population estimates. Such input data enable the calculation of key quantities like exposure to risk, death rates and life tables of national populations, which can also be found in the HMD [36]. The period of time for which complete data is available varies across countries or areas but usually involves at least a couple of decades. Most countries or areas are highly industrialized and relatively wealthy.

Two further points are noteworthy. First, we disregard ages younger than 20 years old to avoid unnecessary complexity. As pointed out by Jarner and Kryger [37], the pattern of infant and child mortality is different from adult mortality. Furthermore, in most developed countries, current levels of young-age mortality are very low.

Second, we will fit the model on 51 years of in-sample annual mortality rates from year 1950 to 2000. We will use the mortality rates from 2001 onwards as out-of-sample data to assess the forecast error of our model. We also have mortality data at ages 20 to 104 for each year of in-sample data, giving $51 \times 85 = 4,335$ data points. This is far greater than the

number of parameters to estimate in our model (provided that the number of states in the Markov model is not excessively large), and there is therefore no danger of overfitting the model to the available data.

4. Model calibration to mortality data

4.1. Outline of calibration procedure

For a given value of N, we fit our model to the mortality data by minimizing the *weighted average quadratic distance* (WAQD) between expected log-mortality and observed log-mortality. The WAQD is defined precisely below, for both the preliminary model and the full model with age effects. In the full model, there are $2N + 86$ parameters: transition intensities $\lambda(0), \ldots, \lambda(N-1)$; mortality improvement factors $\gamma(0), \ldots, \gamma(N)$; age effect factors $b_{20}, \ldots, b_{104}$ at ages $20, \ldots, 104$ respectively. Ideally, all these parameter values would be found by simultaneously minimizing the WAQD wrt all the parameters. Unfortunately, this is either not computationally feasible or very time-consuming, unless N is small.

We proceed using a pragmatic approach instead. Consider first the preliminary model where age effects are not included. We fix the transition intensities between the various 'Alive' states to be constant across all the states, $\lambda(k) \equiv \lambda$ for $k \in [0, N)$, and we then derive mathematically an expression for the optimal estimates of the mortality improvement factors by minimizing the WAQD wrt $\{\gamma(k), k \in [0, N]\}$. We then perform a grid search for the minimum WAQD on a grid spanned by λ, refining the grid at any local minimum. For reasonably small N, this is feasible because we have an easily and quickly calculated closed-form expression for the optimal $\{\gamma(k), k \in [0, N]\}$.

Fixing λ and then minimizing the WAQD wrt $\{\gamma(k), k \in [0, N]\}$ might at first sight appear to be an implausible shortcut. It is, in fact, perfectly justifiable. A direct analogy is to Girsanov's change-of-measure theorem (Itô [38], p. 1535): one can change the probability measure underlying a standard Brownian motion, but then distort its state space by imposing a local drift which reverses the change in measure. In financial mathematics, this underpins the pricing of securities in complete markets when an artificial risk-neutral probability measure is used along with a risk-free rate to discount the security payoffs in the state space (Shreve [39], p. 216).

Thus armed with the optimal estimates of λ and $\{\gamma(k), k \in [0, N]\}$ in the preliminary model, we turn to estimating all the parameters, including the age effect factors $\{b_x, x \in [20, 104]\}$, in the full model. We will show that we can use the preliminary model parameters as the input or first-stage values to a recursive scheme which minimizes the WAQD. Repeated substitution leads to numerical convergence to our final estimates of all the parameter values.

In the next sections, we describe in greater detail the procedure that we use for estimation, first for the preliminary model and then for the full model.

4.2. Preliminary model without age effects

Denote the state of the Markov chain at time t by X_t. Let $p_{0k}(t)$ be the transition probability of a life being in state k in the middle of year t given that the life starts in state 0 at time 0, i.e. $p_{0k}(t) = \mathbb{P}[X_{t+1/2} = k \mid X_0 = 0]$. (The observed instantaneous rate of mortality for a given year estimates the rate of mortality at the midpoint of the year.)

The observed rate of mortality $\widehat{\mu}_{x,t}$ at age x in year t is calculated as $\widehat{\mu}_{x,t} = d_{x,t}/E^c_{x,t}$. Here, $d_{x,t}$ is the number of deaths recorded at age x last birthday during calendar year t. Furthermore, $E^c_{x,t}$ is the exposure-to-risk at age x last birthday during year t (that is, the total time lived by people aged x last birthday in calendar year t) [40, p. 95-96]. This data is obtained from the HMD [36]. Similar to Lee and Carter [1], the log of the estimated mortality rate $\overline{\mu}(x)$ at age x is then obtained as a weighted average of the log of the observed mortality rates across the years of observation:

$$\ln[\overline{\mu}(x)] \;=\; \frac{\sum_{t=0}^{2000-Y} w_{x,t} \ln \widehat{\mu}_{x,t}}{\sum_{t=0}^{2000-Y} w_{x,t}}, \tag{1}$$

with Y being the starting year of the period of observation and the year 2000 being the last year. The estimated mortality rate at age x from the mortality data in the HMD [36] is then taken to be the standard or baseline mortality at age x in the Markov model, i.e. in Figure 1, $\mu^s_x = \overline{\mu}(x)$. In equation (1) above, $w_{x,t}$ denotes the weight assigned to the exposure in year t at age x. It is common to use $w_{x,t} = (Var[\widehat{\mu}_{x,t}])^{-1}$, which can be estimated by $w_{x,t} = \left(E^c_{x,t}\right)^2/d_{x,t}$ [41, p. 321-322].

The WAQD pertaining to calendar year t is defined as

$$W_t \;=\; \sum_{k=0}^{N} p_{0k}(t) \sum_{x=20}^{104} w_{x,t} \left(\sum_{j=0}^{k} \gamma(j) + \ln \overline{\mu}(x) - \ln \widehat{\mu}_{x,t} \right)^2 . \tag{2}$$

This is similar to the objective functions used by Lee and Carter [1] and Pitacco *et al.* [40, p. 190]. The term in parentheses on the r.h.s. of equation (2) is evidently the difference between the estimated log-mortality rate and the observed log-mortality rate plus the cumulative log-change in mortality rates by state k which is reached at time t, at a given age x. Because we may have more observations at certain ages than at other ages (e.g. there are more younger individuals than older individuals in the population), the quadratic deviation is then weighted by exposure weights. An expectation is then computed by summing the preceding quantity over the whole state space weighted by the probability of reaching state k by time t, given the starting state 0 at time 0. The total WAQD is then found by summing the right hand side of equation (2) over the years from the start of observation in year Y to the end in year 2000, i.e. $W = \sum_{t=0}^{2000-Y} W_t$.

As explained in section 4.1, the first step in our calibration procedure is to minimize the WAQD wrt $\{\gamma(k), k \in [0, N]\}$. Proposition 1 sets this out below. It is helpful to recall and introduce some notation ahead of the statement and proof of Proposition 1. From section 2, $\Gamma(k) = \sum_{j=0}^{k} \gamma(j)$, for $k = 0, \ldots, N$, is the cumulative log-change in mortality rate by state k of the Markov chain, i.e. the log-change in mortality rate from state 0 to state k. Recall also that $p_{0k}(t) = \mathbb{P}[X_t = k \mid X_0 = 0]$ is the transition probability of being in state k at time t given the starting state 0 at time 0. Let the transition probability vector from state 0 over time t be $\mathbf{p}(t) = (p_{00}(t), p_{01}(t), \ldots, p_{0N}(t))^{\mathrm{T}}$, and let $\mathbf{P}(t) = \mathrm{diag}(\mathbf{p}(t)) \in \mathbb{R}^{(N+1)\times(N+1)}$, i.e. $\mathbf{P}(t)$ is a diagonal matrix whose leading diagonal is made up of elements of $\mathbf{p}(t)$.

