Mortality Density Forecasts: An Analysis of Six Stochastic Mortality Models

Andrew J.G. Cairns[ab], David Blake[c], Kevin Dowd[c], Guy D. Coughlan[de], David Epstein[d], and Marwa Khalaf-Allah[d]

January 6, 2011

Abstract

This paper develops a framework for developing forecasts of future mortality rates. We discuss the suitability of six stochastic mortality models for forecasting future mortality and estimating the density of mortality rates at different ages. In particular, the models are assessed individually with reference to the following qualitative criteria that focus on the plausibility of their forecasts: 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 model. An important, though unsurprising, conclusion is that a good fit to historical data does not guarantee sensible forecasts. We also discuss the issue of model risk, common to many modelling situations in demography and elsewhere. We find that even for those models satisfying our qualitative criteria, there are significant differences between both central forecasts of mortality rates at different ages and the distributions surrounding those central forecasts.

Keywords: Plausibility, fan charts, model risk, forecasting, model selection criteria.

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

[b]Corresponding author: E-mail A.Cairns@ma.hw.ac.uk

[c]Pensions Institute, Cass Business School, City University, 106 Bunhill Row, London, EC1Y 8TZ, United Kingdom.

[d]Pension Advisory Group, JPMorgan Chase Bank, 125 London Wall, London, EC2Y 5AJ, United Kingdom.

1 Introduction

The last twenty years has seen a growing range of models for forecasting mortality. Early work on stochastic models by McNown and Rogers (1989) and Lee and Carter (1992) has been followed by:

- developments on the statistical foundations by, for example, Lee and Miller (2001), Brouhns et al. (2002), Booth et al. (2002a), Czado et al. (2005), Delwarde et al. (2007), and Li et al. (2009); and
- the development of new stochastic models by Booth et al. (2002a,b, 2005), Cairns et al. (2006b) (CBD), Renshaw and Haberman (2006), Hyndman and Ullah (2007), Cairns et al. (2009), Plat (2009) and Debonneuil (2010).

These stochastic models vary significantly according to a number of key elements: number of sources of randomness driving mortality improvements at different ages; assumptions of smoothness in the age and period dimensions; inclusion or not of cohort effects; estimation method.

A number of studies have sought to draw out more formal comparisons between a number of these models. Some of these limit themselves to comparison of some variants of the Lee-Carter model (Lee and Miller, 2001, and Booth et al., 2002a,b, 2005). Hyndman and Ullah (2007) compare out-of-sample forecasting performance of Lee-Carter and its Lee-Miller and Booth-Maindonald-Smith variants with a new class of multifactor models. CMI (2005, 2006, 2007), compare the Lee-Carter, Renshaw and Haberman and P-splines models. Extension of these types of analysis has been extended to a wider range of models with substantially different characteristics by the present authors, of which this paper is one part.

Cairns et al. (2009) focused on quantitative and qualitative comparisons of eight stochastic mortality models (see 1 in Section 2), based on their general characteristics and ability to explain *historical* patterns of mortality. The criteria employed included: quality of fit, as measured by the Bayes Information Criterion (BIC); ease of implementation; parsimony; transparency; incorporation of cohort effects; ability to produce a non-trivial correlation structure between ages; robustness of parameter estimates relative to the period of data employed.

Complementing this, Dowd et al. (2010a,b) carry out a range of formal, out-of-sample backtesting and goodness-of-fit tests using English and Welsh males mortality data. They find that some models fare better under some criteria than others, but that no single model can claim superiority under all the criteria considered. In any event, different patterns of mortality improvements in different countries means that models that are best for one country might not be as suitable for another. Finally, this paper focuses on the *ex ante* plausibility and robustness of forecasts produced by the different models. The present paper, therefore, focuses on the *ex*

ante qualitative aspects of forecasts, while the previous works (Cairns et al., 2009, Dowd et al., 2010a,b) focus on the *ex post* quantitative.

Building on the analyses of historical data of Cairns et al. (2009) and Dowd et al. (2010a,b), the present paper focuses on *ex ante* qualitative aspects of mortality forecasts and the distribution of results around central forecasts. Specifically, we introduce a number of qualitative criteria that focus on the plausibility of forecasts made by different models.

Often in this paper, we will refer to the concept of *biological reasonableness* (which was first raised in Cairns et al. 2006a). The concept is not intended to refer to criteria based on hard scientific (biological or medical) facts. Instead, it is intended to cover a wide range of subjective criteria, related to biology, medicine and the environment. What the modeller needs to do is look at the results and ask the question: *what mixture of biological factors, medical advances and envorinmental changes would have to happen to cause this particular set of forecasts?* As one example, the upper set of projections in Figure 4 at age 85 looks rather more unusual than the two lower sets of projections under a particular model. Under the upper scenario, we would have to think of a convincing biological, medical or environmental reason why, *with certainty*, age 85 mortality rates are going to deteriorate to 1960's levels. If the modeller cannot think of any good reason why this might happen, then she must rule out the model (at least with its current method of calibration) on grounds of biological unreasonableness.

Besides biological reasonableness, we also consider the issue of *the plausibility of forecast levels of uncertainty in projections at different ages*. The objective here is to judge whether or not the pattern of uncertainty at different ages is consistent with historical levels of variability at different ages: we can sometimes conclude that a particular model is less plausible on the basis of forecast levels of uncertainty.

An important additional issue concerns the *robustness of forecasts* relative to the choice of sample period and age range. If we make a small change either to the sample period (for example, when we add in the latest mortality data) or to the age range, we would normally expect to see, with a robust model, only modest changes in the forecasts at all ages. Where a model is found to lack robustness with one sample population, there is a danger that it will lack robustness if applied to another sample population and should, therefore, either be used with great care or not used at all.

Although application of such a wide ranging set of model selection criteria will eliminate some models, we will demonstrate that mortality forecasting is no different from many other modelling problems where model risk is significant: mortality forecasters should acknowledge this fact and make use of multiple models rather than pretend that it is sufficient to make forecasts based on any single model.

1.1 Plan for this paper

We will consider qualitative assessment criteria that allow us to examine the *ex ante* plausibility of the forecasts generated by six stochastic mortality models, illustrating with national population data for England & Wales (EW) for an age group consisting of 60-89 year old males and estimated over years 1961-2004. This is supplemented by a more brief discussion of forecasts for the equivalent US dataset. We focus on higher ages because our current principal research interest is the longevity risk facing pension plans and annuity providers.

We will concentrate on six of the models discussed by Cairns et al. (2009): these are labelled in Table 1 as M1, M2, M3, M5, M7 and M8. Models M2, M3, M7 and M8 include a cohort effect and these emerged in Cairns et al. (2009) as the best fitting, in terms of BIC, of the eight models considered on the basis of male mortality data from EW and the US for the age group under consideration. M2 is the Renshaw and Haberman (2006) extension of the original Lee-Carter model (M1), M3 is a special case of M2, and M7 and M8 are extensions of the original CBD model (M5). The original Lee-Carter and CBD models had no cohort effect, and provide useful benchmarks for comparison with the four models involving cohort effects. M4 is not considered any further in this study because of its low BIC and qualitative rankings for these datasets in Cairns et al. (2009, Table 3). (M4 focuses on identifying the smooth underlying trend. However, this means that it is not as good as the other models at capturing short-term deviations from this trend.) Although M3 is a special case of M2, we include it here because it had a relatively high BIC ranking for the US data, and because it avoids a problem with the robustness of parameter estimates for M2 identified by CMI (2007), Cairns et al. (2009), and Dowd et al. (2010a,b). M6 was also dropped from the original set of eight models: M6 is a special case of M7, and M7 was found to be stable and to deliver consistently better and more plausible results than M6.

The structure of the paper is as follows. In Section 2, we specify the stochastic processes needed for forecasting the term structure of mortality rates for each of the models. Results for the different models using EW male mortality data are compared and contrasted in Section 3. Section 5 examines two applications of the forecast models, namely applications to survivor indices and annuity prices, and makes additional comments on model risk and plausibility of the forecasts. Each model is then tested for the robustness of its forecasts in Section 4. Finally, in Section 6, we summarise an analysis for US male mortality data: our aim is to draw out features of the US data that are distinct from the EW data. Section 7 concludes.

Model	formula
M1	$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}$
M2	$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x}^{(3)}$
M3	$\log m(t,x) = \beta_x^{(1)} + n_a^{-1}\kappa_t^{(2)} + n_a^{-1}\gamma_{t-x}^{(3)}$
M5	logit $q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x})$
M7	logit $q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \kappa_t^{(3)}\left((x-\bar{x})^2 - \hat{\sigma}_x^2\right) + \gamma_{t-x}^{(4)}$
M8	logit $q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \gamma_{t-x}^{(3)}(x_c - x)$

Table 1: Formulae for six out of the original eight mortality models investigated by Cairns et al. (2009). The functions $\beta_x^{(i)}$, $\kappa_t^{(i)}$, and $\gamma_{t-x}^{(i)}$ are age, period and cohort effects, respectively. $\bar{x}$ is the mean age over the range of ages being used in the analysis. $\hat{\sigma}_x^2$ is the mean value of $(x-\bar{x})^2$. n_a is the number of ages.

