Pensions
Institute

DISCUSSION PAPER PI-0802

Evaluating the goodness of fit of stochastic mortality models

Kevin Dowd, Andrew J.G. Cairns, David Blake, Guy D. Coughlan, David Epstein, and Marwa Khalaf-Allah

December 2010

ISSN 1367-580X

The Pensions Institute
Cass Business School
City University
106 Bunhill Row London
EC1Y 8TZ
UNITED KINGDOM

http://www.pensions-institute.org/

Insurance: Mathematics and Economics 47 (2010) 255–265

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Evaluating the goodness of fit of stochastic mortality models

Kevin Dowd [a,*], Andrew J.G. Cairns [b], David Blake [a], Guy D. Coughlan [c], David Epstein [c], Marwa Khalaf-Allah [c]

[a] *Pensions Institute, Cass Business School, 106 Bunhill Row, London, EC1Y 8TZ, United Kingdom*

[b] *Maxwell Institute for Mathematical Sciences, Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom*

[c] *Pension ALM Group, JPMorgan Chase Bank, 125 London Wall, London EC2Y 5AJ, United Kingdom*

ARTICLE INFO

Article history:
Received July 2009
Received in revised form
June 2010
Accepted 18 June 2010

Keywords:
Goodness of fit
Mortality models
Standard normality

ABSTRACT

This study sets out a framework to evaluate the goodness of fit of stochastic mortality models and applies it to six different models estimated using English & Welsh male mortality data over ages 64–89 and years 1961–2007. The methodology exploits the structure of each model to obtain various residual series that are predicted to be iid standard normal under the null hypothesis of model adequacy. Goodness of fit can then be assessed using conventional tests of the predictions of iid standard normality. The models considered are: Lee and Carter's (1992) one-factor model, a version of Renshaw and Haberman's (2006) extension of the Lee–Carter model to allow for a cohort-effect, the age-period-cohort model, which is a simplified version of the Renshaw–Haberman model, the 2006 Cairns–Blake–Dowd two-factor model and two generalized versions of the latter that allow for a cohort-effect. For the data set considered, there are some notable differences amongst the different models, but none of the models performs well in all tests and no model clearly dominates the others.

1. Introduction

In an earlier study, Cairns et al. (2009) examined eight different stochastic mortality models. The models were estimated on both English & Welsh and US male mortality data (over ages 60–89) and were assessed for their ability to explain *historical* patterns of mortality using both qualitative and quantitative criteria; the latter consisted primarily of Bayesian Information Criterion (BIC) rankings complemented by nesting tests in the cases where one model is a special case of another.

The present study builds on this work in proposing a more complete and systematic methodology for establishing the quantitative goodness of fit (GOF) of six of the above models based on formal hypothesis testing[1]:

- the one-factor Lee–Carter model (Lee and Carter, 1992), denoted M1 in Cairns et al. (2009)
- Renshaw and Haberman's generalization of the Lee–Carter model to incorporate a cohort-effect (Renshaw and Haberman, 2006), denoted M2
- the age-period-cohort (APC) model which is a simplification of the Renshaw–Haberman model (Currie, 2006) (see, also Osmond, 1985; Jacobsen et al., 2002), denoted M3
- the two-factor Cairns–Blake–Dowd (CBD) model of Cairns et al. (2006a), denoted M5
- two different generalizations of the CBD model incorporating a cohort-effect, denoted M6 and M7.

More specifically, we use what we know about the structure of each model to construct the following series that are predicted to be (at least approximately) independently and identically distributed standard normal (hereafter abbreviated to 'iid N(0, 1)') under the null hypothesis:

- Standardized mortality rate residuals or *mortality residuals* for short. The mortality residuals are the differences between the realized (or actual) mortality rates for any given set of ages and years and their model-generated equivalents (i.e., fitted values). Once standardized, these are predicted to be approximately iid N(0, 1) under the null hypothesis.
- Standardized residuals of the model's unobservable state variables (SVs) or *SV residuals* for short. The SVs are the stochastic factors driving the dynamics of the model, and, once standardized, are also assumed to be approximately iid N(0, 1).
- Standardized residuals for the prices (or fair values) of mortality-dependent financial instruments derived from the model (or *price residuals* for short), where the residuals concerned are the differences between these prices and their model-based equivalents, and these too should be approximately iid N(0, 1) under the null hypothesis.

* Corresponding author.
E-mail addresses: Kevin.Dowd@hotmail.co.uk (K. Dowd), a.cairns@ma.hw.ac.uk (A.J.G. Cairns), d.blake@city.ac.uk (D. Blake), guy.coughlan@jpmorgan.com (G.D. Coughlan), david.epstein@jpmorgan.com (D. Epstein), marwa.khalafallah@jpmorgan.com (M. Khalaf-Allah).

[1] The reason for excluding two of the eight models is explained in Section 7 below.

0167-6687/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.insmatheco.2010.06.006

Each model was estimated using LifeMetrics data for the mortality rates of English & Welsh males[2] for ages from 64 to 89 and spanning the years 1961 to 2007.[3] As such, the results presented herein are not necessarily representative of what might be obtained for other data sets. They do, however, serve to illustrate both the methodology and the potential weaknesses in certain stochastic mortality models.

The paper is organized as follows. Section 2 explains our notation and Section 3 outlines the models to be considered. Section 4 outlines and implements the testing framework for the models' mortality residuals, while Section 5 does the same for each model's SV residuals. Section 6 provides some test results for the price of an illustrative mortality-dependent financial contract, namely a period term annuity. Section 7 presents two comparisons: a comparison with the findings of our own earlier studies and a comparison with some recent studies by other researchers testing the out-of-sample performance of stochastic mortality models. Section 8 concludes.

2. Notation

We begin with some notation, and distinguish between the following mortality rates:

- $q(t,x)$ = true (and unobserved) mortality rate, i.e., the probability of death between times t and $t+1$ for individuals aged x at time t;
- $\tilde{q}(t,x)$ = crude estimate of year-t mortality rate based on observed deaths and exposures data;
- $\bar{q}(t,x)$ = estimated year-t mortality rate based on data up to and including year t, and using a specified mortality model (i.e., the fitted value from the model);
- $\tilde{m}(t,x)$ = crude estimate of year-t death rate (i.e. the observed number of deaths divided by the average population size aged x last birthday during year t).

The crude mortality rate $\tilde{q}(t,x)$ is linked to the crude death rate, $\tilde{m}(t,x)$, via $\tilde{q}(t,x) = 1-\exp\left(-\tilde{m}(t,x)\right)$.

The models that we consider involve the following SVs:

- $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$ are the true (unobserved) age, period and cohort-effects, given that the relevant specified model is true;
- $\bar{\beta}_x^{(i)}$, $\bar{\kappa}_t^{(i)}$ and $\bar{\gamma}_c^{(i)}$ are their estimates, given data from years t_0 to t_1 and ages x_0 to x_1, and which are used to calculate the $\bar{q}(t,x)$;
- $\hat{\beta}_x^{(i)}$, $\hat{\kappa}_t^{(i)}$ and $\hat{\gamma}_c^{(i)}$ are their one-step ahead forecasts given data from years t_0 to t_1-1 and ages x_0 to x_1.

The cohort-effects are estimated for years of birth c_0 to c_1, where the year of birth is equal to $c = t - x$.

3. The stochastic mortality models

The models examined in this study are the following:

Model M1

Model M1, the original Lee–Carter model, postulates that the true underlying death rate, $m(t,x)$, satisfies the following equation:

$$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} \tag{1}$$

where the state variable $\kappa_t^{(2)}$ follows a one-dimensional random walk with drift (Lee and Carter, 1992):

$$\kappa_t^{(2)} = \kappa_{t-1}^{(2)} + \mu + CZ_t^{(2)} \tag{2}$$

in which μ is a constant drift term, C is a constant volatility and $Z_t^{(2)}$ is a one-dimensional iid N(0, 1) error.