Proposition 1. *The values of $\gamma(k)$ which minimize the weighted average quadratic distance (WAQD) between expected and observed mortality are given by*

$$\gamma(k) \;= \boldsymbol{\imath}_{k+1}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{c} - \boldsymbol{\imath}_{k}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{c} \tag{3}$$

for $k = 1, \ldots, N$, and $\gamma(0) = \boldsymbol{\imath}_1^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{c}$.

In the above, $\boldsymbol{\imath}_k$ is a column vector of zeros except for 1 in row k. Further, $\mathbf{A} = \sum_{t=0}^{2000-Y} h_1(t)\, \mathbf{P}(t)$, with $h_1(t) = \sum_{x=20}^{104} w_{x,t}$. Also, $\mathbf{c} = \sum_{t=0}^{2000-Y} h_2(t)\, \mathbf{p}(t)$, with $h_2(t) = \sum_{x=20}^{104} w_{x,t} (\ln \widehat{\mu}_{x,t} - \ln \overline{\mu}(x))$.

Proof. [Proof of Proposition 1] The total WAQD is

$$W \;=\; \sum_{t=0}^{2000-Y} \sum_{k=0}^{N} p_{0k}(t) \sum_{x=20}^{104} w_{x,t} [\Gamma(k) + \ln \overline{\mu}(x) - \ln \widehat{\mu}_{x,t}]^2 . \tag{4}$$

Expanding the term in square brackets in the equation above, and using $h_1(t)$ and $h_2(t)$ as defined in Proposition 1, as well as $h_3(t) = \sum_{x=20}^{104} w_{x,t} (\ln \widehat{\mu}_{x,t} - \ln \overline{\mu}(x))^2$, we can simplify the WAQD to

$$W \;=\; \sum_{t=0}^{2000-Y} \sum_{k=0}^{N} p_{0k}(t) \left[\Gamma(k)^2 h_1(t) \;-\; 2\Gamma(k) h_2(t) \;+\; h_3(t) \right]. \tag{5}$$

Now, $h_1(t)$, $h_2(t)$ and $h_3(t)$ can be taken out of the inner summation. Let $\mathbf{\Gamma} = (\Gamma(0), \Gamma(1), \ldots, \Gamma(N))^{\mathrm{T}}$. Using $\mathbf{p}(t)$ and $\mathbf{P}(t)$ as defined just before Proposition 1, we obtain

$$W \;=\; \sum_{t=0}^{2000-Y} \Big[h_1(t) \mathbf{\Gamma}^{\mathrm{T}} \mathbf{P} \mathbf{\Gamma} \;-\; 2 h_2(t) \mathbf{p}^{\mathrm{T}} \mathbf{\Gamma} \;+\; h_3(t) \mathbf{p}^{\mathrm{T}} \mathbf{1} \Big], \tag{6}$$

where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^{N+1}$. Note that we suppress the dependence of $\mathbf{p}(t)$ and $\mathbf{P}(t)$ on t in the notation hereinafter, for the sake of clarity.

The elements of the diagonal matrix $\mathbf{P}$ are positive, hence $\mathbf{P}$ is positive definite. (The elements in the leading diagonal of $\mathbf{P}$ are the non-zero transition probabilities to the various 'Alive' states of the Markov chain. The eigenvalues of the diagonal matrix $\mathbf{P}$ are positive. By a well-known theorem of matrices—see, for example, Theorem 2 of Johnson [42], or Itô [38], p. 996— $\mathbf{P}$ is therefore positive definite.) Furthermore, $h_1(t) = \sum_{x=20}^{104} w_{x,t} > 0$ since $w_{x,t} > 0$. From the quadratic form in equation (6), we conclude that the existence and uniqueness of a minimum in W wrt. $\mathbf{\Gamma}$ are guaranteed.

$$\frac{\partial W}{\partial \mathbf{\Gamma}} \;=\; \sum_{t=0}^{2000-Y} \left[2 h_1(t) \mathbf{P} \mathbf{\Gamma} - 2 h_2(t) \mathbf{p} \right] \;=\; 2 \mathbf{A} \mathbf{\Gamma} - 2 \mathbf{c}, \tag{7}$$

where $\mathbf{A}$ and $\mathbf{c}$ are defined in the Proposition. To minimize W, we solve $\partial W / \partial \mathbf{\Gamma} = \mathbf{0}$, where $\mathbf{0} \in \mathbb{R}^{N+1}$ is a column vector of zeros, giving $\mathbf{\Gamma} = \mathbf{A}^{-1} \mathbf{c}$. Note that $\mathbf{A} = \sum_{t=0}^{2000-Y} h_1(t)\, \mathbf{P}(t)$ is invertible since $h_1(t) \neq 0$, and $\mathbf{P}$ is non-singular. ($\mathbf{P}$ is non-singular since its leading diagonal elements are non-zero, as discussed above. By a well-known theorem of matrices—see for example Perlis [43], p. 72—the determinant of $\mathbf{P}$, a diagonal matrix, is the product of these non-zero elements, and is therefore non-zero.)

Finally, since $\Gamma(k) = \sum_{j=0}^{k} \gamma(j)$, it follows that $\gamma(0) = \Gamma(0) = \boldsymbol{\iota}_1^{\mathrm{T}} \mathbf{\Gamma}$ and $\gamma(k) = \Gamma(k) - \Gamma(k-1) = \left(\boldsymbol{\iota}_{k+1}^{\mathrm{T}} \mathbf{\Gamma} \right) - \left(\boldsymbol{\iota}_k^{\mathrm{T}} \mathbf{\Gamma} \right)$ for $k = 1, \ldots, N$. □

As explained in the outline of our calibration procedure in section 4.1, we make some simplifying assumptions for the sake of parsimonious modelling and to keep the estimation as straightforward as possible. First, we assume that the exposure weights $w_{x,t}$ at age x and time t, as introduced in equation (1) and used in the WAQD in equation (2), are set equal to one, $w_{x,t} \equiv 1$, for the sake of simplicity. As argued by Pitacco *et al.* [40, p. 190], using weights that are not exogenous, in that they depend on the random number of deaths, is questionable, especially for stochastic mortality models in contrast to the static 'life tables' used for insurance pricing purposes. Second, we assume henceforth that the transition intensity $\lambda > 0$, from one 'Alive' state to the next, is constant not just in time but also over the state space. We still require a numerical search procedure for λ when minimizing the WAQD, but the overall estimation procedure is simplified. In particular, $p_{0k}(t)$ can be expressed simply as:

$$p_{0k}(t) \;=\; \frac{1}{k!} (\lambda t)^k e^{-\lambda t}, \quad \text{for } k = 0, \ldots, N. \tag{8}$$

Equation (8) above follows from the fact that, conditional on no death occurring, the transitions out of any 'Alive' state $k < N+1$ in the Markov process are restricted to those of a time-homogeneous pure birth process with rate λ.