2 Forecasting with stochastic mortality models

We take six stochastic mortality models which, on the basis of fitting to historical data, appear to be suitable candidates for forecasting future mortality at higher ages, and prepare them for forecasting. To do this, we need to specify the stochastic processes that drive the age, period and (if present) cohort effects in each model.

We define $m(t,x)$ to be the death rate in year t at age x, and $q(t,x)$ to be the corresponding mortality rate, with the relationship between them given by $q(t,x) = 1 - \exp[-m(t,x)]$. The models considered are outlined in Table 1.

All but M5 require the use of one or more identifiability constraints (see Appendix A, Section A.1), and age-, period- and cohort-effect parameter values are estimated using a Newton-Raphson type of iterative scheme (see Appendix A, Section A.2).

The six models investigated in this paper, while representative of the class of time-series age-period-cohort models, are not the only models in their class. For example, we do not consider the models proposed by Booth et al. (2002a,b) (a multifactor age-period extension of M1), Hyndman and Ullah (2007) (a similar type of extension to M1 that smooths the mortality data across ages before fitting the model), Plat (2009) (adapting features of both M2 and M7 facilitating extension to lower ages), and Butt and Haberman (2010) (further variations on Plat, 2009).

Before we discuss the forecasting models in detail, we add a word of caution. The stochastic models used here have been found to be appropriate for the specific datasets used in this paper. Potential users of these and other models should not assume that the same stochastic models will be appropriate for other datasets, and so

the process leading to the selection of appropriate time series models for the period and cohort effects should never be bypassed.

2.1 Age effects

The age effects, $\beta_x^{(i)}$, are either non-parametric and estimated from historical data (M1-M3), or assume some particular functional form (M5-M8). Further, we focus on forecasts of mortality within the same range of ages used to estimate the underlying models, so it is not necessary in this paper to simulate or extrapolate the age effects.

2.2 Period effects

Random-walk processes have been widely used to drive the dynamics of the period effect ever since the introduction of the original Lee-Carter (1992) model. The method used to estimate the model has been refined by subsequent authors in order to improve the fit and place the model on more secure statistical foundations (see, for example, Brouhns et al., 2002, Booth et al., 2002a, Czado et al., 2005, and de Jong and Tickle, 2006).

Following CBD, we use a multivariate random walk with drift and correlated innovations to drive the dynamics of the period effect: that is,

$$\kappa_t^{(i)} = \kappa_{t-1}^{(i)} + \mu_\kappa^{(i)} + \sigma_\kappa^{(i)} Z_\kappa^{(i)}(t)$$

where the $\mu_\kappa^{(i)}$ are the drifts, the $\sigma_\kappa^{(i)}$ are the volatilities, and the $Z_\kappa^{(i)}(t)$ are standard normal innovations that are correlated across the components, i, but independent through time. This model appears to be consistent with the data (see the plots of the $\kappa_t^{(i)}$ in Cairns et al. (2009)). However, more general ARIMA models might provide a better fit statistically to some datasets. For example, CMI (2007) uses an ARIMA(1,1,0) process for the period effect in the Lee-Carter model (M1) and an ARIMA(2,1,0) process for the period effect in the Renshaw and Haberman model (M2).

Additionally, it is appropriate to consider the suitability, or otherwise, of the random-walk model using biological reasonableness as a criterion. We can take M7 as an example where the random-walk assumption could be questioned. For example, a positive drift in $\kappa_t^{(3)}$ could result in the mortality curve between ages 60 and 89 adopting a 'U' shape at some point in the future, which we see as unlikely or even implausible from a biological point of view. From a qualitative point of view, therefore, it might be appropriate to model both $\kappa_t^{(2)}$ and $\kappa_t^{(3)}$ as mean reverting processes (see Cairns et al., 2008b, Section 4.4). However, this might pose difficulties as the limited data available could result in considerable paramater uncertainty. In practice, for the specific paramaterisations of the multivariate random-walk model we

have investigated, the age 60 to 89 mortality curve retains a biologically reasonable shape up to at least a forecasting horizon of 50 years. In practical terms, therefore, we are comfortable with the continued use of the multivariate random-walk model.

2.3 Cohort effects

The principal challenge we face in building a stochastic mortality model that can be used for forecasting lies in specifying the dynamic process driving the cohort effect. As a simple starting point, we follow previous studies (e.g., Renshaw and Haberman, 2006, and CMI, 2007), and assume that the cohort effect, $\gamma_{t-x}^{(i)}$, has dynamics that are independent of the period effect, $\kappa_t^{(i)}$.

Fitted values for the cohort effects for models M2, M3, M7 and M8 can be seen in Figure 1 (dots). It is clear from looking at these plots, that a simple random-walk process is unlikely to be appropriate, leading us to consider a variety of alternatives.

For each of models M2, M3, M7 and M8, we considered a full range of ARIMA(p, d, q) models with $d = 0, 1, 2$ and $p, q = 0, 1, 2, 3, 4$ as candidates for the cohort effects. In the case of M8, we also considered an AR(1) model around a linear drift. The Bayes Information Criterion (BIC) was calculated for each ARIMA model and, based on this information, we drew up a short list of suitable candidates for each of M2, M3, M7 and M8. In some circumstances ARIMA models with the highest BIC were rejected because they gave rise to implausible forecasts. In these cases, the preferred ARIMA models tended to be simpler (lower p and q values) in order to produce more plausible forecasts. Only M7 produced a clear single option for the cohort effect (AR(1)). For each of M2, M3 and M8, we analysed two versions denoted M2A, M2B, M3A, M3B, M8A and M8B (see Cairns *et al.*, 2008a, for further details).

The models are summarised in Table 2.

2.4 Estimation of cohort effects

Our main focus in this paper is making forecasts based on data for EW males aged 60 to 89 over the period 1961 to 2004. If we genuinely have the true model and a very large population, then the single observation for age 60 in 2004 will be sufficient for us to get an accurate estimate of the 1944 (i.e. $2004-60$) cohort effect. In reality, we are exposed to model risk, and even the EW males population has significant noise in its death counts. To counter these effects, Cairns et al. (2009) proposed that cohorts with fewer than 5 observations be excluded from the estimation procedure, to prevent overfitting of the cohort effect for these cohorts. In this case, therefore, the four most recent (1941 to 1944) and earliest (up to 1880) cohorts were excluded (see Cairns et al., 2009, Section 2.1). To ensure the eight models in Cairns et al.

Model	$\gamma_c =$	Model for the cohort effect	
M2A	$\gamma_c^{(3)}$	$ARIMA(0,2,1)$	$\gamma_c = 2\gamma_{c-1} - \gamma_{c-2} + \phi_0 Z_\gamma(c) + \phi_1 Z_\gamma(c-1)$
M2B	$\gamma_c^{(3)}$	$ARIMA(1,1,0)$	$\gamma_c = \gamma_{c-1} + \mu_\gamma + \alpha_\gamma(\gamma_{c-1} - \gamma_{c-2} - \mu_\gamma) + \phi_0 Z_\gamma(c)$
M3A	$\gamma_c^{(3)}$	$ARIMA(0,2,1)$	$\gamma_c = 2\gamma_{c-1} - \gamma_{c-2} + \phi_0 Z_\gamma(c) + \phi_1 Z_\gamma(c-1)$
M3B	$\gamma_c^{(3)}$	$ARIMA(1,1,0)$	$\gamma_c = \gamma_{c-1} + \mu_\gamma + \alpha_\gamma(\gamma_{c-1} - \gamma_{c-2} - \mu_\gamma) + \phi_0 Z_\gamma(c)$
M7	$\gamma_c^{(4)}$	$ARIMA(1,0,0) \equiv AR(1)$	$\gamma_c = \mu_\gamma + \alpha_\gamma(\gamma_{c-1} - \mu_\gamma) + \phi_0 Z_\gamma(c)$
M8A	$\gamma_c^{(3)}$	$AR(1)$ around a linear drift	$\gamma_c - \delta_1 c = \mu_\gamma + \alpha_\gamma(\gamma_{c-1} - \delta_1(c-1) - \mu_\gamma) + \phi_0 Z_\gamma(c)$
M8B	$\gamma_c^{(3)}$	$AR(1)$ with no drift	$\gamma_c = \mu_\gamma + \alpha_\gamma(\gamma_{c-1} - \mu_\gamma) + \phi_0 Z_\gamma(c)$

Table 2: ARIMA models for the cohort effect for models M2, M3, M7 and M8. The $Z_\gamma(c)$ are independent and identically distributed standard normal innovations, that are independent of the period effect innovations, $Z_\kappa^{(i)}(t)$.