Model M2B[4]

This model, which is a particular extension of the Lee–Carter model to allow for a cohort-effect, postulates that $m(t,x)$ satisfies:

$$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_c^{(3)} \tag{3}$$

where the state variable $\kappa_t^{(2)}$ follows (2) and $\gamma_c^{(3)}$ is a cohort-effect. We follow Cairns et al. (2010) and CMI (2007) and model the cohort-effect, $\gamma_c^{(3)}$, as an ARIMA(1,1,0) process that is independent of $\kappa_t^{(2)}$:

$$\Delta\gamma_c^{(3)} = \mu_\gamma + \alpha_\gamma\left(\Delta\gamma_{c-1}^{(3)} - \mu_\gamma\right) + \sigma_\gamma Z_c^{(\gamma)}. \tag{4}$$

Model M3B[5]

This model is a simplified version of M2B and postulates that $m(t,x)$ satisfies:

$$\log m(t,x) = \beta_x^{(1)} + \kappa_t^{(2)} + \gamma_c^{(3)} \tag{5}$$

where the variables (including the cohort-effect) are the same as for M2B.

Model M5

M5 is a reparameterized version of the CBD two-factor mortality model (Cairns et al., 2006a). This model postulates that the mortality rate $q(t,x)$ satisfies:

$$\operatorname{logit} q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) \tag{6}$$

where $q(t,x) = 1-\exp(-m(t,x))$ and $\bar{x}$ is the average of the ages used in the dataset, and where the state variables now follow a two-dimensional random walk with drift:

$$\kappa_t = \kappa_{t-1} + \mu + CZ_t \tag{7}$$

where μ is a constant 2×1 drift vector, C is now a constant 2×2 upper triangular 'volatility' matrix (or, more precisely, the Choleski 'square root' matrix of the variance–covariance matrix), and Z_t is a two-dimensional standard normal variable, each component of which is independent of the other.[6]

Model M6

M6 is a generalized version of M5 with a cohort-effect, i.e.,

$$\operatorname{logit} q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \gamma_c^{(3)} \tag{8}$$

where the κ_t process follows (7) and the $\gamma_c^{(3)}$ process follows (4).

Model M7

Our last model, M7, is another generalized version of M5 with a cohort-effect, i.e.,

$$\operatorname{logit} q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \kappa_t^{(3)}((x-\bar{x})^2 - \sigma_x^2) + \gamma_c^{(4)} \tag{9}$$

where the state variables κ_t in this case follow a three-dimensional random walk with drift, σ_x^2 is the variance of the age range used in the dataset, and $\gamma_c^{(4)}$ is a cohort-effect that is modelled as an AR(1) process.[7]

[2] See Coughlan et al. (2007) and www.lifemetrics.com for the data and a description of LifeMetrics. The original source of the data was the UK Office for National Statistics.

[3] The under-64s were excluded because it is the mortality rates of older people that are of the greatest financial significance to pension funds and annuity providers – and this is our main interest in conducting this series of studies on stochastic mortality models – and the mortality rates of those over age 89 were excluded because of poor data reliability. We would also emphasise that models M5–M7 were specifically designed for the higher age ranges, whereas the other models considered in this study were designed to fit younger ages as well.

[4] M2B is a version of M2 that assumes an ARIMA(1,1,0) process for the cohort-effect.

[5] M3B is also a version of M3 that assumes an ARIMA(1,1,0) process for the cohort-effect.

[6] The reparameterization of the original model is $\kappa_t^{(2)} = A_2(t)$ and $\kappa_t^{(1)} - \kappa_t^{(2)}\bar{x} = A_1(t)$, where $A_1(t)$ and $A_2(t)$ are the state variables of the original model. An additional difference between the original CBD model and the reparameterized version M5 is that x in M5 refers to age at time t, whereas in the original CBD model it refers to age at some initial time 0.

[7] The generalization, therefore, incorporates an additional quadratic age effect as well as a cohort-effect.

4. Assessing the goodness of fit of the mortality residuals

Assessing goodness of fit involves three stages: estimation, implementation and testing.

4.1. Estimation

We start by selecting a lookback window on which to base our initial estimates. We choose a rolling 20-year window comprising the current and previous 19 years' historical observations.[8,9] We also need a suitable age range on which to fit the model, and we choose the age range 64–89.

For each model, we then estimate the parameters and obtain estimates of the unobserved SVs $\bar{\beta}_x^{(i)}$, $\bar{\kappa}_t^{(i)}$ and $\bar{\gamma}_c^{(i)}$ (as appropriate) and obtain model-based estimates of the mortality rate $\bar{q}(t, x)$. In the present context, the sequence of 20-year rolling windows gives us estimates for 27 years between 1981 and 2007.

The mortality residual is calculated as the difference between $\tilde{q}(t, x)$ and $\bar{q}(t, x)$. If the underlying random variable, the number of deaths, follows the assumption of a Poisson distribution (as, for example, assumed by Brouhns et al., 2002, and Li et al., 2009), then the distribution of deaths can be approximated by a normal distribution as the population size and the number of deaths gets 'large', as seems reasonable when we consider the size of the male population of England & Wales. If a model's estimates are adequate, the mortality residuals should also be approximately normal. The standardized mortality residuals – found by subtracting the residual mean and dividing the result by the residual standard deviation – are then predicted to be approximately iid N(0, 1).[10]

By way of example, and to make our discussion of estimation issues more concrete, consider the case of model M1 (whose structure is set out in Eqs. (1) and (2) above):

1. We first take the exposures and deaths data from 1961 to 1981 and fit the model to obtain estimates for the age effects $\bar{\beta}_x^{(1)}$ and $\bar{\beta}_x^{(2)}$ and the period-effect $\bar{\kappa}_t^{(2)}$ (see Eq. (1) above).
2. We then insert these into (1) to obtain the model-based death rate, $\bar{m}(t, x)$, and thence the model-based mortality rate, $\bar{q}(t, x)$, and the mortality residual $\tilde{q}\,(t, x) - \bar{q}\,(t, x)$ for 1981.
3. We repeat this process using data for 1962–1982 to get the mortality residual for 1982; we repeat again using data for 1963–1983 to obtain the 1983 mortality residual, and carry on in the same manner until we use data for 1987–2007 to obtain the 2007 mortality residual.

The other models are estimated in comparable ways.

4.2. Implementation

Let $D(t, x)$ be the number of deaths between t and $t + 1$ at age x last birthday, and let $E(t, x)$ be the corresponding exposures. From these, we calculate the crude death rates $\tilde{m}(t, x) = D(t, x)/E(t, x)$. Given the Poisson assumption about deaths and given that the expected number of deaths is large, the number of deaths is approximately normal with mean and variance both equal to $\bar{m}(t, x)E(t, x)$. It follows that for each model, the standardized mortality residuals

$$\varepsilon(t, x) = \frac{\tilde{m}(t, x) - \bar{m}(t, x)}{\sqrt{\bar{m}(t, x)/E(t, x)}} \tag{10}$$

should be approximately iid N(0, 1) under the null hypothesis.[11,12] Moreover, we would expect this prediction to hold both when we follow any given age from one year to the next and when we compare the death rates for different ages during the same year. Thus, the matrix of $\varepsilon(t, x)$ terms should be approximately iid N(0, 1) in both dimensions.

We then have $26 \times 27 = 702$ observations in the $\varepsilon(t, x)$ matrix (i.e., we have observations for each of 26 different ages spanning 64–89, over 27 different years spanning 1981–2007).

4.3. Test results

The hypothesis tests used in this section aim to identify whether the mortality residuals described above are consistent with iid N(0, 1) as predicted under the null hypothesis. We then carry out the following tests on the matrix of mortality residuals:

- A t-test of the prediction that their mean should be 0;
- A variance ratio (VR) test of the prediction that the variance should be 1 (see Cochrane, 1988; Lo and MacKinley, 1988, 1989); and
- A Jarque–Bera normality test based on the skewness and kurtosis predictions (see Jarque and Bera, 1980).

In addition, we also test the prediction that the residuals have zero correlation both across adjacent ages and across adjacent years. These tests are based on the test statistic $\rho\sqrt{N-2}/(1-\rho^2)$, where ρ is the relevant correlation coefficient, which is distributed under the null hypothesis as a t-distribution with $N - 2$ degrees of freedom. Note that we have 26 cross-age correlations (that between ages 64 and 65, that between ages 65 and 66, and so on) and 27 cross-year correlations (that between 1981 and 1982, that between 1982 and 1983, and so on).

Our test results are presented in Table 1. The upper section of this Table shows the sample moments and size. The middle section shows the p-values associated with mean, variance and normality predictions. The third shows the percentages of cross-age and cross-year correlation test results that are significant at the 1% level. If the null hypothesis of zero correlation held in each case, then we would expect these percentages to be 'close' to 1%.