Since $p_{0k}(t)$ in equation (8) features a maximum wrt λ (at $\lambda = k/t$ provided $k > 0$, $t > 0$) and no minimum, it is worth investigating whether W can indeed be minimized wrt λ, i.e. it is worth investigating the existence of an optimal estimate of λ using WAQD.

Denoting by $H(k,t)$ the expression in the square brackets in equation (5), we can express the WAQD in a compact fashion: $W = \sum_{t=0}^{2000-Y} \sum_{k=0}^{N} p_{0k}(t) H(k,t)$. Since $\partial p_{0k}(t)/\partial\lambda = (\lambda t)^k e^{-\lambda t}(k-\lambda t)/(\lambda k!)$, it follows that

$$\frac{\partial W}{\partial \lambda} = \sum_{t=0}^{2000-Y} \sum_{k=0}^{N} \frac{1}{\lambda k!} (\lambda t)^k e^{-\lambda t} (k - \lambda t) H(k,t). \tag{9}$$

An analytical expression for λ in the solution of $\partial W/\partial\lambda = 0$ is difficult to find, especially for a Markov chain with a large state space (large N), but numerical estimates can easily be computed.

As for the existence of a minimum in W wrt λ, we note that $\partial^2 p_{0k}(t)/\partial\lambda^2 = (\lambda t)^k e^{-\lambda t}[(k-\lambda t)^2 - k]/(\lambda^2 k!)$, so that

$$\frac{\partial^2 W}{\partial \lambda^2} = \sum_{t=0}^{2000-Y} \sum_{k=0}^{N} \frac{1}{\lambda^2 k!} (\lambda t)^k e^{-\lambda t} [(k - \lambda t)^2 - k] H(k,t). \tag{10}$$

For $k \geq 0$ and $t > 1$, it is easy to see that, disregarding the term in square brackets in equation (10), the summand inside the double summation in equation (10) is positive. In particular, $H(k,t) > 0$ since it is identical to the innermost summand in equation (4). Whether W is convex wrt λ therefore rests on a weighted sum of terms in $[(k-\lambda t)^2 - k]$. We cannot formally show the existence of a minimum, but the above analysis serves two purposes. First, it reassures us that the absence of a minimum in $p_{0k}(t)$ does not rule out a minimum in W. Second, it illustrates the difficulty in deriving the optimal parameter estimates analytically, thereby justifying our two-step estimation procedure.

4.3. Full model with age effects

For the more comprehensive model allowing for age effects, i.e. with structure $\mu_x^{(i+1)} = \mu_x^{(i)} \exp(b_x \gamma(i))$ as spelled out in section 2, the WAQD pertaining to calendar year t is

$$\widetilde{W}_t = \sum_{k=0}^{N} p_{0k}(t) \sum_{x=20}^{104} w_{x,t} \left(b_x \sum_{j=0}^{k} \gamma(j) + \ln \overline{\mu}(x) - \ln \widehat{\mu}_{x,t} \right)^2. \tag{11}$$

As before, the total WAQD from year Y to year 2000 is then $\widetilde{W} = \sum_{t=0}^{2000-Y} \widetilde{W}_t$.

The WAQD then needs to be optimized with respect to the mortality improvement factors $\gamma(k)$ for $k = 0,\ldots,N$, as well as the age effects b_x for $x = 20,\ldots,104$. As in the proof of Proposition 1, let the cumulative sum of the mortality improvements be $\Gamma(k) = \sum_{j=0}^{k} \gamma(j)$, for $k = 0,\ldots,N$, and let $\boldsymbol{\Gamma} = (\Gamma(0),\Gamma(1),\ldots,\Gamma(N))^{\mathrm{T}}$. Furthermore, define $\mathbf{b} = (b_{20}, b_{21}, \ldots, b_{104})^{\mathrm{T}}$. The following system of equations has to be solved:

$$\frac{\partial \widetilde{W}}{\partial \boldsymbol{\Gamma}} = 0 \quad \Leftrightarrow \quad \Gamma(k) = \boldsymbol{\iota}_{k+1}^{\mathrm{T}} \widetilde{\mathbf{A}}^{-1} \widetilde{\mathbf{c}}, \tag{12a}$$

$$\frac{\partial \widetilde{W}}{\partial \mathbf{b}} = 0 \quad \Leftrightarrow \quad b_x = \boldsymbol{\iota}_{x-19}^{\mathrm{T}} \mathbf{B}^{-1} \mathbf{d}. \tag{12b}$$

In equation (12a), $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{c}}$ are, respectively, versions of $\mathbf{A}$ and $\mathbf{c}$ (as defined above in Proposition 1) which are modified to allow for the age effects. Specifically, $\widetilde{\mathbf{A}} = \sum_{t=0}^{2000-Y} \widetilde{h}_1(t)\mathbf{P}(t)$ with $\widetilde{h}_1(t) = \sum_{x=20}^{104} w_{x,t} b_x^2$, while $\widetilde{\mathbf{c}} = \sum_{t=0}^{2000-Y} \widetilde{h}_2(t)\mathbf{p}(t)$ with $\widetilde{h}_2(t) = \sum_{x=20}^{104} w_{x,t} b_x (\ln \widehat{\mu}_{x,t} - \ln \overline{\mu}(x))$. The vector $\mathbf{p}(t)$ and the matrix $\mathbf{P}(t)$ are unchanged from Proposition 1.

In equation (12b), $\mathbf{B} = \sum_{t=0}^{2000-Y} h_4(t)\mathbf{V}(t)$. Here, $h_4(t) = \sum_{k=0}^{N} p_{0k}(t)\Gamma(k)^2$. We observe that $h_4(t)$ is the second moment of $\Gamma(X_t)$ where X_t is the random state of the Markov chain at time t, i.e. $h_4(t) = \mathbb{E}[\Gamma(X_t)^2]$. Furthermore, $\mathbf{V}(t) = \mathrm{diag}(\mathbf{w}(t))$, i.e.

N	$N/50$	$\widehat{\lambda}$
25	0.5	0.77
50	1	1.42
100	2	2.69
150	3	3.91
200	4	5.06
300	6	7.27
400	8	9.45

Table 1. Optimal parameter values for the transition intensity from one 'Alive' state to the next in the preliminary Markov model, for different values of N, i.e. for different number of states in the Markov chain. 51 years of observations are used from starting year $Y = 1950$.

$\mathbf{V}(t)$ is a diagonal matrix whose leading diagonal is made up of elements of $\mathbf{w}(t) = (w_{20,t}, w_{21,t}, \ldots, w_{104,t})^{\mathrm{T}}$.

Finally, in equation (12b), we also have $\mathbf{d} = \sum_{t=0}^{2000-Y} h_5(t)\mathbf{z}(t)$. Here, $h_5(t) = \sum_{k=0}^{N} p_{0k}(t)\,\Gamma(k)$, and we observe that $h_5(t)$ is the first moment of $\Gamma(X_t)$ (compare with $h_4(t)$ above which was the second moment). Furthermore, $\mathbf{z}(t) = (z_{20,t}, z_{21,t}, \ldots, z_{104,t})^{\mathrm{T}}$ where $z_{j,t} = w_{j,t}(\ln \widehat{\mu}_{j,t} - \ln \overline{\mu}(x))$ for $j \in [20, 104]$.