(2009) were considered on a consistent basis, the Lee-Carter and CBD models that have no cohort effect were also fitted to the same dataset.

2.5 Interplay between age, period and cohort effects

A recurring theme in this paper is the possibility that cohort effects might be partially or completely replaced by well-chosen age and period effects. As an example, with M3, a linear cohort effect can be completely replaced by linear adjustments to the age and period effects. In other cases, application of an identifiabilty constraint transforms the cohort effect but does not eliminate it. In further cases, there is no identifiability constraint that can be applied. Nevertheless, from time to time, we remark that an observed cohort effect points us towards the use of a more complex model that has additional age and period effects, maybe with a simplified cohort effect. This discussion then raises the question of whether a cohort effect is needed at all, given the possibility that it could be replaced by additional age-period effects. However, this brings us back, first, to what type of effects we believe to be appropriate for inclusion in a stochastic mortality model, and, second, to the relative parsimony of the models that we attempt to fit.

It is appropriate at this point to review two distinct philosophies underlying model building. Any model, however sophisticated, can only ever be a crude and imperfect representation of reality. But is it better to start with a simple model and expand it as its weaknesses emerge? Or is it better to begin with a very general model and attempt to simplify the model in the light of results obtained, recognising that the principle of parsimony dictates that a simple but well-specified model is preferred to a more complex model? Mortality modelling has traditionally adopted the former approach, but other disciplines, such as economics and finance, have adopted the

latter, so-called general-to-specific modelling framework (see, e.g., Campos et al. (2005) and Bauwens and Sucarrat (2008)). Clearly, general-to-specific modelling has advantages, since more general models are able to encompass simpler models, whereas the opposite will not be true.

Returning again to mortality modelling, the general-to-specific approach would begin by considering a wide range of factors that the model builder believes could determine mortality rates in the population of interest. These would certainly include gender, year of birth, current age, education, occupation, health status, lifestyle indicators, ethnicity and so on. Given the difficulties of measuring some of these factors and given the computational problems of estimating models with such a large number of factors, the model builder will inevitably be drawn to choosing a simpler model that still tries to capture the influence of (at least some of) the wider group of factors.

For the national populations considered here, we have what we believe to be reliable deaths and exposures data by age, calendar year and gender. Age and calendar year allows us to identify individual cohorts. Subdivisions by covariates such as social class, educational attainment and smoking status tend to be much less reliable, at least over long periods of time, and so these covariates tend not to be included in studies focusing on population mortality forecasting.

The relevance of the general-to-specific approach here is as follows. We have argued in Section 1 why there should be a cohort effect in all populations. The approach therefore suggests that we should first include it as a component in at least some models and then test if this effect is significant or not. Additionally, we need to consider if a well-designed and significant cohort effect results in a more parsimonious model than, say, a model with additional age-period effects.

3 Forecasts and model comparisons

We now proceed to compare the forecasting results for EW for the nine models M1, M2A, M2B, M3A, M3B, M5, M7, M8A and M8B. (Corresponding results for US males are presented and discussed in Section 6.) To do this, we will present fan charts of the forecasts produced by the models. Each fan chart illustrates the forecast output from the stochastic mortality models by dividing the simulated densities into 5% quantile bands. Fan charts give us the opportunity to explore any distinctive visual features of each model's forecasts, as well as any differences between them. This, in turn, will give us a first indication of the degree of model risk. These visual comparisons are supplemented by a range of quantitative and qualitative diagnostics which will increase our confidence in some models and question the suitability of others for our purposes.

We consider the plausibility of model forecasts by assessing the biological reasonable-

ness of: the projections of the future term-structure of mortality; projected period and cohort effects; and forecast levels of uncertainty relative to historical levels of uncertainty. These three criteria are, of course, closely related, but it is useful to think about each separately. Although 'plausibility' is a rather subjective concept that is difficult to define, the forecasts produced by some of the models turn out to be so obviously implausible that they can be ruled out for use with this specific dataset. In Section 4, we consider a fourth criterion, namely, the robustness of model forecasts in the face of changes to the historical data sets used to calibrate the model; this continues a discussion initiated by Cairns et al. (2009) who considered the robustness of parameter estimates.

An examination of Figures 1 to 3 reveals the following:

- Figure 1 shows fan charts for the cohort effects for each model. (M1 and M5 have no cohort effect and are not plotted.) We can see that M2A's and M3A's fans have a distinctively different shape from the other models, and expand without limit. The same is true for M2B's and M3B's fans, although this is less obvious from the plots. These are a result of the second- and first-order differencing in these models, respectively. The fans for M2B and M3B seem plausible, whereas the fans for M2A and M3A seem less so, because of the rapidity with which they spread out.

  The differences between the fan charts for M8A and M8B reflect differences in the trend in $\gamma_c^{(3)}$ (which trend the latter model sets to zero). Both models' fans converge to a finite width, a consequence of using a stationary AR(1) process for the cohort effect. However, model M8A's fan is slightly narrower, and this reflects the fact that the lack of a constraint on the drift allows the estimation procedure to achieve a tighter fit than is the case with M8B.

  The different structure of each model inevitably means that each chart is visually distinctive. This might be a sign that model risk is significant, although this cannot be fully established before we investigate some key output variables.

- In Figure 1, M2A, M3A and M8A all incorporate a linear trend. As remarked earlier (Section 2.3), a linear trend can be converted into a mixture of age-period effects. If these cannot be merged into existing age-period effects, this might imply that the model is deficient in the following sense: the age-cohort effect is being used to compensate for an inadequate number of age-period components. It might not be sufficient, for example, to augment the Lee-Carter model, M1, solely by the addition of an age-cohort component, as in M2A. Rather, it might be more appropriate to extend the Lee-Carter model by adding an age-period component as well as, or instead of, an age-cohort component, with a further requirement that any cohort effect has no drift. We do not consider such an extension in this paper.

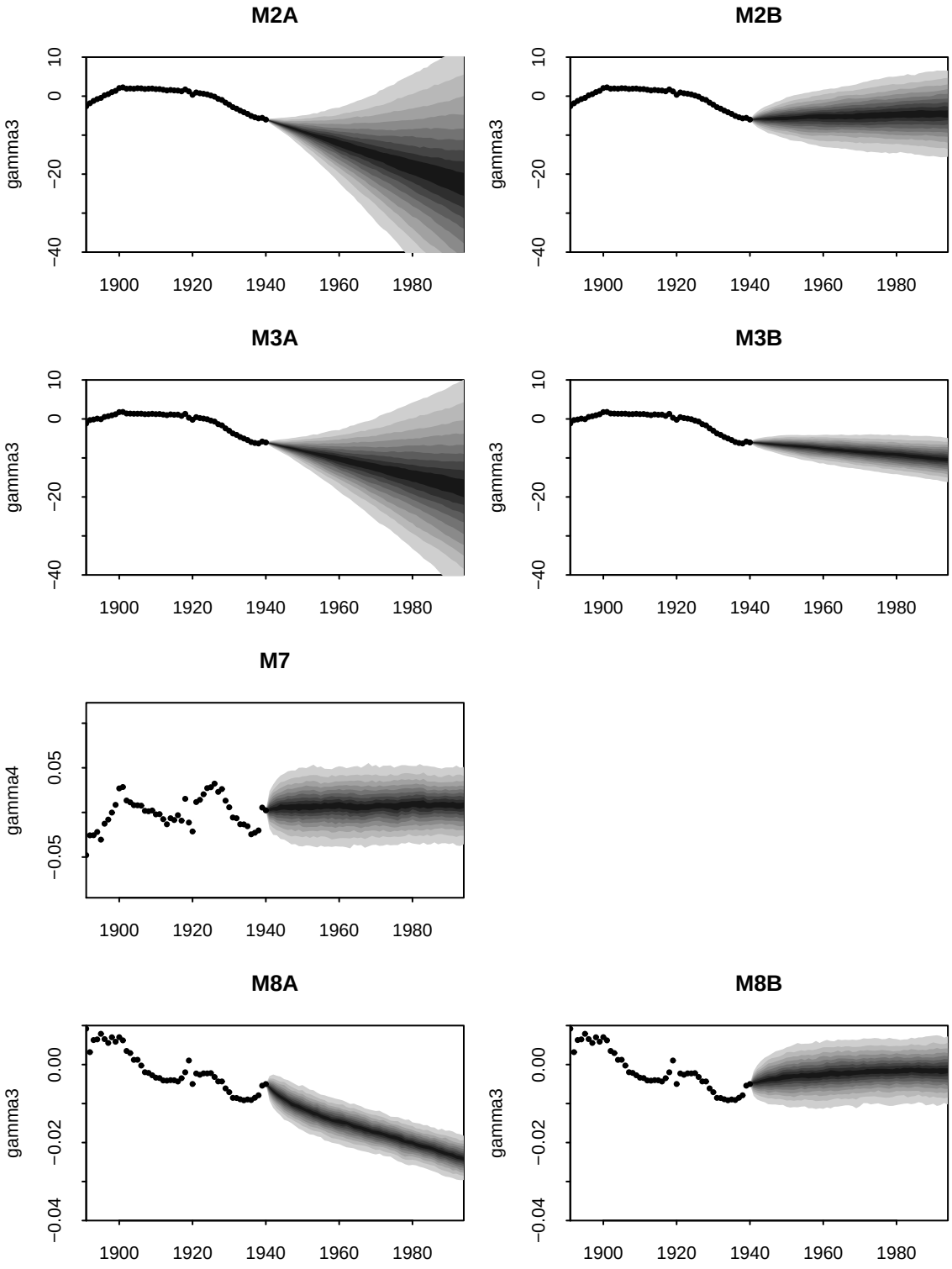

Figure 1: England & Wales, males: Fan charts for the projected cohort effect. For M1 and M5, there is no cohort effect so no fan charts have been plotted. The dots show estimates of the cohort effect fitted to the historical dataset.