The results in Table 1 suggest that the models perform quite poorly: the normality prediction is always decisively rejected and, with the exception of M2B, so too are the variance predictions. The correlation predictions are also rejected more frequently than they should be under the null, but there are notable differences: M7 and M2B perform best on this test and M1 and M3B worst.

[8] We chose a 20-year lookback window for estimating the models as a compromise between having a longer lookback which would increase the efficiency of the estimated parameters and a shorter lookback which would reduce any potential bias in the parameter estimates that would arise if the mortality data used for estimation incorporated one or more breaks in trend. Booth et al. (2002a) favour using a lookback window that extends back to the most recent break in trend, while Hyndman and Ullah (2007, p. 4953) recognize that there is a case for modifying the lookback window "due to the presence of substantial outliers in the fitting period". We experimented with both 10-year and 20-year lookback windows and concluded that a 20-year lookback window provided the best compromise.

[9] We could also have a chosen a window that expands over time to take account of the fact that our data accumulate over time. Having started with 20 observations to obtain our estimates for 1981, we might have used 21 observations to obtain estimates for 1982, and so forth. However, an expanding window would complicate the underlying statistics. A rolling fixed-length window is more straightforward to deal with.

[10] For convenience, we use the term 'tested for iid N(0, 1)' as shorthand for 'tested for the predictions of iid N(0, 1)', where these predictions are those of a zero mean, a unit variance, a zero skewness, a kurtosis equal to 3, and, of course, independent and identically distributed.

[11] We say 'approximately', in part, because we are using estimates of the SVs rather than their true values, in part, because there are likely to be measurement errors in the data (e.g., estimates of exposures are likely to be subject to errors) and, in part, because the assumed Poisson process with a fixed 'arrival' or mortality rate at any point in time is likely to be an over-simplification of reality.

[12] The reader will also note that (10) strictly refers to death-rate rather than mortality-rate residuals. However, the former will have the same distribution as the latter, so, for expositional purposes, it is convenient to ignore the difference between them.

Table 1
Test results for standardized mortality residuals $\varepsilon(t, x)$: five stochastic mortality models.

	M1	M2B	M3B	M5	M6	M7
Sample moments						
Mean	−0.0315	−0.0094	−0.0014	−0.0801	0.0221	−0.0084
Variance	3.5194	0.9286	3.1179	3.7529	2.3690	1.9829
Skewness	−1.0394	−0.4919	−0.6346	−1.0394	−1.0394	−1.0394
Kurtosis	9.2363	4.8350	6.7453	9.2363	9.2363	9.2363
N	702	702	702	702	702	702
P-values of sample moments						
Mean	0.6566	0.7962	0.9828	0.2736	0.7036	0.8741
Variance	0.0000	0.1764	0.0000	0.0000	0.0000	0.0000
Normality	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Percentages of correlation results significant at 1% level						
By adjacent ages	30.8	3.8	26.9	11.5	15.4	0.0
By adjacent years	22.2	7.4	37.0	22.2	3.7	7.4

Notes: Based on 27 annual observations spanning 1981–2007 for ages 64–89.

5. Assessing the goodness of fit of the state variable residuals

5.1. Estimation

The derivation of the test results for the SV residuals is complicated by the fact that the SVs are unobservable. We therefore need to obtain estimates of the unobserved state variables ($\bar{\kappa}_t^{(i)}$ and $\bar{\gamma}_c^{(i)}$) using 20 years of data up to and including year t. If we had direct observations of the state variables ($\tilde{\kappa}_t^{(i)}$ and $\tilde{\gamma}_c^{(i)}$) in the same way that we have direct observations of the mortality rates, $\tilde{q}(t, x)$, we could have proceeded in the same way as in the previous section: we would have obtained the period-effect residuals as $\tilde{\kappa}_t^{(i)} - \kappa_t^{(i)}$ and the cohort-effect residuals as $\tilde{\gamma}_c^{(i)} - \bar{\gamma}_c^{(i)}$. However, this is not possible because $\tilde{\kappa}_t^{(i)}$ and $\tilde{\gamma}_c^{(i)}$ are not directly observable. We therefore need proxies for these observations, and we obtain these proxies using 1-step ahead forecasts based on a model estimated using 20 years of data up to and including year $t - 1$. If we denote these forecasts by $\hat{\kappa}_t^{(i)}$ and $\hat{\gamma}_c^{(i)}$, the estimated period-effect residuals become $\hat{\kappa}_t^{(i)} - \bar{\kappa}_t^{(i)}$ and the estimated cohort-effect residuals become $\hat{\gamma}_c^{(i)} - \bar{\gamma}_c^{(i)}$. We now standardize each of these series by subtracting its estimated mean and dividing the result by its estimated one-period-ahead standard deviation. The resulting standardized SV residual series are then each predicted to be approximately iid N(0, 1) under the null hypothesis.

For each model, we have one or more sets of standardized SV residuals. The number of standardized SV residual series depends on the model – it is equal to the number of period-effects (which varies from 1 to 3) and the number of cohort-effects (which is either 0 or 1) in each model. The number of standardized SV residual series in each model therefore varies from 1 to 4.

As an aside, the fact that the model is re-estimated for each year in our sample period means that we are working with estimates for μ and C that are regularly updated. Accordingly, in the discussion below, we let $\bar{\mu}_t$ and $\bar{C}_t$ denote their estimates based on data up to and including year t.

5.2. Implementation

We now consider each model in turn.

Model M1

For M1, we use (2) to obtain estimated values of $\kappa_t^{(2)}$ (i.e., $\bar{\kappa}_t^{(2)}$) and 1-step ahead forecasts of $\kappa_t^{(2)}$ (i.e., $\hat{\kappa}_t^{(2)}$), viz.: [13]

$$\bar{\kappa}_t^{(2)} = \bar{\kappa}_{t-1}^{(2)} + \bar{\mu}_{t-1} + \bar{C}_{t-1}\bar{Z}_t^{(2)} \tag{11}$$

$$\hat{\kappa}_t^{(2)} = \bar{\kappa}_{t-1}^{(2)} + \bar{\mu}_{t-1}. \tag{12}$$

Substituting (12) into (11) and rearranging gives the standardized SV residuals:

$$\bar{Z}_t^{(2)} = \bar{C}_{t-1}^{-1}(\bar{\kappa}_t^{(2)} - \hat{\kappa}_t^{(2)}). \tag{13}$$

In (13), $\bar{\kappa}_t^{(2)}$ is the estimated value of $\kappa_t^{(2)}$ based on data from $t - 20$ up to and including time t, and $\hat{\kappa}_t^{(2)}$ is the 1-step ahead forecasted value of $\kappa_t^{(2)}$ based on data from $t - 20$ up to and including time $t - 1$. This gives us 27 values of $\bar{Z}_t^{(2)}$ and, under the null hypothesis, these are predicted to be iid N(0, 1).

Model M2B

For M2B, we obtain the standardized SV residuals $\bar{Z}_t^{(2)}$ using (13), and we model the cohort-effect $\gamma_c^{(3)}$ and recover the standardized cohort-effect residuals $\bar{Z}_c^{(\gamma)}$ using (4). Both standardized residual series $\bar{Z}_t^{(2)}$ and $\bar{Z}_c^{(\gamma)}$ are predicted to be iid N(0, 1).

We can also test the properties of both sets of estimated residuals simultaneously. Since $\bar{Z}_t^{(2)}$ and $\bar{Z}_c^{(\gamma)}$ should each be iid N(0, 1) and independent of each other, statistical theory tells us that the sum of squares of 2 independent N(0, 1) variates is distributed as a chi-squared with 2 degrees of freedom. It therefore follows that:

$$\bar{Y}_t = [\bar{Z}_t^{(2)}]^2 + [\bar{Z}_c^{(\gamma)}]^2 \sim \chi_2^2$$
$$\bar{p}_t = F\left(\bar{Y}_t\right) \sim \text{ iid U(0, 1)} \tag{14}$$

where $F(.)$ is the distribution function for a chi-squared with 2 degrees of freedom. Under the null, the series $\bar{p}_t$ should be distributed as iid standard uniform (or iid U(0, 1)). If we wished to, we could then test this series using standard uniformity tests such as Kolmogorov–Smirnov, Kuiper, Lilliefors, etc.[14] However, testing is easier (and we have more tests available) if we put $\bar{p}_t$ through the following transformation:

$$\bar{h}_t = \Phi^{-1}(\bar{p}_t) \sim \text{ iid N(0, 1)} \tag{15}$$

where $\Phi(.)$ is the distribution function for a standard normal variable. This transformation gives us an 'observed' series $\bar{h}_t$ that is distributed as iid N(0, 1) under the null. We can then test whether $\bar{h}_t$ is iid N(0, 1).