The system of equations (12a)–(12b) can be solved numerically by successive substitution as follows. At the first stage, start with the preliminary model wihout age effects, i.e. $b_x = 1$ at all ages $x \in [20, 104]$. Cumulative changes to mortality $\Gamma(k)$ in all states $k \in [0, N]$ can then be calculated using Proposition 1. These are then substituted into the r.h.s. of equation (12b) yielding second-stage values for b_x at all ages $x \in [20, 104]$. In turn, these are substituted into the r.h.s. of equation (12a), leading to second-stage values for $\Gamma(k)$ in all states $k \in [0, N]$, and so on, until convergence is reached.

4.4. Results of calibration

Our calibration procedure generates estimates for all the parameters of our Markov model, but it does not tell us what the size of the Markov chain should be, i.e. the number of states. We suggest that this can be determined by numerical experimentation. For the preliminary model, the calibration has been run with several values of N.

We find that, as N increases, the minimum WAQD decreases. This is as anticipated because the more states there are, the more parameters are involved, and the better the fit. We also find that, as N increases, our (WAQD-minimizing) estimate of λ increases. This is illustrated in Table 1, where we have 51 years of annual mortality data, from 1950 onwards, on which the model is calibrated and we choose N to be equal to 50 times 0.5, 1, 2... Adding states without changing λ means that the probability of eventually entering the last few "Alive" states will become smaller and eventually negligible. In order to significantly improve the fit, these probabilities need to be sufficiently different from zero, which is achieved by increasing λ so that the process traverses as much of the state space as possible.

Finally we also find that, as N increases, the estimated log-changes to mortality rates, $\gamma(k)$, $k \in [0, N]$, decrease (results not shown here for economy of space). More states lead to more transitions if the transition intensity λ increases. To compensate for this, the impact of each transition should be smaller.

For the full Markov model with $N = 50$, parameterized from 51 years of mortality data from 1950, we estimate $\lambda = 1.29$. Tables 2 and 3 show the optimal parameter values for $\Gamma(k)$, $k \in [0, N]$ and b_x, $x \in [20, 104]$ respectively. We observe from Table 2 that $\Gamma(0) = \gamma(0)$ is strongly positive, indicating that mortality in the initial state is much higher than the average mortality across the period of investigation. This is self-evidence since mortality rates decrease during the period. For the same reason, it is unsurprising that Γ decreases as a function of state, reaching negative values upon reaching state 33. At that point,

State k	$\Gamma(k)$	State k	$\Gamma(k)$	State k	$\Gamma(k)$
0	37.21046234	17	10.16128483	34	−1.282926153
1	34.89110164	18	9.43041363	35	−2.123815014
2	31.30856332	19	8.708274309	36	−2.999173623
3	27.90632886	20	8.001132141	37	−3.904036891
4	25.19136301	21	7.314907495	38	−4.831988911
5	23.04653425	22	6.653009452	39	−5.775537066
6	21.28096167	23	6.015309668	40	−6.726552135
7	19.7463605	24	5.398173284	41	−7.676727809
8	18.36637132	25	4.79519957	42	−8.618012162
9	17.11439091	26	4.19826336	43	−9.542970976
10	15.98204253	27	3.598533779	44	−10.44505611
11	14.96112827	28	2.987293383	45	−11.31876798
12	14.0375066	29	2.356516698	46	−12.15971639
13	13.19117416	30	1.699253023	47	−12.96459607
14	12.39925669	31	1.009887793	48	−13.7311004
15	11.64013218	32	0.284343793	49	−14.45779864
16	10.89723104	33	−0.479749777	50	−19.13930163

Table 2. Optimal parameter values for $\Gamma(k) = \sum_{i=0}^{k} \gamma(i)$ at each state $k \in [0, 50]$ for the full Markov model with $\lambda = 1.29$, $N = 50$, parameterized from 51 years of mortality data from 1950.

mortality is below average and will reduce further. The values in Table 3 follow a less monotone pattern. However, we can observe that b_x is relatively high, and therefore mortality improvements more pronounced, for young ages, up to age 45, say, compared to later ages. Also note that the values for b_x are small for $x \geq 90$, so for very high ages mortality improvements do not have a very significant impact. This suggests the possible existence of an upper limit to lifespan. This is in line with expectation-based methods, where mortality reductions are observed to be greater for younger ages, see e.g. [44].

5. Forecasting

5.1. Forecasting procedure

To test the forecasting power of our Markov model, we calibrate it on (in-sample) mortality data from 1950 until 2000, and then forecast mortality from 2001 onwards. We can then compare our forecast mortality rates to (out-of-sample) mortality data from 2001.

In order to create the forecasts, the Markov model must be augmented by new 'Alive' states. This is akin to forecasting a Markov counting process and adding integer states. Whilst we can use the estimated transition intensity $\widehat{\lambda}$ and the estimated age effect factors $\widehat{b}_x$, $x \in [20, 104]$ from the model calibration stage, the mortality change factors $\{\gamma(k)\}$ over the augmented states must themselves be forecast.

In order to project the mortality change factors, we employ an innovations state space model (Hyndman *et al.* [30]). This is a richer class of models than the classical Holt-Winters exponential smoothing model (Hyndman *et al.* [45]). Trends and seasonal components, which may be of either additive or multiplicative form, are simultaneously estimated (Ord *et al.* [46]). We use the bias-corrected Akaike Information Criterion (cAIC) to select the best model in the class of innnovations state space models: this turns out to be the so-called "damped trend" model (McKenzie and Gardner [47]; Hyndman *et al.* [30], p. 48), which is reported to be highly successful in terms of forecast accuracy when applied to different types of data (Makridakis and Hibon [48]; Gardner and McKenzie [49]; Fildes [50]). Parameter estimation is performed via maximum likelihood estimation, and both point forecasts and prediction intervals can be generated. Refer to Appendix A for details.

For the full model with age effects and $N = 50$, a plot of $\Gamma(k)$, along with forecasts and confidence intervals, is shown in Figure 2.

Age x	b_x	Age x	b_x	Age x	b_x	Age x	b_x
20	0.014261355	42	0.014812453	64	0.009601416	86	0.011695384
21	0.015866235	43	0.014119139	65	0.009402891	87	0.010812765
22	0.017292544	44	0.013620313	66	0.00894529	88	0.010131609
23	0.018935807	45	0.013317686	67	0.009874399	89	0.00983736
24	0.018301737	46	0.013626309	68	0.010134124	90	0.008982384
25	0.01955903	47	0.013238125	69	0.010570269	91	0.0078466
26	0.019086171	48	0.013587528	70	0.010295663	92	0.007807907
27	0.018720848	49	0.013508995	71	0.009893928	93	0.007240778
28	0.018955846	50	0.012225902	72	0.011135998	94	0.006614019
29	0.019767372	51	0.011260071	73	0.01159652	95	0.00612189
30	0.019052062	52	0.011779514	74	0.012122628	96	0.005999284
31	0.017292291	53	0.011757458	75	0.012241192	97	0.005087022
32	0.018619477	54	0.011429729	76	0.012617455	98	0.005715154
33	0.016869444	55	0.009619608	77	0.012072695	99	0.004075368
34	0.017544975	56	0.010643902	78	0.012779524	100	0.004026139
35	0.016240706	57	0.009609826	79	0.012945643	101	0.004006924
36	0.015965233	58	0.010172365	80	0.012194586	102	0.00183354
37	0.016163254	59	0.009579479	81	0.011769632	103	0.003193209
38	0.016294748	60	0.008964135	82	0.012136992	104	−0.00459685
39	0.015642919	61	0.00845602	83	0.0122118		
40	0.015160679	62	0.009601562	84	0.012377463		
41	0.014380544	63	0.010045288	85	0.011700719		

Table 3. Optimal parameter values for b_x, $x \in [20, 104]$ for the full Markov model with $\lambda = 1.29$, $N = 50$, parameterized from 51 years of mortality data from 1950.