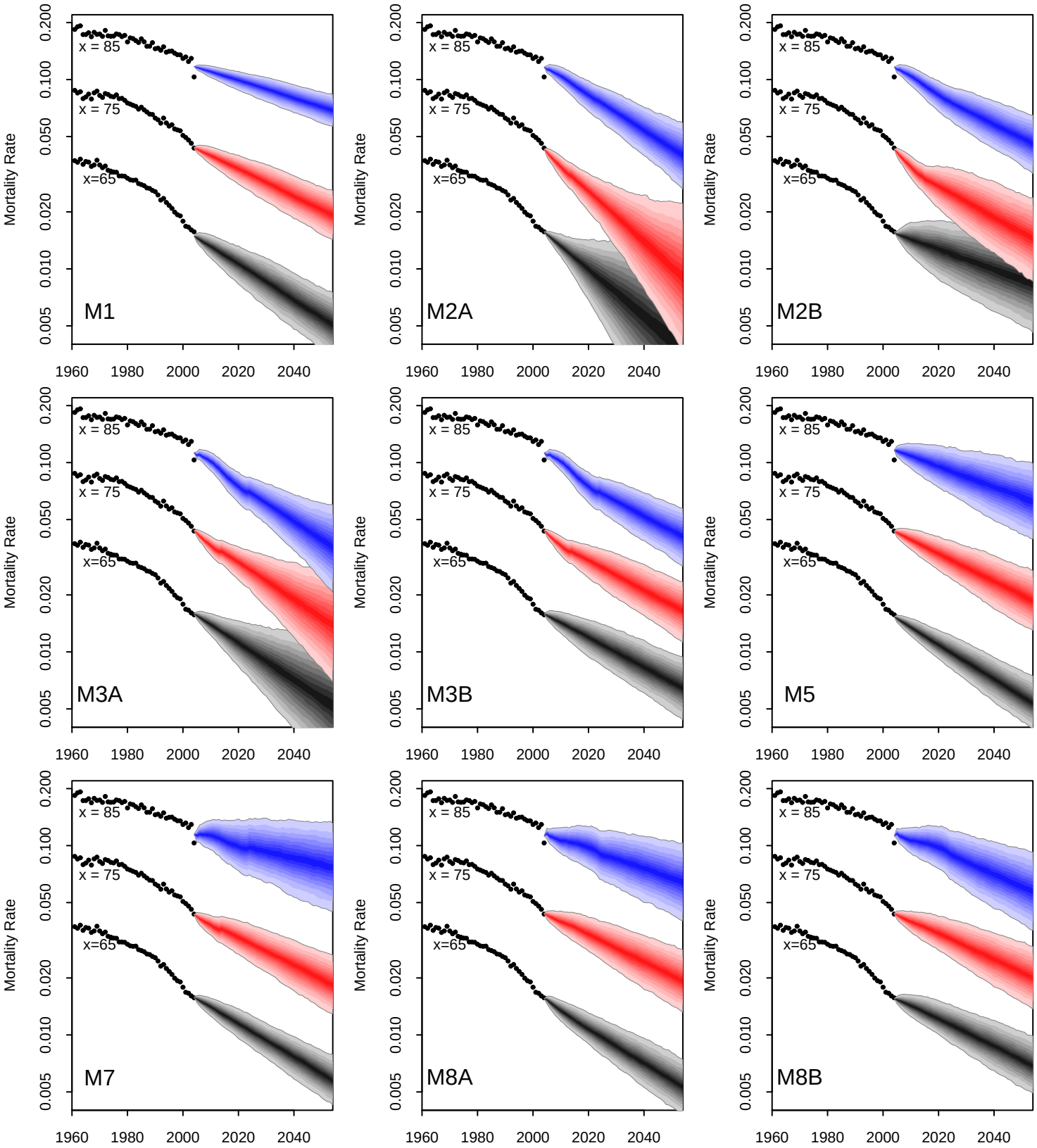

Figure 2: England & Wales, males: Mortality rates, $q(t, x)$, for models M1, M2A, M2B, M3A, M3B, M5, M7, M8A and M8B for ages $x = 65$ (bottom fan), 75 (middle fan), and 85 (top fan). The dots show historical mortality rates for 1961 to 2004.

- Figure 2 shows fan charts for mortality rates at ages 65, 75 and 85 for each of the nine models. In each case, except for M1 and M5, the central trend at age 65 seems relatively smooth, while, at age 85, it wobbles around until 2025 (most obviously, see the upwards kink in the age 85 fan in 2025 in the M3A/B plots). This is because the central trend is linked to the estimated cohort effect, $\gamma_c^{(3)}$ ($\gamma_c^{(4)}$ for M7). The cohort effect has been estimated for years of birth up to 1940. At age 85, the mortality rate is influenced by the estimated cohort effect right up to 2025 when the 1940 cohort reaches age 85. After 2025, age-85 mortality rates depend on smooth projections of the cohort effect. At age 65, the smoother projected cohort effect is evident almost immediately.

- Figure 2 allows us to make an interesting comparison between model M1, on the one hand, and M5, M7, M8A and M8B, on the other. With M1, the age-85 fans are narrower than the age-65 fans. The opposite is true for models M5, M7, M8A and M8B. For these models, the predicted uncertainty is consistent with the greater observed volatility in age-85 mortality rates between 1961 and 2004 than in age-65 mortality rates over the same period (see Appendix B). The contrasting result for M1 occurs because it has a single stochastic period effect, $\kappa_t^{(2)}$. For M1, the widths of the fans are proportional to the age effect, $\beta_x^{(2)}$ and this is unlikely to satisfy the criterion of biological reasonableness. The shape of the fitted $\beta_x^{(2)}$ curve tends to be influenced primarily by relative rates of improvement at different ages over the historical observation period. Historical improvements have been lower at higher ages, forcing $\beta_x^{(2)}$ to be lower at higher ages (see Cairns et al, 2009, Figure 7), so causing the fans at higher ages to be narrower, rather than wider.

  Similarly, fans for M2A, M2B and M3A are noticeably wider at age 65 than age 85.

- Figure 3 allows us to make a more detailed comparison of the mortality fans produced by the different models by overlaying the fans for six out of the nine model variants under consideration: M1, M2B, M3B, M5, M7 and M8B.

  At age 65 (bottom graph), all but the M2B fans have roughly equal width. The central trends, however, are noticeably different. For example, the difference in trend between M5 and M7 equates to a difference in the rate of improvement in the age-65 mortality rate of 0.3% per annum.(Specifically, for age 65, the M5 improvement rate was 2.1% per annum, while for M7 the improvement rate was 1.8% per annum.)