[13] When we use the 20-year window to obtain the $\hat{\kappa}_t^{(2)}$ forecasts, we need to ensure that any constraints in the estimation process are used in a fashion consistent with the way in which the $\bar{\kappa}_t^{(2)}$ estimates were obtained. Thus, for M1, we use the constraints $\sum_{t=1961}^{1980} \kappa_t^{(2)} = 0$ and $\sum_{x=x_0}^{x_1} \beta_x^{(2)} = 1$ for both $\bar{\kappa}_t^{(2)}$ and $\hat{\kappa}_t^{(2)}$.

[14] For more on these tests, see, e.g., Dowd (2005, chapter 15 appendix).

Model M3B

The standardized SV residuals for M3B are obtained in exactly the same way as for M2B.

Model M5

For model M5, we use (7) to obtain the 2 × 1 vector $\bar{\kappa}_t$ and the 1-step ahead forecasts $\hat{\kappa}_t$:

$$\bar{\kappa}_t = \bar{\kappa}_{t-1} + \bar{\mu}_{t-1} + \bar{C}_{t-1}\bar{Z}_t \tag{16}$$

$$\hat{\kappa}_t = \bar{\kappa}_{t-1} + \bar{\mu}_{t-1} \tag{17}$$

$$\bar{Z}_t = \bar{C}_{t-1}^{-1}(\bar{\kappa}_t - \hat{\kappa}_t). \tag{18}$$

Under the null, each standardized SV residual series, $Z_t^{(1)}$ and $Z_t^{(2)}$, is iid N(0, 1) and independent of the other.

We now test $Z_t^{(1)}$ and $Z_t^{(2)}$ for iid standard normality using conventional tests, and additionally apply a standard correlation test to check the prediction that these have a zero correlation.

As with M2B and M3B, we can also test the properties of both sets of standardized residuals simultaneously. In this case, under the null hypothesis,

$$\bar{Y}_t = [\bar{Z}_t^{(1)}]^2 + [\bar{Z}_t^{(2)}]^2 \sim \chi_2^2$$

$$\bar{p}_t = F\left(\bar{Y}_t\right) \sim \text{ iid U}(0, 1) \tag{19}$$

$$\bar{h}_t = \Phi^{-1}\left(\bar{p}_t\right) \sim \text{ iid N}(0, 1). \tag{20}$$

We now test $\bar{h}_t$ for iid N(0, 1).

Model M6

Following the same logic, for M6 we obtain

$$\bar{Z}_t = \bar{C}_{t-1}^{-1}(\bar{\kappa}_t - \hat{\kappa}_t) \tag{21}$$

which gives us two sets of standardized SV residuals $Z_t^{(1)}$ and $Z_t^{(2)}$ that are predicted to be iid N(0, 1) and independent of each other. As with the previous model, we test $Z_t^{(1)}$ and $Z_t^{(2)}$ for iid zero correlation standard normality.

As with M2B and M3B, we also obtain the corresponding standardized cohort-effect residuals that are also predicted to be iid N(0, 1). It then follows that

$$\bar{Y}_t = [\bar{Z}_t^{(1)}]^2 + [\bar{Z}_t^{(2)}]^2 + [\bar{Z}_c^{\gamma}]^2 \sim \chi_3^2$$

$$\bar{p}_t = F\left(\bar{Y}_t\right) \sim \text{ iid U}(0, 1) \tag{22}$$

$$\bar{h}_t = \Phi^{-1}\left(\bar{p}_t\right) \sim \text{ iid N}(0, 1) \tag{23}$$

which we then test for iid N(0, 1).

Model M7

M7 is similar but involves three sets of standardized SV residuals, $Z_t^{(1)}$, $Z_t^{(2)}$ and $Z_t^{(3)}$, which are predicted to be iid N(0, 1) and to have zero correlations. M7 also involves standardized cohort-effect residuals $\bar{Z}_c^{(\gamma)}$.[15] Applying the same logic as before then gives us:

$$\bar{Y}_t = [Z_t^{(1)}]^2 + [Z_t^{(2)}]^2 + [Z_t^{(3)}]^2 + [Z_c^{(\gamma)}]^2 \sim \chi_4^2$$

$$\bar{p}_t = F\left(\bar{Y}_t\right) \sim \text{ iid U}(0, 1) \tag{24}$$

$$\bar{h}_t = \Phi^{-1}\left(\bar{p}_t\right) \sim \text{ iid N}(0, 1) \tag{25}$$

where $F(.)$ is now the distribution function for a chi-squared with 4 degrees of freedom. As in earlier cases, we then test $\bar{h}_t$ for iid N(0, 1).

Table 2
Results for the standardized residuals of the state variable $\bar{Z}_t^{(2)}$: model M1.

Sample moments	
Mean	−0.375
Variance	0.955
Skewness	0.033
Kurtosis	2.916
N	27
Test of mean prediction	
P-value mean *t*-test statistic	0.057
Test of variance ratio prediction	
P-value variance ratio test statistic	0.944
Test of normality prediction	
P-value Jarque–Bera test statistic	0.994
Test of temporal independence	
Pearson correlation $(t+1, t)$	−0.545
P-value correlation	0.001*

Notes: Based on 27 annual observations spanning 1981–2007 for ages 64–89. All tests are two-sided except for the JB test which is inherently one-sided. If ρ is the correlation coefficient, $\rho\sqrt{N-2}/(1-\rho^2)$ is distributed under the null as a t-distribution with $N-2$ degrees of freedom.
* Indicates significance at the 1% level.

5.2.1. Test results

Model M1 Table 2 presents the sample moments and the test results for M1's standardized SV residual series, $\bar{Z}_t^{(2)}$, and these results are compatible with the null hypothesis of standard normality. However, the null hypothesis of temporal independence is strongly rejected. Altogether, there are four p-values reported for M1, and, of these, one is significant at well under the 1% level. If we treat any p-values below 1% as a 'fail', then, by this criterion, M1 has a 'failure' rate of 25%.

Model M2B Table 3 presents the sample moments and test results for each of $\bar{Z}_t^{(2)}$ and $\bar{Z}_c^{(\gamma)}$ for model M2B. This model performs very poorly by these tests: both series score p-values of 0 for the variance and normality tests – and the sample moments of the $\bar{Z}_c^{(\gamma)}$ bear no resemblance to the predictions. Similarly, the $\bar{h}_t$ test results in Table 4 lead us to reject the null hypothesis that the standardized residuals are jointly iid N(0, 1).

Note that there are 12 p-values reported for M2B, and of these six are below 1%, implying a 'failure' rate of 50%.

To investigate further, Figs. 1 and 2 give the QQ plots[16] for the model's two standardized SV residual series, $\bar{Z}_t^{(2)}$ and $\bar{Z}_c^{(\gamma)}$, and Fig. 3 gives a plot of empirical vs. predicted $\bar{p}_t$. We can see that all three Figures show extremely poor fits: the two QQ plots have one or more very extreme outliers (especially for the cohort-effect plot in Fig. 8) and do not lie close to the 45° line; the $\bar{p}_t$ plot in Fig. 9 clearly does not lie anywhere close to its predicted 45° line either. There are therefore very clear problems with both this model's standardized residuals series.

[15] Note, however, that $\bar{Z}_c^{(\gamma)}$ now refers to the standardized residual of the $\gamma_c^{(4)}$ process rather than that of the $\gamma_c^{(3)}$. The context makes it clear which gamma process $\bar{Z}_c^{(\gamma)}$ is referring to.

[16] A QQ plot is a plot of the empirical quantiles of a distribution against their predicted counterparts, where the latter in this case are based on the prediction of standard normality. QQ plots give a useful visual indicator of whether the empirical quantiles are consistent with the predicted ones: under the null, we would expect the plots to lie fairly close to the 45° line. Note that we do not report the QQ and associated plots for models other than M2B, as these are all compatible with the underlying null hypotheses. The plots for M2B, on the other hand, are more informative.