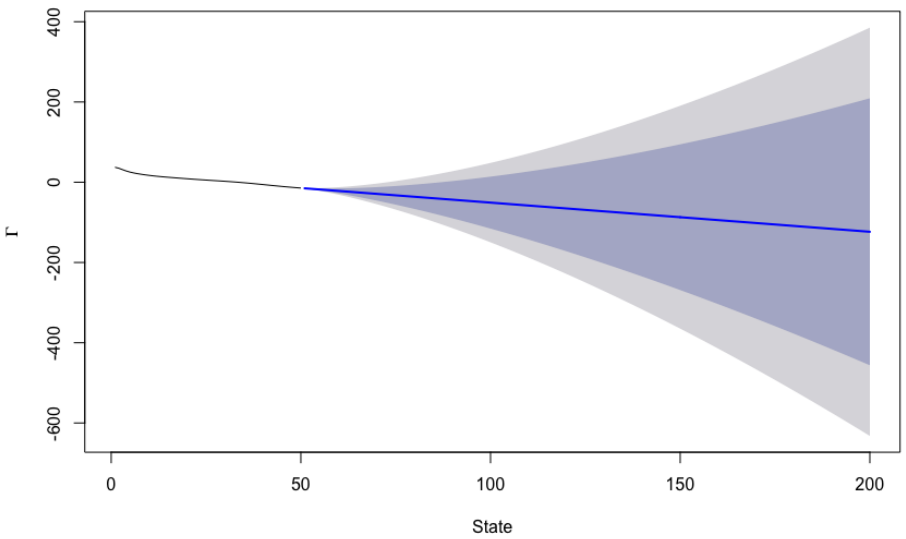

Figure 2. Plot of cumulative mortality change factor $\Gamma(k)$ for the full model with age effects and $N = 50$. Calibrated values from in-sample mortality data (starting year $Y = 1950$) are shown up to state 50, and forecasts and confidence intervals are shown for later states.

5.2. Forecast accuracy

The forecast accuracy of our Markov model may be assessed by comparing the actual mortality rates in the out-of-sample years from 2001 onwards to the forecast mortality rates in these out-of-sample years according to our model. As in the seminal Lee-Carter stochastic mortality model [1], we wish to compare the log of the instantaneous mortality rates.

Let $\widetilde{\mu}(x,t)$ be a random variable denoting the instantaneous mortality rate at age x in an out-of-sample year t according to our model. Our central forecast of the log-mortality rate is $\mathbb{E}[\ln \widetilde{\mu}(x,t)]$. By summing the squared deviation between our central forecast and the observed log-mortality over all ages, the forecast error in an out-of-sample year t can therefore be measured as:

$$\sum_{x=20}^{104} \Big(\mathbb{E}[\ln \widetilde{\mu}(x,t)] - \ln \widehat{\mu}_{x,t}\Big)^2. \tag{13}$$

In the Markov model, from Figure 1, the log-mortality rate at age x when in state k is

$$\ln \mu_x^{(k)} \;=\; \ln \mu_x^{(k-1)} + b_x \gamma(k) \;=\; \ln \mu_x^s + \sum_{j=0}^{k} b_x \gamma(j), \tag{14}$$

where the standard or baseline mortality is estimated from the mortality data in the HND [36], $\mu_x^s = \overline{\mu}(x)$, as explained near equation (1). Thus, the expected log-mortality at age x and time t in the Markov process is the log-mortality at age x when in state k in equation (14), weighted by the probability that the process is in state k after t years given the starting state 0, summed over all possible values of state k:

$$\mathbb{E}[\ln \widetilde{\mu}(x,t)] \;=\; \sum_k p_{0k}(t) \left(\sum_{j=0}^{k} b_x \gamma(j) + \ln \overline{\mu}(x) \right). \tag{15}$$

Substituting the expected log-mortality in equation (15) above into equation (13) gives the forecast error in an out-of-sample year t:

$$\sum_{x=20}^{104} \left[\left(b_x \sum_k p_{0k}(t) \sum_{j=0}^{k} \gamma(j) \right) + \ln \overline{\mu}(x) - \ln \widehat{\mu}_{x,t} \right]^2, \tag{16}$$

noting that $\sum_k p_{0k}(t) = 1$. Note that the error calculated in equation (16) above for forecasting purposes is subtly different from the WAQD in equation (11) used for calibration purposes. Note also that the forecast error can be readily and exactly calculated without need for simulations or approximations.

Table 4 lists the forecast errors at different out-of-sample years for three models: a naïve model where mortality is static and remains as in year 2000; the Markov model calibrated with $N = 50$ and $Y = 1950$ and the Markov model calibrated with $N = 10$ and $Y = 1990$. Recall from section 2 that there are $N + 2$ states in total: an initial 'Alive' state with preliminary mortality, a terminal 'Dead' state, and N 'Alive' states with improved mortality. So the second and third Markov models in Table 4 have 52 and 12 states in total respectively. Note also that these models are the full Markov models which allow for age effects.

We observe from Table 4 that the forecast errors for all three models generally increase the further out one is in the out-of-sample period, as one might anticipate. Judging by the total forecast errors over all the out-of-sample years (in the bottom row of Table 4), the Markov models clearly outperform the naïve model. This lends credibility to our Markov modelling approach.

Somewhat surprisingly, the Markov model with fewer states outperforms the other Markov model in Table 4. However, our initial investigations show that it is not clear-cut

Year	Naïve	$N = 50$	$N = 10$
2001	0.348	1.054	0.349
2002	0.341	1.150	0.299
2003	0.511	1.443	0.387
2004	0.743	1.718	0.545
2005	1.147	1.568	0.685
2006	1.530	2.013	0.955
2007	1.870	2.069	1.079
2008	1.761	2.427	0.992
2009	3.008	3.149	1.767
2010	3.641	2.972	1.927
2011	4.207	3.390	2.116
2012	4.985	2.868	2.412
2013	5.140	3.171	2.334
2014	5.421	3.701	2.269
2015	5.110	3.332	2.406
2016	4.642	3.574	2.192
Total	44.405	39.599	22.714

Table 4. Forecast errors at different out-of-sample years and total forecast error for three models: a naïve model with static mortality, the Markov model with $N = 50$, and the Markov model with $N = 10$. The Markov models are the full model incorporating age effects, and N determines the size of the state space of the model.

that fewer states lead to better forecasts. Further research will be required to be more conclusive. The results in Table 4 serve mainly to illustrate that the Markov model is a viable model that can perform well in terms of forecasting mortality.