  The differences in trend are even bigger at age 85 (M5 versus M7: 0.6% per annum). But at age 85, we also see a noticeable difference between the spreads of the M1, M3B, M5, M7 and M8B fans. M1 has the narrowest fan for reasons already mentioned earlier. M5, M7 and M8B are closer in terms of the width of the fans. M7, with three random period effects, has the widest fan, with the

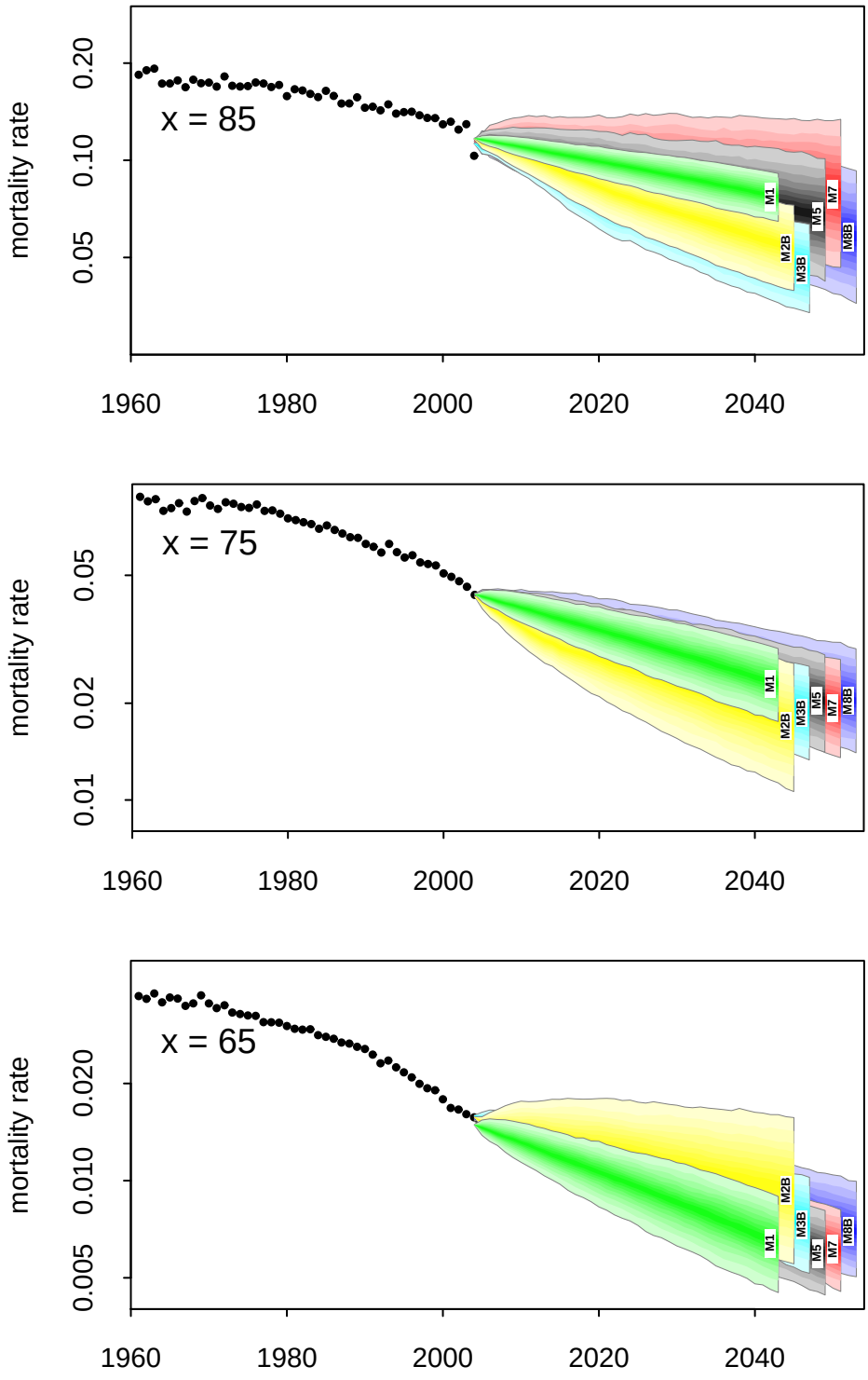

Figure 3: England & Wales, males: Mortality rate fan charts, $q(t, x)$, for models M1 (shortest and uppermost fan), M2B (second top fan), M3B (third fan) M5 (fourth), M7 (fifth), and M8A (sixth and rearmost fan) for ages $x = 65$, 75, and 85. The dots show historical mortality rates for 1961 to 2004.

high degree of uncertainty at age 85 resulting from a mixture of the variances of and covariances between the $\kappa_t^{(i)}$ and $\beta_x^{(i)}$ terms.

- Similar comparisons were made of M2A versus M2B, M3A versus M3B, and M8A versus M8B (see Figure 2). In all cases, we found that the choice of model did have at least a moderate impact on forecasts. In each case, central projections at age 65 were most susceptible to the choice of model, reflecting differences in the central trend in the cohort effect (Figure 1). Fans for mortality rates at ages 65 and 75 under M2A were much wider than those under M2B (the same holds for M3A and M3B), reflecting the greater uncertainty under the ARIMA(0,2,1) model. In contrast, fan widths under M8A and M8B were reasonably similar, reflecting earlier remarks about future uncertainty in the M8 cohort effect.

In terms of the suitability of the models for the dataset under consideration, we can summarise as follows: The figures reveal reasonable consistency of forecasts between M3B, M5, M7 and M8B, all of which pass the plausibility criterion, but with sufficient differences for model risk to be recognised as a significant issue. The figures also lead us to question the plausibility of the forecasts produced by M1 and M2 for this dataset, since they imply that forecasts of mortality at age 85 are much less uncertain than at age 65, contrary to historical evidence (Appendix B).

4 Robustness of projections

We now assess the projections from models M1, M2B, M3B, M5, M7, M8A and M8B for robustness relative to the sample period used in estimating the model. For each model, we compare three sets of simulations:

- Scenario 1: (A) The underlying model is first fitted to mortality data from 1961 to 2004. (B) The stochastic model for the $\kappa_t^{(i)}$period effects and the $\gamma_{t-x}^{(i)}$ cohort effects is then fitted to the full set of values resulting from (A) (44 $\kappa_t^{(i)}$'s and 60 $\gamma_{t-x}^{(i)}$'s).
- Scenario 2: (A) The underlying model is first fitted to mortality data from 1961 to 2004. (B) The stochastic model for the $\kappa_t^{(i)}$ period effects and the $\gamma_{t-x}^{(i)}$ cohort effects is then fitted to a restricted set of values resulting from (A) (the final 24 $\kappa_t^{(i)}$'s and the final 45 $\gamma_{t-x}^{(i)}$'s; i.e. the same number as Scenario 3).
- Scenario 3: (A) The underlying model is first fitted to mortality data from 1981 to 2004. (B) The stochastic model for the $\kappa_t^{(i)}$period effects and the $\gamma_{t-x}^{(i)}$ cohort effects is then fitted to the full set of values resulting from (A) (24 $\kappa_t^{(i)}$'s and 45 $\gamma_{t-x}^{(i)}$'s).

Typical results are presented here for models M2B, M3B and M7 (Figures 4 to 6 respectively).

If the period and cohort effects were, in fact, observable, then we would be using the same 24 $\kappa_t^{(i)}$'s and the same 45 $\gamma_{t-x}^{(i)}$'s to generate the fans under scenarios 2 and 3, implying that the two sets of fans should be the same. In reality the age, period and cohort effects need to be estimated. Therefore, noise in the data, combined with the fact that the models are approximations to a more complex reality, means that estimates of these effects will be sensitive to the choice of scenario. The subsequent estimation errors then feed through to differences between the scenario 2 and 3 forecasts. However, if a model is robust, then we would expect these differences to be modest, and for the two sets of fans (2 and 3) to have similar median trajectories and similar spreads.

Our investigations allow us to make the following observations (see Figures 4 to 6 by way of example):

- In most cases, the central trajectory of the mortality fans is closely connected to the start and end years used to fit the simulation model for the period effects. (Indeed, for a pure random walk process, the median forecast is a straight line extrapolation of the line connecting the first and the last observations.) For example, if the central projections in the scenario 1 fans are extrapolated