Table 3

Results for the standardized residuals of the state variables $\bar{Z}_t^{(2)}$ and $\bar{Z}_c^{(\gamma)}$: model M2B.

Sample moments		
	$\bar{Z}_t^{(2)}$	$\bar{Z}_c^{(\gamma)}$
Mean	−0.198	−10.866
Variance	11.454	1001.433
Skewness	−1.768	−2.910
Kurtosis	13.068	10.666
N	27	27
Test of mean prediction[1]		
P-value mean t-test statistic	0.764	0.086
Test of variance ratio prediction		
P-value variance ratio test statistic	0.000*	0.000*
Test of normality prediction		
P-value Jarque–Bera test statistic	0.000*	0.000*
Test of temporal independence[2]		
Pearson correlation $(t+1, t)$	0.029	0.637
P-value correlation	0.886	0.000*

Notes: As per notes to Table 2.

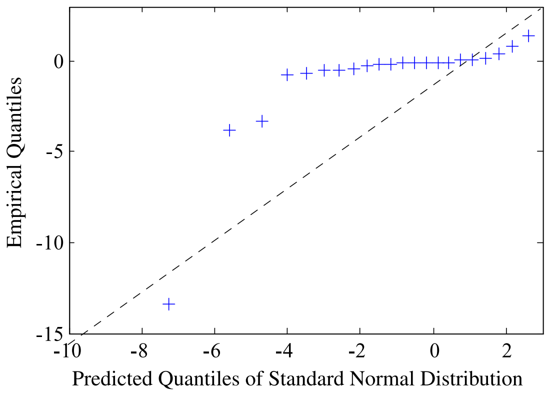

Fig. 1. QQ plot for $\bar{Z}_t^{(2)}$: model M2B. Note: Based on 27 annual $\bar{Z}_t^{(2)}$ observations of model M2B over the period 1981–2007 and ages 64–89.

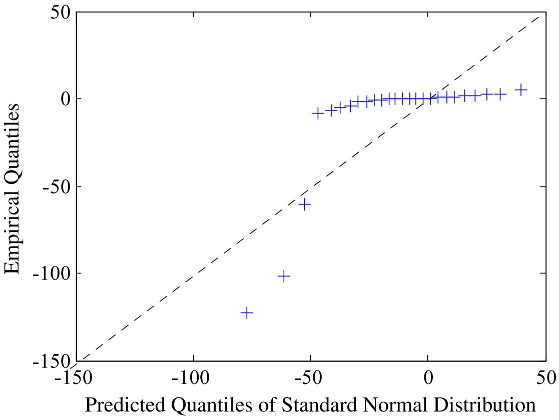

Fig. 2. QQ plot for $\bar{Z}_c^{(\gamma)}$: model M2B. Note: Based on 27 annual $\bar{Z}_c^{(\gamma)}$ observations of model M2B over the period 1981–2007 and ages 64–89.

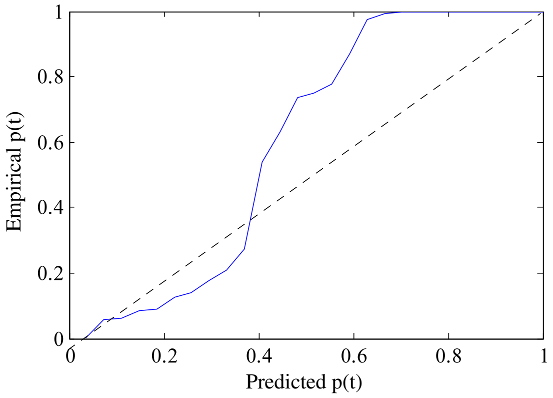

Fig. 3. Plot of empirical vs. predicted $\bar{p}_t$: model M2B. Note: Based on 27 annual $\bar{p}_t$ observations of model M2B over 1981–2007 and ages 64–89. $\bar{p}_t = F\left([\bar{Z}_t^{(1)}]^2 + [\bar{Z}_c^{(\gamma)}]^2\right)$, where $F(.)$ is χ_2^2.

Table 4

Results for the predicted standard normal variate $\bar{h}_t$: model M2B.

Sample moments	
Mean	1.056
Variance	4.934
Skewness	0.055
Kurtosis	1.435
N	27
Test of mean prediction	
P-value mean t-test statistic	0.020
Test of variance ratio prediction	
P-value variance ratio test statistic	0.000*
Test of normality prediction	
P-value Jarque–Bera test statistic	0.250
Test of temporal independence	
Pearson correlation $(t+1, t)$	0.142
P-value correlation	0.477

Notes: $\bar{h}_t = \Phi^{-1}(\bar{p}_t)$, where $\bar{p}_t = F\left([\bar{Z}_t^{(2)}]^2 + [\bar{Z}_c^{(\gamma)}]^2\right)$, $F(.)$ is the χ_2^2 distribution function, and $\Phi(.)$ is the standard normal distribution function. Note, however, that in 10 cases, the estimated value of $\bar{h}_t$ was 1. Since the normal inverse of 1 is undefined, these values were reduced to 0.9999 for the purposes of computing the results in this Table. Otherwise as per notes to Table 2.

It is worth pausing for a moment to consider why M2B produces such poor results. If the model and fitting procedure were robust, then adding in one year's data should only have a small impact on the estimated age, period and cohort-effects. However, it was found with M2B – but not with any of the other models considered in this study – that adding one extra year of data could lead the model to jump from one set of fitted values for the cohort-effect to a completely different set.[17] This problem is most likely explained by the likelihood function having multiple maxima. The changes in parameter values then reflect a jump in the fitting algorithm from one maximum to another.[18]

Model M3B

Table 5 presents the moments and test results for the standardized residuals for M3B. As with M2B, we have 12 reported p-values, but in this case only three are significant at the 1% level. M3B therefore has a 'failure' rate of 25%.

[17] These claims are borne out by graphs of fitted parameter values (not included here), which show considerable instability for M2B. By contrast, graphs of the fitted parameter values for other models are all stable. For further discussion of the stability problem, see Cairns et al. (2009). The authors of CMI Working Paper 25 encountered similar problems. To quote from their study: "the fitted cohort parameters do not appear to be stable as the age range fitted is changed" (CMI, 2007, p. 18, para 7.18); "when backtesting a dataset or fitting a different age range, we were unable to find a set of starting parameter values that consistently worked for different subsets of the data. Where a number of sets of starting parameter values worked for a particular dataset, we also found that the fitted values could differ materially" (CMI, 2007, p. 19, para 7.21).

[18] These jumps, in turn, lead to the fitted standardized residuals having some very extreme values as shown in Figs. 1–3 and Tables 3 and 4.

Table 5
Results for the standardized residuals of the state variables $\bar{Z}_t^{(2)}$ and $\bar{Z}_c^{(\gamma)}$: model M3B.

Sample moments		
	$\bar{Z}_t^{(2)}$	$\bar{Z}_c^{(\gamma)}$
Mean	0.139	−0.189
Variance	0.798	2.201
Skewness	0.179	−0.085
Kurtosis	2.821	4.050
N	27	27
Test of mean prediction[1]		
P-value mean *t*-test statistic	0.426	0.513
Test of variance ratio prediction		
P-value variance ratio test statistic	0.490	0.001*
Test of normality prediction		
P-value Jarque–Bera test statistic	0.914	0.529
Test of temporal independence[2]		
Pearson correlation $(t+1, t)$	−0.608	−0.055
P-value correlation	0.000*	0.785

Notes: As per notes to Table 2.

Table 6
Results for the predicted standard normal variate $\bar{h}_t$: model M3B.

Sample moments	
Mean	0.168
Variance	2.030
Skewness	0.408
Kurtosis	3.251
N	27
Test of mean prediction	
P-value mean *t*-test statistic	0.546
Test of variance ratio prediction	
P-value variance ratio test statistic	0.003*
Test of normality prediction	
P-value Jarque–Bera test statistic	0.664
Test of temporal independence	
Pearson correlation $(t+1, t)$	0.000
P-value correlation	0.998

Notes: $\bar{h}_t = \Phi^{-1}(\bar{p}_t)$, where $\bar{p}_t = F\left([\bar{Z}_t^{(2)}]^2 + [\bar{Z}_c^{(\gamma)}]^2\right)$, $F(.)$ is the χ_2^2 distribution function, and $\Phi(.)$ is the standard normal distribution function.
* Indicates significance at the 1% level.