6. Applications in life insurance and pensions

In the life insurance and pensions business, models of mortality are of critical importance. Commonly adopted measures of mortality changes include distributions of expectation of life and distributions of present values of annuities at future durations and ages. A key advantage of the Markov approach in this paper is that such indices can be calculated exactly by solving Thiele's differential equations (Dickson *et al.* [29, p. 266]), These differential equations enable an insurer to calculate the reserves that it needs to hold when it sells a portfolio of life insurance policies. With the Markov chain approach, we can add the different states directly to Thiele's differential equations and solve the differential equations numerically at multiple durations. This does not require any simulations, unlike other models of stochastic mortality such as the Lee-Carter model [1].

We give a brief example to illustrate this. If $\bar{e}_x^{(j)}$ denotes the complete expectation of life of an x year old in state j, the appropriate Thiele's differential equation would be for $j \in \{0, \ldots, N\}$:

$$\frac{d}{dt}\bar{e}_{x+t}^{(j)} \;=\; -1 \;-\; \sum_{k=j+1}^{N+1} \lambda\left(\bar{e}_{x+t}^{(k)} - \bar{e}_{x+t}^{(j)}\right) \;+\; \mu_{x+t}^{(s)} \exp[b_{x+t}\Gamma(j)]\bar{e}_{x+t}^{(j)}. \tag{17}$$

For $j = N + 1$, this equation reduces to

$$\frac{d}{dt}\bar{e}_{x+t}^{(N+1)} \;=\; -1 \;+\; \mu_{x+t}^{(s)} \exp[b_{x+t}\Gamma(N+1)]\bar{e}_{x+t}^{(N+1)}, \tag{18}$$

while

$$\frac{d}{dt}\bar{e}_{x+t}^{(N+2)} \;=\; 0. \tag{19}$$

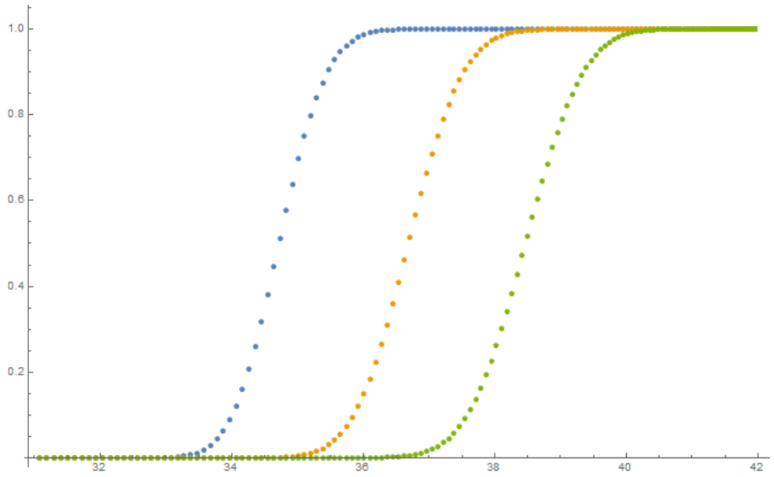

Figure 3. Cumulative distribution function of complete expectation of life for a 50-year old at durations 25 (blue), 40 (orange) and 55 (green).

The appropriate boundary conditions are $\bar{e}_\omega^{(j)} = 0$ for $j \in \{0, .., N+2\}$, where ω denotes the limiting age of a life. In addition, $\mu^{(s)}$ and b are defined for non-integer ages by applying polynomial interpolation between integer ages.

For durations 25, 40 and 55 (so calendar years 2000, 2015 and 2030), the cumulative distribution functions (CDFs) of complete expectation of life are displayed for ages 50 and 80 in Figures 3 and 4, respectively. For a 50-year old, the mean life expectancies are 34.77, 36.73 and 38.50, respectively. For an 80-year old, they are 8.97, 10.03 and 11.05, respectively.

From Figures 3 and 4, we notice that the CDFs move to the right as duration goes up. This is not surprising, when we consider the extrapolative nature of forecasting. Mortality improvements have been observed during the periods of observation, so we would expect mortality improvements to continue in future years. The variability of remaining lifetime is for age 80 than for age 50, due to the more limited remaining life span. Figures 3 and 4 capture the variability of future remaining lifetimes, and therefore the number of years that annuities or pensions will remain payable. They can therefore help pension and annuity providers to determine the amount of capital to hold to cover longevity risk. The CDFs can also help national local governments with future general public planning (health care needs, etc.).

7. Conclusion

In this paper, we introduce a Markov chain model for stochastic mortality based on time-homogeneous continuous-time mortality changes, and we demonstrate its advantages in terms of flexibility and ease of calibration. We model age-independent changes in mortality by means of transitions across several 'Alive' states, along with a terminal 'Dead' state. Our preliminary model considers mortality improvements in population-wide mortality, whereas our full model allows for age effects in mortality improvements.

We use female mortality statistics drawn from the Human Mortality Database to calibrate our models using a two-step estimation procedure. In the first step, we obtain a closed-form solution to the minimization of a weighted average quadratic distance (WAQD) with respect to the cumulative log-change in mortality rates, and we then numerically estimate the transition intensities. Our investigation shows that the choice of total number of 'Alive' states is critical. On the one hand, the greater the number of states, the better the fit to the data. One the other hand, a model with a higher number of states means that more mortality change factors are to be estimated and may lead to overfitting.

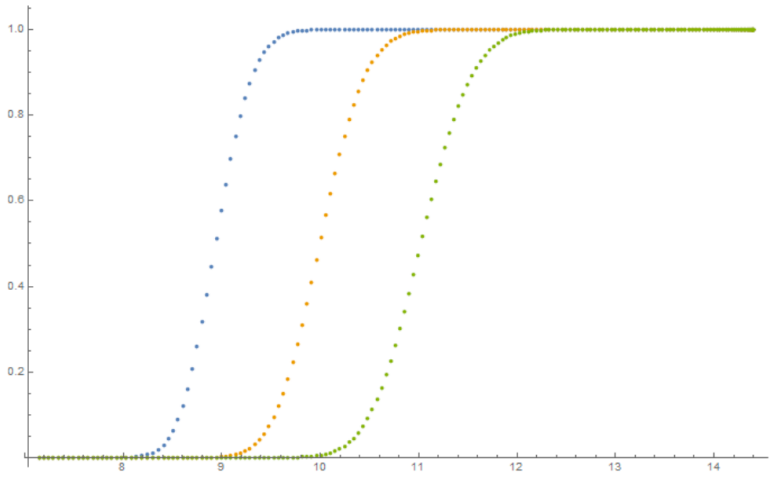

Figure 4. Cumulative distribution function of complete expectation of life for an 80-year old at durations 25 (blue), 40 (orange) and 55 (green).

We calibrate the models on in-sample mortality data from 1950 until 2000, and then forecast mortality from 2001 onwards. Our forecast can then be compared with the out-of-sample data. We employ an innovations state space model, in particular the damped trend model, to project the mortality change factors. We use these forecasts along with the estimated transition intensity and age effect factors for forecasting. We compare the actual mortality rates to the forecast mortality rates for out-of-sample data to find the forecast error for three models: naïve (with static mortality rates), full Markov model with 50 states and starting year 1950, and full Markov model with 10 states and starting year 1990. The Markov models exhibit a lower forecast error than the static model. As expected, the forecast error increases as we move further out of sample.