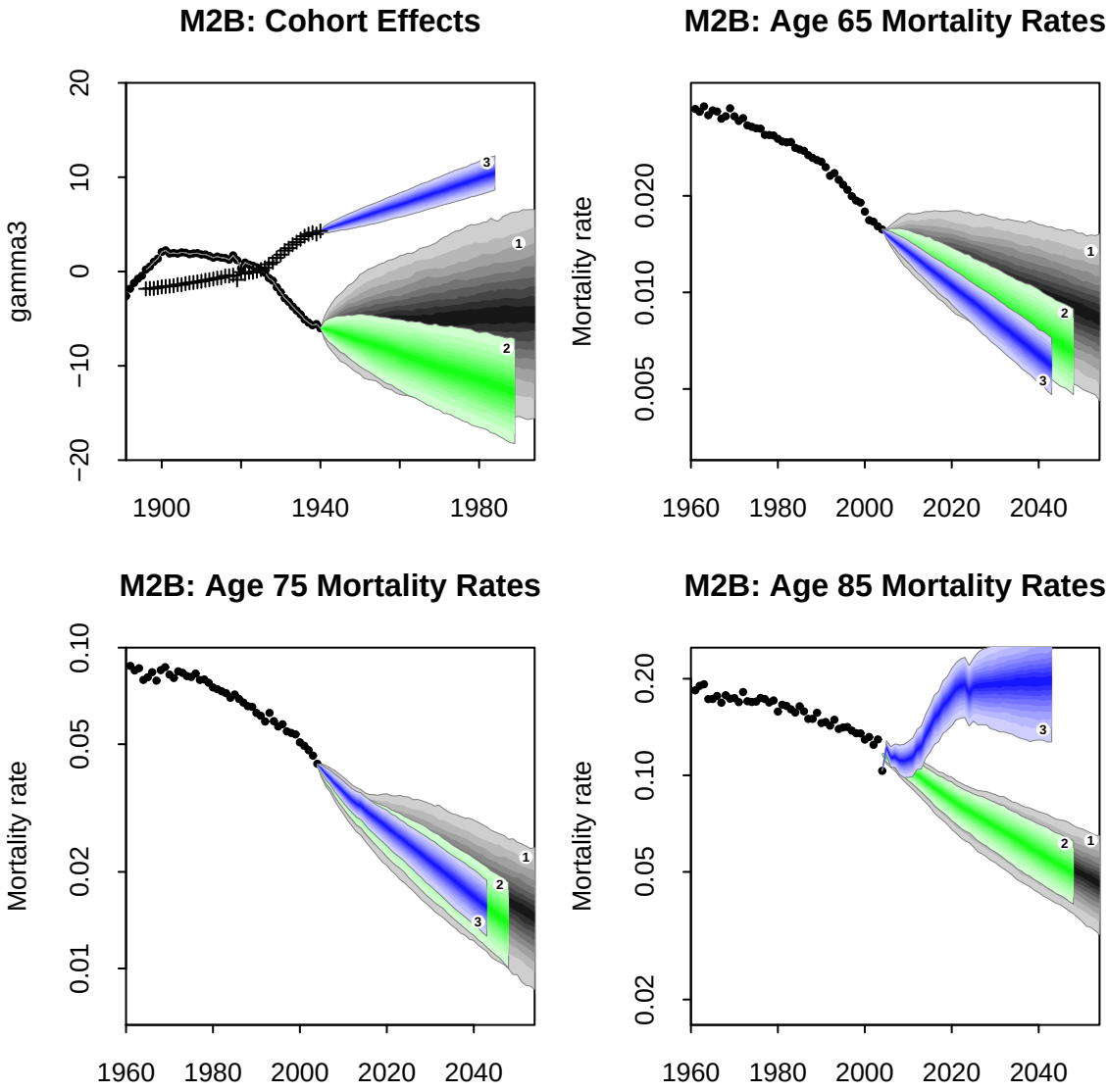

Figure 4: England & Wales, males: Model M2B. Cohort effect and mortality rates for ages 65, 75 and 85. **Notes:** Dots and rearmost fan (label 1 in each plot): Scenario 1, historical data from 1961 to 2004 used to estimate the historical $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 44 $\kappa_t^{(2)}$ values and the 60 $\gamma_c^{(3)}$ values. Dots and middle fan (label 2): Scenario 2, historical data from 1961 to 2004 used to estimate the historical $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 24 most-recent $\kappa_t^{(2)}$ values and the 45 most-recent $\gamma_c^{(3)}$ values. Crosses and uppermost fan (label 3): Scenario 3, historical data from 1981 to 2004 used to estimate the historical $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 24 fitted $\kappa_t^{(2)}$ values and the full 45 fitted $\gamma_c^{(3)}$ values.

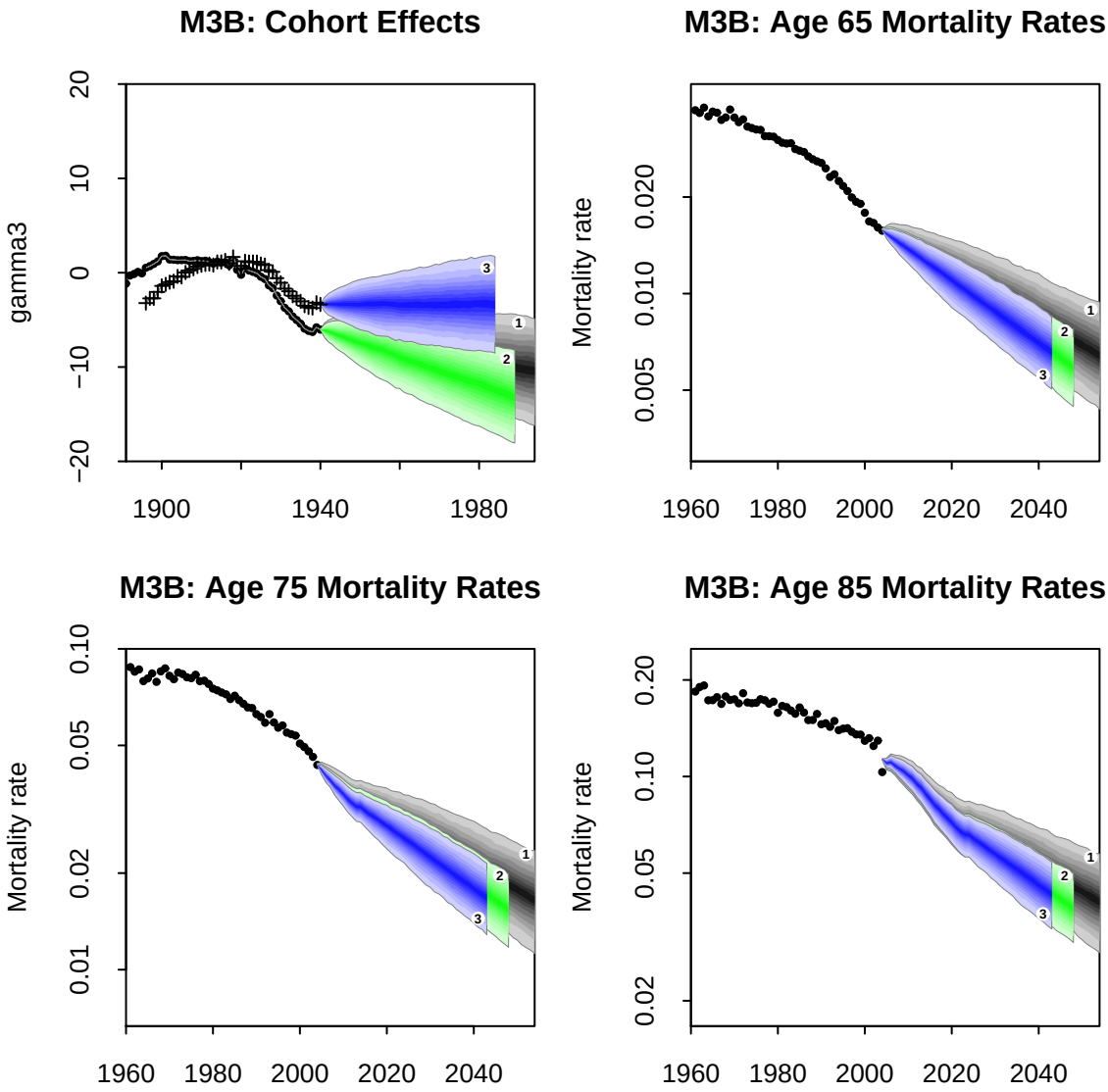

Figure 5: England & Wales, males: Model M3B. Cohort effect and mortality rates for ages 65, 75 and 85. Notes: see Figure 4.