Table 7
Results for the standardized residuals of the state variables $\bar{Z}_t^{(1)}$ and $\bar{Z}_t^{(2)}$: model M5.

Sample moments		
	$\bar{Z}_t^{(1)}$	$\bar{Z}_t^{(2)}$
Mean	−0.337	0.555
Variance	0.720	1.301
Skewness	0.163	−0.036
Kurtosis	3.039	2.257
N	27	27
Test of mean prediction		
P-value mean *t*-test statistic	0.049	0.018
Test of variance ratio prediction		
P-value variance ratio test statistic	0.305	0.278
Test of normality prediction		
P-value Jarque–Bera test statistic	0.941	0.731
Test of temporal independence		
Pearson correlation $(t+1, t)$	−0.539	0.124
P-value correlation	0.001*	0.533
Correlation between $\bar{Z}_t^{(1)}$ and $\bar{Z}_t^{(2)2}$		
Pearson correlation	−0.028	
P-value correlation	0.890	

Notes: As per notes to Table 2.

Table 8
Results for the predicted standard normal variate $\bar{h}_t$: Model M5.

Sample moments	
Mean	0.083
Variance	1.422
Skewness	−0.135
Kurtosis	2.623
N	27
Test of mean prediction	
P-value mean *t*-test statistic	0.721
Test of variance ratio prediction	
P-value variance ratio test statistic	0.150
Test of normality prediction	
P-value Jarque–Bera test statistic	0.886
Test of temporal independence	
Pearson correlation $(t+1, t)$	0.114
P-value correlation	0.570

Notes: As per notes to Table 6.

Model M5

Table 7 presents the sample moments and the test results for $\bar{Z}_t^{(1)}$ and $\bar{Z}_t^{(2)}$ based on M5, and Table 8 presents the sample moments and test results for M5's $\bar{h}_t$ series. M5 has 13 *p*-values of which only 1 is significant at the 1% level: M5 therefore has a 'failure rate' of 7.7%.

Model M6

Tables 9 and 10 present the comparable results for M6. This model has 17 *p*-values of which two are significant at the 1% level: M6 therefore has a 'failure rate' equal to 11.7%.

Model M7

Tables 11 and 12 present the corresponding results for M7. For this model we have 23 *p*-values, of which 2 are significant. Hence, M7 has a failure rate equal to 8.7%.

5.3. Summary of Section 5 results

The results of applying the state variable GOF tests to the six models are summarized in Table 13, which shows the proportions of test results for each model that are significant at the 1% level. It also shows the implied ranking by this criterion: M5 comes a little ahead of M7, which in turn comes a little ahead of M6. M1 and M3B then follow as equal second to last, and M3B comes last.

6. Assessing the goodness of fit of model-based annuity price residuals

Our final test of the adequacy of the models is to test the goodness of fit of the prices (or fair values) of financial assets that depend on model-based mortality forecasts. To illustrate, we consider the case of a period term annuity for males aged 65, payable until age 90.[19] We will assume the cashflows on

[19] A period term annuity is one that has a fixed term and ignores future mortality improvements. That is, for valuation purposes the annuity's future cash flows are calculated purely from the latest period mortality rates. We consider term annuities ceasing at age 90 because models M1, M2B and M3B, having been fitted to data from ages 60 to 89, apply to mortality rates from ages 60 to 89 only. Their semi-parametric structure means that there is no natural way to use them to extrapolate mortality rates beyond age 89.

Table 9
Results for the standardized residuals of the state variables $\bar{Z}_t^{(1)}$, $\bar{Z}_t^{(2)}$ and $\bar{Z}_c^{(\gamma)}$: model M6.

Sample moments			
	$\bar{Z}_t^{(1)}$	$\bar{Z}_t^{(2)}$	$\bar{Z}_c^{(\gamma)}$
Mean	−0.121	0.516	−0.393
Variance	0.755	1.090	2.572
Skewness	0.359	0.361	−0.134
Kurtosis	3.225	2.229	5.549
N	27	27	27
Test of mean prediction[1]			
P-value mean t-test statistic	0.475	0.016	0.214
Test of variance ratio prediction			
P-value variance ratio test statistic	0.383	0.684	0.000*
Test of normality prediction			
P-value Jarque–Bera test statistic	0.728	0.533	0.025
Test of temporal independence[2]			
Pearson correlation $(t+1, t)$	−0.503	−0.138	−0.059
P-value correlation	0.002*	0.487	0.770
Correlation between $\bar{Z}_t^{(1)}$ and $\bar{Z}_t^{(2)}$		−0.090	
P-value of correlation between $\bar{Z}_t^{(1)}$ and $\bar{Z}_t^{(2)}$		0.652	

Notes: As per notes to Table 2.

Table 10
Results for the predicted standard normal variate $\bar{h}_t$: model M6.

Sample moments	
Mean	0.292
Variance	1.336
Skewness	−0.196
Kurtosis	2.092
N	27
Test of mean prediction	
P-value mean t-test statistic	0.201
Test of variance ratio prediction	
P-value variance ratio test statistic	0.235
Test of normality prediction	
P-value Jarque–Bera test statistic	0.577
Test of temporal independence	
Pearson correlation $(t+1, t)$	−0.073
P-value correlation	0.716

Notes: $\bar{h}_t = \Phi^{-1}(\bar{p}_t)$, where $\bar{p}_t = F\left([\bar{Z}_t^{(1)}]^2 + [\bar{Z}_t^{(2)}]^2 + [\bar{Z}_c^{(\gamma)}]^2\right)$, $F(.)$ is the χ_2^2 distribution function, and $\Phi(.)$ is the standard normal distribution function. Note, however, that in 1 case, the estimated value of $\bar{h}_t$ was 1, which was reduced to 0.9999 for the purposes of computing the results in this Table. Otherwise as per notes to Table 6.

Table 11
Results for the standardized residuals of the state variables $\bar{Z}_t^{(1)}$, $\bar{Z}_t^{(2)}$, $\bar{Z}_t^{(3)}$ and $\bar{Z}_c^{(\gamma)}$: model M7.

Sample moments				
	$\bar{Z}_t^{(1)}$	$\bar{Z}_t^{(2)}$	$\bar{Z}_t^{(3)}$	$\bar{Z}_c^{(\gamma)}$
Mean	−0.330	0.302	0.007	0.098
Variance	0.771	0.863	1.601	2.554
Skewness	0.195	0.793	0.059	−0.126
Kurtosis	2.965	3.044	2.728	5.747
N	27	27	27	27
Test of mean prediction				
P-value mean t-test statistic	0.062	0.103	0.978	0.753
Test of variance ratio prediction				
P-value variance ratio test statistic	0.421	0.671	0.054	0.000*
Test of normality prediction				
P-value Jarque–Bera test statistic	0.917	0.243	0.952	0.014
Test of temporal independence				
Pearson correlation $(t+1, t)$	−0.600	−0.075	0.081	−0.332
P-value correlation	0.000*	0.708	0.687	0.074
Correlations				
	$\bar{Z}_t^{(1)}$	$\bar{Z}_t^{(2)}$	$\bar{Z}_t^{(3)}$	
$\bar{Z}_t^{(1)}$	1			
$\bar{Z}_t^{(2)}$	−0.232	1		
$\bar{Z}_t^{(3)}$	0.172	−0.131	1	
P-values of correlations[2]				
	$\bar{Z}_t^{(1)}$	$\bar{Z}_t^{(2)}$	$\bar{Z}_t^{(3)}$	
$\bar{Z}_t^{(1)}$	1			
$\bar{Z}_t^{(2)}$	0.230	1		
$\bar{Z}_t^{(3)}$	0.383	0.512	1	

Notes: As per notes to Table 2.

Table 12
Results for the predicted standard normal variate $\bar{h}_t$: model M7.