Finally, we present an application of our Markov approach to life insurance and pensions. Key mortality change indicators like the distributions of life expectancy and expected present values of annuities are easily calculated using Thiele's differential equations. This should facilitate the estimation and management of longevity risk by life insurers, pension providers and others.

The main novelty of our model is the application of both Markov chains and innovations state space models to the mortality forecasting problem. Our method has many advantages including flexibility and ease of parameterisation. With regard to flexibility, as many mortality improvement states as required can be added to the model. As demonstrated in this work, the model can be easily calibrated to real mortality data. Life expectancies and reserves required for life insurance and pensions are also easily computed, without recourse to simulations. Therefore, our model can help practitioners forecast mortality and manage longevity risk more easily.

Our work has some limitations that require further investigation and exploration. Our model is fitted by minimizing the WAQD, but other criteria could be used for this purpose. In addition, the transition intensities are assumed to be constant between the various 'Alive' states. Furthermore, our estimation method has a limitation due to the two-step method that we utilize, whereby mortality rates are first estimated and then an innovations state space model is used for projection. The model also disregards idiosyncratic shocks to mortality such as Covid mortality.

For future work, we intend to undertake a more rigorous and systematic investigation into the combination of factors, such as the number of states, period of investigation and transition intensities in the model, that delivers the best forecast accuracy. This will also enable us to compare the performance of our model with that of a mainstream one like the Lee-Carter model. In this paper, only point forecasts have been used in judging the

forecasting power of our model. Another topic for future research would involve using also information about prediction intervals, reflecting the parameter uncertainty of the model.

Author Contributions: Conceptualization, J.S. and I.O.; methodology, J.S. and I.O.; formal analysis, J.S. and I.O.; investigation, J.S. and I.O.; resources, I.O.; data curation, J.S.; software, J.S., I.O. and M.K.; calibration, J.S. and I.O.; writing—original draft preparation, J.S., I.O.and M.K.; writing—review and editing, I.O. and M.K.; visualization, J.S. and I.O.; supervision, J.S. and I.O.; project administration, I.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data on population estimates and mortality rates is available on the Human Mortality Database: https://www.mortality.org/ (accessed on 5 March 2020).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Innovations state space model

In order to forecast the coefficients $\{\gamma(k)\}$, we assume that they are realizations of a stochastic process indexed by the stage $k \in \mathbb{N}$ of mortality improvement. The stochastic process is described by an innovations state space model, which is described briefly here. For details, see Hyndman *et al.* [30].

The innovations state space model can be written by means of an observation equation

$$\gamma(k) = \ell_{k-1} + \phi b_{k-1} + \varepsilon_k, \tag{A1}$$

and two state equations

$$\ell_k = \ell_{k-1} + \phi b_{k-1} + \alpha\varepsilon_k, \tag{A2}$$
$$b_k = \phi b_{k-1} + \beta\varepsilon_k. \tag{A3}$$

Here, ℓ_k denotes the level of the data, superposed on a trend b_k, along with additive noise ε_k which is identically Normally distributed with zero mean and variance σ_ε^2. The parameters α and β are smoothing parameters for the level and trend respectively, whilst ϕ controls the speed at which the trend flattens out.

Three basic specifications exist, depending on the values of the three parameters: $\{\alpha \in (0,1),\ \beta \equiv \phi \equiv 0\}$ or $\{\alpha, \beta \in (0,1),\ \phi \equiv 1\}$ or $\{\alpha, \beta, \phi \in (0,1)\}$. This is extended to a total of 10 specifications by allowing one or both of the trend and the error to enter multiplicatively into the observation and state equations: for details, see Hyndman *et al.* [30]. (Seasonal components can also be incorporated, but visual inspection does not reveal any seasonality, so this is ignored here.)

We omit the last estimated value of $\gamma(k)$, pertaining to the terminal state of our Markov model. Since this last state is an absorbing state, γ_N is an outlier as a result of boundary effects. We then choose the best model by minimizing the bias-corrected Akaike Information Criterion (cAIC).

Parameter values are found by maximizing likelihood, as described by Hyndman *et al.* [30]. Initial values of the state variables are chosen according to a heuristic scheme which is empirically verified by Hyndman *et al.* [45]. Parameter estimates, along with initialization values and cAIC values, are displayed in Table A1. (If a parameter value appears as 0 or 1, this means that the model specification is such that the parameter is identical to 0 or 1, respectively.)

Point forecasts are readily calculated by substitution and iteration in the observation and state equations (A1)–(A3), with the error term replaced by its mean of zero. Confidence levels can also be calculated since closed-form expressions for conditional variances are known for the cAIC-minimizing models that are specified for our data (see Hyndman *et al.* [30]). A plot of $\Gamma(k)$, with forecasts and confidence intervals, is shown in Figure 2.

Y	N	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\phi}$	ℓ_0	b_0	σ_ε	cAIC
1950	50	0.9983	0.9983	1	39.55	-2.3105	0.2439	61.69
1990	10	0.0001	0.0001	1	4.4896	-0.5626	0.3163	18.08

Table A1. MLE parameter estimates ($\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\phi}$), initialization values (ℓ_0, b_0), standard error σ_ε of innovations, and bias-corrected Akaike Information Criterion (cAIC) for innovations state space model fitted to $\Gamma(k)$ for the two Markov models described in section 5.2.