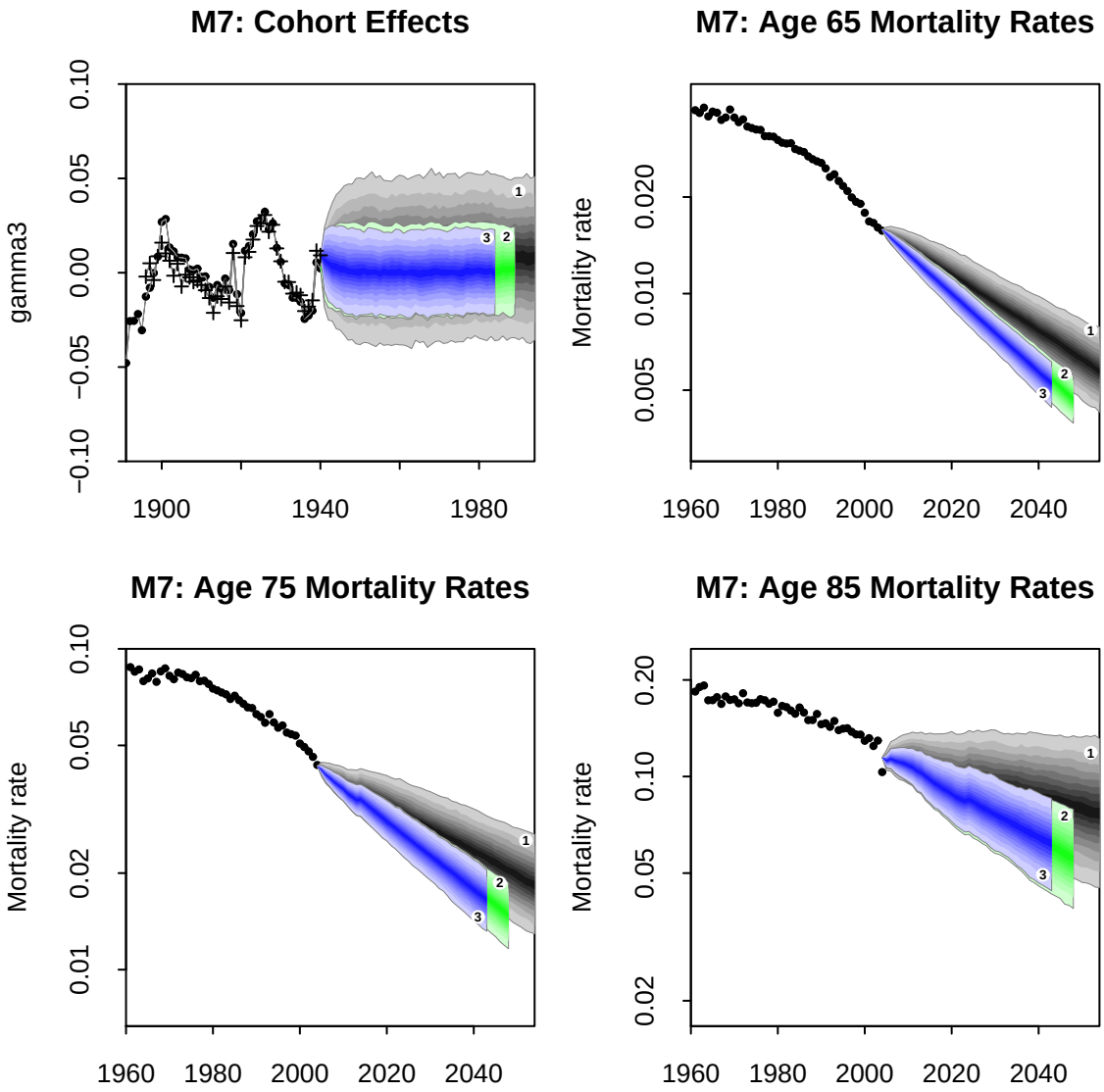

Figure 6: England & Wales, males: Model M7. Cohort effect and mortality rates for ages 65, 75 and 85. Notes: see Figure 4.

backwards from 2004, then the extrapolation starts off below the dots but then reconnects around about 1961. For the scenario 2 and 3 fans, this backwards extrapolation will be approximately aligned with the line connecting the 1981 and 2004 observations.

Since the historical data display an *apparent* change in trend (whether or not this change in trend is genuine, or just the result of statistical variation), it is inevitable that, for all models, fans based on data from 1961 to 2004 will differ from those based on data from 1981 to 2004.

- In most cases, the scenario 1 fans are wider than the scenario 2 and 3 fans, reflecting the greater volatility in mortality rates that can be seen in the years 1961 to 1980. Greater volatility in the mortality data leads to greater volatility in the estimates of the underlying period effects, $\kappa_t^{(i)}$. This, in turn, leads to higher estimates for the variances in the random-walk model for the period effects. Finally, this leads to greater uncertainty in future mortality rates. The scenario 2 and 3 fans draw on estimates of the period effects that cover the less-volatile years.

- For M2B (Figure 4), we can also see very significant differences between the scenario 2 and 3 fans, most obviously at age 85 where there is a clear problem with the scenario 3 fan. The explanation for the implausible shape of the scenario 3 fan at age 85 lies partly with the fitted values for $\beta_x^{(3)}$. Using data from 1961 to 2004, the fitted $\beta_x^{(3)}$ is entirely positive (see Cairns et al., 2009, Figure 4). When we use data from 1981 to 2004 (see Cairns et al., 2009, Figure 4), the fitted $\beta_x^{(3)}$ is very different, taking negative values below age 77 and positive values above (and these are larger in magnitude as well). Figure 4 also shows that $\gamma_c^{(3)}$ is increasing more steeply after year of birth 1925. When this is combined with the negative values for $\beta_x^{(3)}$ up to age 77, this implies falling cohort mortality. But as the post-1925 steepening in $\gamma_c^{(3)}$ feeds through to the higher ages during the *forecasting* period 2004 to 2024, it combines with *positive* values for $\beta_x^{(3)}$ resulting in sharply deteriorating mortality (Figure 4, scenario 3 fans). In contrast, when we use data from 1961 to 2004, since $\beta_x^{(3)}$ is positive at all ages, the post-1925 steepening in $\gamma_c^{(3)}$ means that mortality rates continue to fall at high ages within the forecasting period 2004-2024 (Figure 4, scenario 2 fans). Thus, the finding in Cairns et al. (2009), that changing from 1961-2004 data to 1981-2004 data resulted in substantially different estimates for the age, period and cohort effects has been shown to have a material impact on key forecasts based on this model.

  This lack of stability would appear to be linked to the shape of the likelihood function for model M2 using this dataset. First, the fitting algorithm is generally slow to converge indicating that the likelihood surface is quite flat in some dimensions. Second, we investigated (but do not report here in detail) how the parameter estimates evolve when we add one calendar year's data at a time.

Occasionally, we see that the parameter values jump to a set of values that are qualitatively quite different from the previous year's estimates, confirming that the likelihood function has multiple maxima with significantly different parameter values. It therefore seems likely that the scenario 3 fan relates to one maximum and the scenario 2 fan to another. Similar problems with the robustness of M2 have been reported by CMI (2007), Cairns et al. (2009), Plat (2009), Dowd et al. (2010a,b) and Debonneuil (2010).

So we can conclude that for the dataset under consideration and for this implementation of M2, the forecasts are not robust relative to how much historical data are used. Further remarks on alternative methods for calibrating M2 come at the end of this section.

- For M7 (Figure 6), the fans look stable. In particular, the scenario 2 and 3 fans are very similar in terms of trajectory and spread. The greater spread of the scenario 1 fans reflects a greater volatility in the $\kappa_t^{(i)}$ prior to 1981. Cairns et al. (2009, Figure 8) had indicated that M7 appeared to be stable relative to the period of data employed. The results here reinforce this conclusion.

  We can see that the scenario 1 mortality fans also have a different mean trajectory from the scenario 2 and 3 fans. However, we consider this to be 'normal' variation given the changing trends in the data.

  For M1, M3B, M5, M8A and M8B, we can come to similar conclusions as M7 for the EW males 1961-2004 and 1981-2004 datasets. In the case of M3B (Figure 5), the fitted and projected cohort effects appears to lack robustness. However, this is simply a result of differences in the identifiability constraint (see Appendix A.1), which has no impact on forecasts of mortality rates.

In summary, for the dataset used here, it appears that M1, M3, M5, M7 and M8 all appear to be reasonably robust relative to the historical data used. However, for M8, see Section 6. M2B forecasts, in contrast, look to be unstable, at least using the present calibration method using full maximum likelihood (as in Cairns et al., 2009).

In the course of our investigations we did also consider the original approach described by Renshaw and Haberman (2006) in which $\beta_x^{(1)}$ is constrained to be equal to $\beta_x^{(1)}$ under the Lee-Carter model (M1). This method was still found to lack robustness, although the jumps between qualitatively different sets of paramater estimates were less frequent than under our method. More recently, however, Butt and Haberman (2010) report on a further variant on the estimation method. They impose the additional constraints that $\beta_x^{(2)}$ and $\beta_x^{(3)}$ be positive, and report that, for EW males data ages 60-89 and 0-89, this produces the consistent and stable parameter estimates using data from 1961 up to 1997, 1999, 2001, 2003, 2005 and 2007. So, for modellers wishing to pursue the use of M2, the Butt and Haberman (2010) variant seems promising.

5 Applications: Survivor index and annuity price

In this section, we switch our attention from forecasts of the underlying mortality rates, $q(t, x)$, to two "derivative" quantities that utilise these forecasts. The first of these is a survivor index, and the second is the price of an annuity (which is, in turn, derived from the survivor index). Forecasts of these will provide additional evidence of possible model risk.

Figure 7 shows the fan charts produced by each model of the future value of the survivor index $S(t, 65)$; this measures the proportion from a group of males aged 65 at the start of 2005 who are still alive at the start of 2005+t. Note that $S(t, 65)$ requires no forecasts of the cohort effect as the relevant cohort effect, $\gamma^{(3)}_{1940}$, is known at the start of 2005. As a consequence, models M2A and M2B produce identical results. The same applies to M3 and M8.

The fans for M1, M2B, M3B, M5, M7 and M8B are superimposed in Figure 7 to aid comparison. There are some differences between the trends and more significant differences between the dispersions: for example, the M7 fan is much wider than the M2B and M3B fans. Again, therefore, model risk, even amongst the more plausible models, cannot be ignored.