Sample moments	
Mean	0.397
Variance	1.858
Skewness	0.963
Kurtosis	3.926
N	27
Test of mean prediction	
P-value mean t-test statistic	0.142
Test of variance ratio prediction	
P-value variance ratio test statistic	0.010
Test of normality prediction	
P-value Jarque–Bera test statistic	0.077
Test of temporal independence	
Pearson correlation $(t+1, t)$	−0.177
P-value correlation	0.369

Notes: $\bar{h}_t = \Phi^{-1}(\bar{p}_t)$, where $\bar{p}_t = F\left([\bar{Z}_t^{(1)}]^2 + [\bar{Z}_t^{(2)}]^2 + [\bar{Z}_t^{(3)}]^2 + [\bar{Z}_c^{(\gamma)}]^2\right)$, $F(.)$ is the χ_2^2 distribution function, and $\Phi(.)$ is the standard normal distribution function. Otherwise as per notes to Table 6.

the annuity are discounted using a fixed discount rate of 4%. We adopt procedures similar to those employed for testing the goodness of fit of the state variables. Take the first 20-year window covering 1961–1980. For this period, each model is used to obtain estimates of the underlying state variables: $\bar{\beta}_x^{(i)}$, $\bar{\kappa}_t^{(i)}$ and $\bar{\gamma}_c^{(i)}$. We then generate 1000 one-period ahead simulations of $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$ (i.e., for 1981). For each simulation and each model, we generate model-based mortality rates, $q(t,x)$, for ages between 65 and 90, and the corresponding period annuity prices, $a(t,x)$.[20] The 1000 simulated values give us an estimate of the one-period-ahead forecast distribution of $a(t,x)$ for each model, and we use this to estimate the mean, $\bar{a}(t,x)$, and the corresponding standard deviation. We then use the crude mortality rates, $\tilde{q}(t,x)$, for 1981 to calculate the 'crude' period annuity price, $\tilde{a}(t,x)$. The annuity residual for each model is then $\tilde{a}(t,x) - \bar{a}(t,x)$ and this is standardized by dividing by the standard deviation of the one-period-ahead forecast distribution of the period annuity price for

[20] Period annuity prices are calculated as follows. We define, first, the model-simulated period survival function $S(t,x,y) = \{1-q(t,x)\} \times \{1-q(t,x+1)\} \times \cdots \times \{1-q(t,y-1)\}$. The simulated period annuity price is then defined as $a(t,x) = \sum_{y=x+1}^{90} S(t,x,y)(1+r)^{-(y-x)}$ where we assume $r = 0.04$. Crude period annuity prices, $\tilde{a}(t,x)$, are calculated in the same way, replacing $q(t,x)$ by $\tilde{q}(t,x)$.

Table 13
Summary of main standardized residual results for the state variables.

Model	Proportion of test results significant at the 1% level	Implied ranking
M1	25.0%	= 5
M2B	50.0%	6
M3B	25.0%	= 5
M5	7.7%	1
M6	10.5%	3
M7	8.7%	2

Notes: Based on the results in Tables 2–12.

that year. This procedure is repeated for the remaining 20-year windows covering 1962–1981, 1963–1982, etc.

The sample moments and moment-based test statistics for the standardized annuity residuals are given in Table 14, and the main highlights are:

- All models give fairly reasonable sample moments for the residuals.
- M2B, M3B, M5 and M6 each fail the iid test at the 1% significance level.
- M1 and M7 pass all tests at the 1% significance level.

These results suggest that there is little to choose between M1 and M7, and the others come afterwards with little to choose between them.

7. Comparisons

7.1. A comparison with the findings of our own earlier studies

The present paper is the fourth in a series of studies that we have conducted whose aim has been to examine different features of a set of stochastic mortality models with the ultimate objective of identifying which, if any, of these models might make suitable candidates for forecasting future mortality rates at high ages. In this section, we briefly compare and summarize the findings from these earlier studies.

The original study, Cairns et al. (2009), examined eight models, the six models considered here plus:

- the P-splines model (Currie et al., 2004; Currie, 2006; CMI, 2006), denoted M4
- a further generalization of the CBD model incorporating a cohort-effect, denoted M8.

The purpose of that study, as mentioned in the introduction, was to use a set of quantitative and qualitative criteria to assess each model's ability to explain *historical* patterns of mortality: quality of fit, as measured by the BIC; ease of implementation; parsimony; transparency; incorporation of cohort-effects; ability to produce a non-trivial correlation structure between ages; and robustness of parameter estimates relative to the period of data employed.

Using English & Welsh male mortality data, the BIC rankings were as follows: 1 = M8, 2 = M7, 3 = M2, 4 = M6, 5 = M3, 6 = M1, 7 = M4 and 8 = M5. We decided to drop M4 from further analysis, in part, because of its low ranking, but more importantly, because of its inability to produce fully-stochastic projections of future mortality rates. We then went on to obtain the following ranking of the remaining models on US male data: 1 = M2, 2 = M7, 3 = M3, 4 = M8, 5 = M6, 6 = M1, 7 = M5. M7 was found to have the most robust and stable parameter estimates over time on both data sets.

The second study, Cairns et al. (2010), focused on the qualitative forecasting properties of these models by evaluating the *ex-ante* plausibility of the models' probability density forecasts in terms of the following qualitative criteria (see also Cairns et al., 2006b): biological reasonableness; the plausibility of predicted levels of uncertainty in forecasts at different ages; and the robustness of the forecasts relative to the sample period used to fit the models. We found that while a good fit to historical data, as measured by the BIC, is a good starting point, it does not guarantee sensible forecasts. In particular, we found that M8 produced such implausible forecasts of US male mortality rates that it could be dismissed as a suitable forecasting model. M2 lacked robustness in its forecasts, while M1 produced forecasts at higher ages that were 'too precise', in the sense of having too little uncertainty relative to historical volatility.[21] The problems with these three models were not evident from simply estimating their parameters: they only became apparent when the models were used for forecasting. M3, M5 and M7 performed well, producing robust and biologically plausible forecasts.[22]

It is also important to examine the *ex post* forecasting performance of the models i.e., to backtest them. This is the subject of our third study, Dowd et al. (forthcoming). Backtesting is based on the idea that forecast distributions should be compared against subsequently realized mortality outcomes and if the realized outcomes are compatible with their forecasted distributions, then this would suggest that the models that generated them are good ones, and *vice versa*. That study discussed four different classes of backtest: those based on the convergence of forecasts through time towards the mortality rate(s) in a given year; those based on the accuracy of forecasts over multiple horizons; those based on the accuracy of forecasts over rolling fixed-length horizons; and those based on formal hypothesis tests that involve comparisons of realized outcomes against forecasts of the relevant densities over specified horizons. We found that models M1, M3B, M5, M6 and M7 perform well most of the time and there is relatively little to choose between them. Model M2B, by contrast, repeatedly showed evidence of instability.

7.2. A comparison with some recent studies testing the out-of-sample performance of stochastic mortality models

A number of other authors have, in recent years, also tackled the question of the forecasting accuracy of various stochastic mortality models.

Booth et al. (2006) consider five variants or extensions of the Lee–Carter model, M1. The models are fitted to both male and female data in 10 countries up to 1985, and then used to project the death rate, $m(t, x)$, and period life expectancy up to 2000. These projections are then compared to the actual death rates between 1986 and 2000 and the forecasting errors are combined in a variety of ways to assess the relative accuracy of the five models. This study, therefore, uses the same expanding horizon procedure (from a fixed starting point (1985)) as used in our backtesting study (Dowd et al., forthcoming). Although more formal statistical tests are also performed by Booth et al. (2006), it is unclear whether or not the assumptions underpinning these tests (such as independence of errors) have been verified, and this undermines the validity of the study's findings somewhat. In contrast, the present study focuses on a sequence of one-year-ahead forecasts allowing us to conduct a series of formal statistical tests in which the assumptions underlying the null hypothesis are known to be valid.

Huang et al. (2008) and Yang et al. (2010) develop a new approach to mortality forecasting using principal component

[21] This has also been noticed by other researchers, e.g., Li et al. (2009).

[22] M6 was dropped from this study because it was a special case of M7, and M7 was found to be stable and to deliver consistently better and more plausible results than M6.

Table 14
Sample moments and *P*-values for standardized annuity price residuals.