References

1. Lee, R.D.; Carter, L.R. Modeling and forecasting US mortality. *Journal of the American Statistical Association* **1992**, *87*, 659–671. doi:10.1080/01621459.1992.10475265.
2. Cairns, A.J.G.; Blake, D.; Dowd, K.; Coughlan, G.D.; Epstein, D.; Ong, A.; Balevich, I. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal* **2009**, *13*, 1–35. doi:10.1080/10920277.2009.10597572.
3. Cairns, A.J.G.; Blake, D.; Dowd, K.; Coughlan, G.D.; Epstein, D.; Khalaf-Allah, M. Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics* **2011a**, *48*, 355–367. doi:10.1016/j.insmatheco.2010.12.005.
4. Dowd, K.; Cairns, A.J.G.; Blake, D.; Coughlan, G.D.; Epstein, D.; Khalaf-Allah, M. Evaluating the goodness of fit of stochastic mortality models. *Insurance: Mathematics and Economics* **2010a**, *47*, 255–265. doi:10.1016/j.insmatheco.2010.06.006.
5. Dowd, K.; Cairns, A.J.G.; Blake, D.; Coughlan, G.D.; Epstein, D.; Khalaf-Allah, M. Backtesting stochastic mortality models: an ex post evaluation of multiperiod-ahead density forecasts. *North American Actuarial Journal* **2010b**, *14*, 281–298. doi:10.1080/10920277.2010.10597592.
6. Haberman, S.; Renshaw, A.E. A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics* **2011**, *48*, 35–55. doi:10.1016/j.insmatheco.2010.09.003.
7. Stoeldraijer, L.; Van Duin, C.; Van Wissen, L.; Janssen, F. Impact of different mortality forecasting methods and explicit assumptions on projected future life expectancy: The case of the Netherlands. *Demographic Research* **2013**, *29*, 323–354.
8. Guibert, Q.; Lopez, O.; Piette, P. Forecasting mortality rate improvements with a high-dimensional VAR. *Insurance: Mathematics and Economics* **2019**, *88*, 255–272.
9. Hunt, A.; Blake, D. On the structure and classification of mortality models. *North American Actuarial Journal* **2021**, *25*, S215–S234. doi:10.1080/10920277.2019.1649156.
10. Booth, H.; Tickle, L. Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science* **2008**, *3*, 3–43. doi:10.1017/S1748499500000440.
11. Lee, R.; Miller, T. Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography* **2001**, *38*, 537–549. doi:10.1353/dem.2001.0036.
12. Booth, H.; Maindonald, J.; Smith, L. Applying Lee-Carter under conditions of variable mortality decline. *Population Studies* **2002**, *56*, 325–336.
13. Brouhns, N.; Denuit, M.; Vermunt, J.K. A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics* **2002**, *31*, 373–393. doi:10.1016/S0167-6687(02)00185-3.
14. Hatzopoulos, P.; Haberman, S. A parameterized approach to modeling and forecasting mortality. *Insurance: Mathematics and Economics* **2009**, *44*, 103–123. doi:10.1016/j.insmatheco.2008.10.008.
15. Hyndman, R.J.; Ullah, M.S. Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis* **2007**, *51*, 4942–4956. doi:10.1016/j.csda.2006.07.028.
16. De Jong, P.; Tickle, L. Extending Lee–Carter mortality forecasting. *Mathematical Population Studies* **2006**, *13*, 1–18. doi:10.1080/08898480500452109.
17. Booth, H.; Hyndman, R.J.; Tickle, L.; De Jong, P. Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research* **2006**, *15*, 289–310.
18. Cairns, A.J.G.; Blake, D.; Dowd, K. A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance* **2006**, *73*, 687–718. doi:10.1111/j.1539-6975.2006.00195.x.
19. Li, H.; O'Hare, C. Semi-parametric extensions of the Cairns–Blake–Dowd model: a one-dimensional kernel smoothing approach. *Insurance: Mathematics and Economics* **2017**, *77*, 166–176. doi:10.1016/j.insmatheco.2017.10.001.
20. Renshaw, A.E.; Haberman, S. A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics* **2006**, *38*, 556–570. doi:10.1016/j.insmatheco.2005.12.001.
21. Plat, R. On stochastic mortality modeling. *Insurance: Mathematics and Economics* **2009**, *45*, 393–404.
22. Cairns, A.J.G.; Blake, D.; Dowd, K.; Coughlan, G.D.; Khalaf-Allah, M. Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin: The Journal of the IAA* **2011b**, *41*, 29–59. doi:10.2143/AST.41.1.2084385.
23. Reither, E.N.; Olshansky, S.J.; Yang, Y. New forecasting methodology indicates more disease and earlier mortality ahead for today's younger Americans. *Health Affairs* **2011**, *30*, 1562–1568. doi:10.1377/hlthaff.2011.0092.

24. Levantesi, S.; Pizzorusso, V. Application of machine learning to mortality modeling and forecasting. *Risks* **2019**, *7*, 26. doi:10.3390/risks7010026.
25. Atance, D.; Debón, A.; Navarro, E. A comparison of forecasting mortality models using resampling methods. *Mathematics* **2020**, *8*, 1550. doi:10.3390/math8091550.
26. Norberg, R. Optimal hedging of demographic risk in life insurance. *Finance and Stochastics* **2013**, *17*, 197–222. doi:10.1007/s00780-012-0182-3.
27. Lin, X.S.; Liu, X. Markov aging process and phase-type law of mortality. *North American Actuarial Journal* **2007**, *11*, 92–109. doi:10.1080/10920277.2007.10597486.
28. Liu, X.; Lin, X.S. A subordinated Markov model for stochastic mortality. *European Actuarial Journal* **2012**, *2*, 105–127. doi:10.1007/s13385-012-0047-3.
29. Dickson, D.C.M.; Hardy, M.R.; Waters, H.R. *Actuarial Mathematics for Life Contingent Risks*, 2nd ed.; Cambridge University Press: New York, 2013.
30. Hyndman, R.J.; Koehler, A.B.; Ord, J.K.; Snyder, R.D. *Forecasting with exponential smoothing: the state space approach*; Springer: Berlin, 2008.
31. Amsler, M.H. Les chaines de Markov des assurances vie, invalidité et maladie. In Proceedings of the Transactions of the 18th International Congress of Actuaries; , 1968; Vol. 5, pp. 731–746.
32. Hoem, J.M. Markov chain models in life insurance. *Blätter der DGVFM* **1969**, *9*, 91–107.
33. Haberman, S.; Pitacco, E. *Actuarial Models for Disability Insurance*; Chapman & Hall, London, UK., 2018.
34. Wolthuis, H. *Life Insurance Mathematics (The Markovian model).*; IAE, Universiteit van Amsterdam, Amsterdam, the Netherlands, 2003.
35. Chiou, J.M.; Müller, H.G. Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association* **2009**, *104*, 572–585.
36. Human Mortality Database, https://www.mortality.org/, 2020.
37. Jarner, S.F.; Kryger, E.M. Modelling adult mortality in small populations: The SAINT model. *ASTIN Bulletin* **2011**, *41*, 377–418. doi:10.2143/AST.41.2.2136982.
38. Itô, K. *Encyclopedic Dictionary of Mathematics*, 2nd ed.; MIT press, 1993.
39. Shreve, S.E. *Stochastic Calculus for Finance II: Continuous-Time Models*; Vol. 11, Springer: New York, 2004.
40. Pitacco, E.; Denuit, M.; Haberman, S.; Olivieri, A. *Modelling Longevity Dynamics for Pensions and Annuity Business*; Oxford University Press: New York, 2009.
41. Benjamin, B.; Pollard, J.H. *The analysis of mortality and other actuarial statistics*; Vol. 3, The Institute of Actuaries, Oxford, 1993.
42. Johnson, C.R. Positive definite matrices. *The American Mathematical Monthly* **1970**, *77*, 259–264.
43. Perlis, S. *Theory of Matrices*; Addison-Wesley, Reading, Massachusetts, 1952.
44. CMIB. Report no. 17 Continuous Mortality Investigation Bureau. Technical report, Institute and Faculty of Actuaries, 1999.
45. Hyndman, R.J.; Koehler, A.B.; Snyder, R.D.; Grose, S. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting* **2002**, *18*, 439–454. doi:10.1016/S0169-2070(01)00110-8.
46. Ord, J.K.; Koehler, A.B.; Snyder, R.D. Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association* **1997**, *92*, 1621–1629. doi:10.1080/01621459.1997.10473684.
47. McKenzie, E.; Gardner, E.S. Damped trend exponential smoothing: a modelling viewpoint. *International Journal of Forecasting* **2010**, *26*, 661–665. doi:10.1016/j.ijforecast.2009.07.001.
48. Makridakis, S.; Hibon, M. The M3-Competition: results, conclusions and implications. *International Journal of Forecasting* **2000**, *16*, 451–476. doi:10.1016/S0169-2070(00)00057-1.
49. Gardner, E.S.; McKenzie, E. Why the damped trend works. *Journal of the Operational Research Society* **2011**, *62*, 1177–1180. doi:10.1057/jors.2010.37.
50. Fildes, R. The evaluation of extrapolative forecasting methods. *International Journal of Forecasting* **1992**, *8*, 81–98. doi:10.1016/0169-2070(92)90009-X.