	M1	M2B	M3B	M5	M6	M7
Sample moments						
Mean	0.406	−0.272	0.328	0.350	0.151	0.397
Variance	0.889	0.753	0.708	0.680	0.747	1.858
Skewness	−0.010	−0.665	0.047	−0.227	−0.282	0.963
Kurtosis	3.199	3.678	2.970	3.436	3.422	3.926
P-values of tests						
Mean test statistic	0.034	0.116	0.053	0.037	0.373	0.142
VR test statistic	0.748	0.377	0.280	0.225	0.365	0.010
JB test statistic	0.9777	0.286	0.994	0.800	0.756	0.077
Corr$(t+1, t)$	0.0163	0.008*	0.000*	0.002*	0.001*	0.369

Notes: Results for males aged 65, payable until age 90, a discount rate of 4% and a sample size of 27, estimated over 1981–2007 and ages 64–89. See also notes to Table 1.

analysis (PCA), which is similar in spirit to the multi-factor extensions of the Lee–Carter model proposed by Booth et al. (2002b). They compare both the in-sample goodness of fit and the out-of-sample forecasting properties of their new PCA model against a number of established models (such as M1, M3 and M5) for a number of countries. Out-of-sample forecasting accuracy is measured along similar lines to Booth et al. (2006). In terms of mean absolute percentage error, the PCA model ranks second after M5 when tested on both male and female mortality rates for Taiwan, Japan, the USA, Canada, the UK and France across ages 60–99 over the period 1970–2005 (Yang et al., 2010).

Sweeting (2009) examines the two state variables $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ in M5 and concludes that they do not follow the random walk assumption proposed by Cairns et al. (2006a,b), but should instead be modelled as a random fluctuation around a trend, where the trend changes periodically. As a consequence, Sweeting shows that projected mortality rates embody much greater uncertainty than previously understood.

8. Conclusions

The present study sets out a framework for systematically evaluating the goodness of fit of stochastic mortality models, and applies it to a set of mortality models estimated using England & Wales male mortality data. If a model fits the data well, certain key residual series – those relating to mortality rates themselves, to the unobserved state variables that drive the dynamics of the model (including the cohort-effect where appropriate), and to the residuals of mortality-dependent financial prices – will, once standardized, be approximately iid N(0, 1). We then test whether the relevant series are compatible with iid N(0, 1).

We find that none of the models considered in this paper performs well in all sets of tests, and no model performs consistently better than the others. For the particular data set used in this analysis, however, we find that:

- For GOF tests of mortality residuals, model M2B performs best, M7 comes second and M6 third, and M1, M3 and M5 come some way behind.
- For the GOF tests of the state variables, M5, M6 and M7 perform best, in that order, although there is not much to choose between them. The other three models somewhat worse, and the worst performer is M2B.
- For the GOF tests of the annuity price residuals, M1 and M7 emerge as the best models and the other models come some way behind.

When we combine these findings with those from our earlier studies, we conclude that some models perform better under some assessment criteria than others, but that no single model can claim to be the victor. Further, different mortality patterns in different countries means that great care must be taken when selecting the best forecasting model for each country.

Three avenues for further work naturally suggest themselves. The first is to examine the dynamic properties of the state variables in more depth – and in particular, to test whether they follow the random walks which they are assumed to follow, and a start in that direction has been made by Sweeting (2009). The second is to test these findings on other mortality data sets. A third avenue of research, which is much more ambitious, is to build a mortality model that is able to take account of the impact of exogenous factors (such as biomedical, environmental, and socio-economic factors) on mortality rates (as per, e.g. Hanewald, 2009) or to apply a mortality model to cause-of-death data (as per Wilmoth, 1998).

Disclaimer

Additional information is available upon request. This report has been partially prepared by the Pension Advisory group, and not by any research department, of JPMorgan Chase & Co. and its subsidiaries (''JPMorgan''). Information herein is obtained from sources believed to be reliable but JPMorgan does not warrant its completeness or accuracy. Opinions and estimates constitute JPMorgan's judgment and are subject to change without notice. Past performance is not indicative of future results. This material is provided for informational purposes only and is not intended as a recommendation or an offer or solicitation for the purchase or sale of any security or financial instrument.

Acknowledgements

The authors thank Lixia Loh and Liang Zhao for excellent research assistance, and two referees for helpful comments. The usual caveat applies.

References

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

Booth, H., Maindonald, J., Smith, L., 2002b. Age-time interactions in mortality projection: aplying Lee–Carter to Australia, Working Papers in Demography, The Australian National University.

Booth, H., Hyndman, R.J., Tickle, L., De Jong, P., 2006. Lee–Carter mortality forecasting: a multi-country comparison of variants and extensions. Demographic Research 15, 289–310.

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

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

Cairns, A.J.G., Blake, D., Dowd, K., 2006b. Pricing death: frameworks for the valuation and securitization of mortality risk. ASTIN Bulletin 36, 79–120.

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

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Khalaf-Allah, M., 2010. A framework for forecasting mortality rates with an application to six stochastic mortality models. Pensions Institute Discussion Paper PI-0801, March.

Cochrane, J.H., 1988. How big is the random walk in GNP? Journal of Political Economy 96, 893–920.

CMI,, 2006. Stochastic projection methodologies: Further progress and P-Spline model features, example results and implications. Working Paper 20, Continuous Mortality Investigation. Available at: http://www.actuaries.org.uk/knowledge/cmi/cmi_wp/wp20.

CMI,, 2007. Stochastic projection methodologies: Lee–Carter model features, example results and implications. Working Paper 25, Continuous Mortality Investigation. Available at: http://www.actuaries.org.uk/knowledge/cmi/cmi_wp/wp25.

Coughlan, G.D., Epstein, D., Ong, A., Sinha, A., Balevich, I., Hevia Portocarrera, J., Gingrich, E., Khalaf-Allah, M., Joseph, P., 2007. LifeMetrics: A toolkit for measuring and managing longevity and mortality risks. Technical Document (JPMorgan, London, 13 March). Available at: www.lifemetrics.com.

Currie, I.D., Durban, M., Eilers, P.H.C., 2004. Smoothing and forecasting mortality rates. Statistical Modelling 4, 279–298.

Currie, I.D., 2006. Smoothing and forecasting mortality rates with P-splines. Presentation to the Institute of Actuaries. www.ma.hw.ac.uk/~iain/research.talks.html.

Dowd, K., 2005. Measuring Market Risk, second ed. John Wiley, Chichester and New York.

Dowd, K., Cairns, A.J.G., Blake, D., Coughlan, G.D., Epstein, D., Khalaf- Allah, M., 2010. Backtesting stochastic mortality models: An *ex-post* evaluation of multi-period-ahead density forecasts. North American Actuarial Journal, forthcoming.

Jacobsen, R., Keiding, N., Lynge, E., 2002. Long-term mortality trends behind low life expectancy of Danish women. J. Epidemiol. Community Health 56, 205–208.

Jarque, C., Bera, A., 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Economics Letters 6, 255–259.

Hanewald, K., 2009. Mortality modeling: Lee–Carter and the macroeconomy. Discussion Paper, Humboldt-Universität zu Berlin, May 19, 2009, available at SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1336888.

Huang, H.-C., Yue, J.C., Yang, S.S., 2008. An empirical study of mortality models in Taiwan. Asia-Pacific Journal of Risk and Insurance 3, 140–154.

Hyndman, R.J., Ullah, M.S., 2007. Robust forecasting of mortality and fertility rates: a functional data approach. Computational Statistics & Data Analysis 51, 4942–4956.

Lee, R.D., Carter, L.R., 1992. Modeling and forecasting US mortality. Journal of the American Statistical Association 87, 659–675.

Li, J.S-H., Hardy, M.R., Tan, K.S., 2009. Uncertainty in mortality forecasting: an extension to the classic Lee–Carter approach. ASTIN Bulletin 39, 137–164.

Lo, A.W., MacKinley, A.C., 1988. Stock prices do not follow random walks: evidence based on a simple specification test. Review of Financial Studies 1, 41–66.

Lo, A.W., MacKinley, A.C., 1989. The size and power of the variance ratio test in finite samples: a Monte Carlo investigation. Journal of Econometrics 40, 203–238.

Osmond, C., 1985. Using age, period and cohort models to estimate future mortality rates. International Journal of Epidemiology 14, 124–129.

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

Sweeting, P., 2009. A Trend-Change Extension of the Cairns–Blake–Dowd Model, Pensions Institute Discussion Paper PI-0904, February.

Wilmoth, J.R., 1998. Is the pace of Japanese mortality decline converging toward international trends? Population and Development Review 24, 593–600.

Yang, S.S., Yue, J.C., Huang, H.-C., 2010. Modeling longevity risks using a principal component approach: a comparison with existing stochastic mortality models. Insurance: Mathematics and Economics 46, 254–270.