The survivor index can be used to calculate the present value of a term annuity payable annually in arrears for a maximum of 25 years to a male aged 65 at the start of 2005. The price is equal to the present value of the survivor index, which, assuming a constant interest rate, is given by:

$$P = \sum_{t=1}^{25} v^t S(t, 65)$$

where v is the discount factor. If we assume a rate of interest of 4% per annum, then the simulated empirical distribution function of P under each of the nine models is plotted in Figure 8. We can see that there are only moderate differences between the models. (see Table 3). This is an interesting finding: although the models can give quite different mortality forecasts, these differences can be attenuated when used in applications.

The analysis was repeated for 2% and 10% interest, and we found that the broad conclusion above was robust: there were only moderate differences between models, with minor changes in the relative positions of the six distributions of the present value.

The calculations were repeated for the present value of a term annuity payable annually in arrears for a maximum of 30 years to a male aged 60 at the start of 2005:

$$P = \sum_{t=1}^{30} v^t S(t, 60).$$

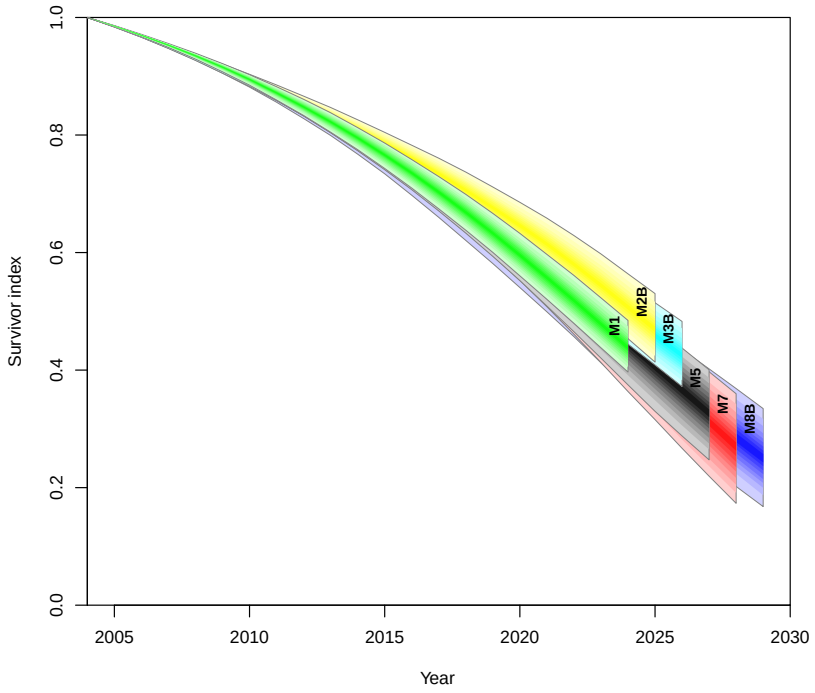

Figure 7: England & Wales, males: Fan charts for the survivor index $S(t, 65)$ for the cohort aged 65 at the start of 2005, for models M1 (uppermost fan), M2B (second top fan), M3B (third), M5 (fourth), M7 (fifth) and M8B (sixth and rearmost fan).

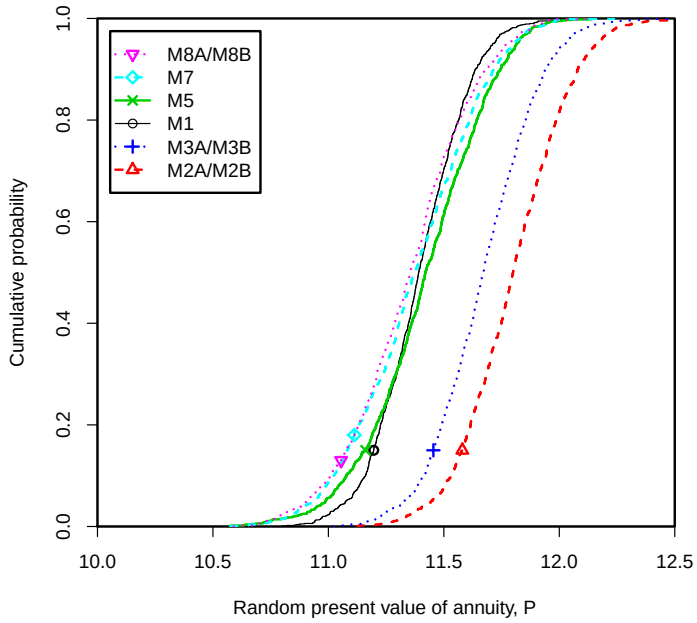

Figure 8: England & Wales, males: Cumulative distribution function of the present value of an annuity payable annually in arrears for a maximum of 25 years to a male aged 65 at the start of 2005, assuming a rate of interest of 4% per annum.

The general conclusions from this additional experiment are much the same as for the age 65 cohort. However, we can make the additional observation that the choice of model for the cohort effect under models M2, M3 and M8 has only a moderate impact on the value of an annuity at age 60.

Model	Mean	St. Dev.	Coefficient of variation
M1	11.396	0.195	1.72%
M2A/M2B	11.802	0.22	1.86%
M3A/M3B	11.673	0.213	1.83%
M5	11.418	0.256	2.24%
M7	11.373	0.267	2.34%
M8A/M8B	11.348	0.26	2.29%

Table 3: England & Wales, males: Mean, standard deviation and coefficient of variation (the standard deviation divided by the mean) of the random present value $P = \sum_{t=1}^{25} v^t S(t, 65)$.

Model	Mean	St. Dev.	Coefficient of variation
M1	13.428	0.222	1.65 %
M2A	13.804	0.26	1.89 %
M2B	13.612	0.340	2.50 %
M3A	13.648	0.257	1.88 %
M3B	13.582	0.257	1.89 %
M5	13.427	0.263	1.96 %
M7	13.331	0.276	2.07 %
M8A	13.393	0.272	2.03 %
M8B	13.312	0.276	2.07 %

Table 4: England & Wales, males: Mean, standard deviation and coefficient of variation of the random present value $P = \sum_{t=1}^{30} v^t S(t, 60)$.

6 Results for US males

In this section, we report briefly on a repeat analysis of US males data from 1968 to 2003. (For a more detailed discussion, see Cairns *et al.*, 2008a.) Our aim in this repeat analysis is to see if the conclusions that we have drawn in Sections 3 to 6 are specific to the England & Wales males dataset or if they might apply more generally to the US population for the same age range and gender.

Much of our analysis threw up similar results to those in earlier sections:

- Models M1, M3, M5 and M7 all continued to produce plausible forecasts.
- Models M1, M3, M5 and M7 all continued to produce similar robust forecasts. For example, the US equivalent of Figure 3 produced a similar set of results for these models, except that the fans for the different models were slightly more spread out.
- Model M2 continued to exhibit robustness problems.

However, our analysis of the final model, M8, revealed some dangers associated with the use of this model. For the remainder of this section, therefore, we focus on model M8 which generates such different results compared with EW males data that we question the validity of M8 for this dataset.

Cairns et al. (2009) noted that, when M8 was applied to US data, projections of mortality rates even for cohorts born before 1943 looked implausible, with mortality rates increasing rather than continuing to fall. Sensitivity tests suggest that the downturn in the fitted $\gamma_c^{(3)}$ around 1920 (see Figure 9, bottom) causes the mortality improvements at ages 75 and 84 (rates at age 85 were not available prior to 1980) to go into reverse, until the 1920 to 1940 fitted cohort effects have worked their way through. It is possible, although unlikely, that this is a genuine effect. A much more likely explanation is that M8 lacks the necessary factors to fit what are age-period effects adequately, and that it compensates for this by overfitting the cohort effect with implausible consequences. A related point concerning M2A and M8A was discussed in Section 3. In this case, however, the lack of a second or third age-period component had less serious consequences.

For the US data, a random-walk process with drift fits better than a stationary AR(1) process (with $\alpha < 1$) around a linear trend (indeed our estimation package struggled to fit any stationary ARIMA model).

Results for model M8A with α fixed at 0.9999 (in effect, a random-walk model) are shown in Figure 9, and these confirm that M8 produces some rather strange mortality forecasts at higher ages. The sharp increase in mortality rates at ages 75 and 84 up to 2014 and 2023, respectively, is solely due to the estimated values of $\gamma_c^{(3)}$ and does not depend on the form of model used to explain the future cohort effect.

The change in direction of the fans (for example, around 2014 for the age 75 fan) corresponds to a change in direction of the $\gamma_c^{(3)}$ process that occurs around 1940 (at the beginning of the projection period: see Figure 9, bottom).

Model fitting and projection under M8A were carried out using some alternative historical data periods (similar to the EW robustness analysis) and these showed that the model produced similar, implausible projections.

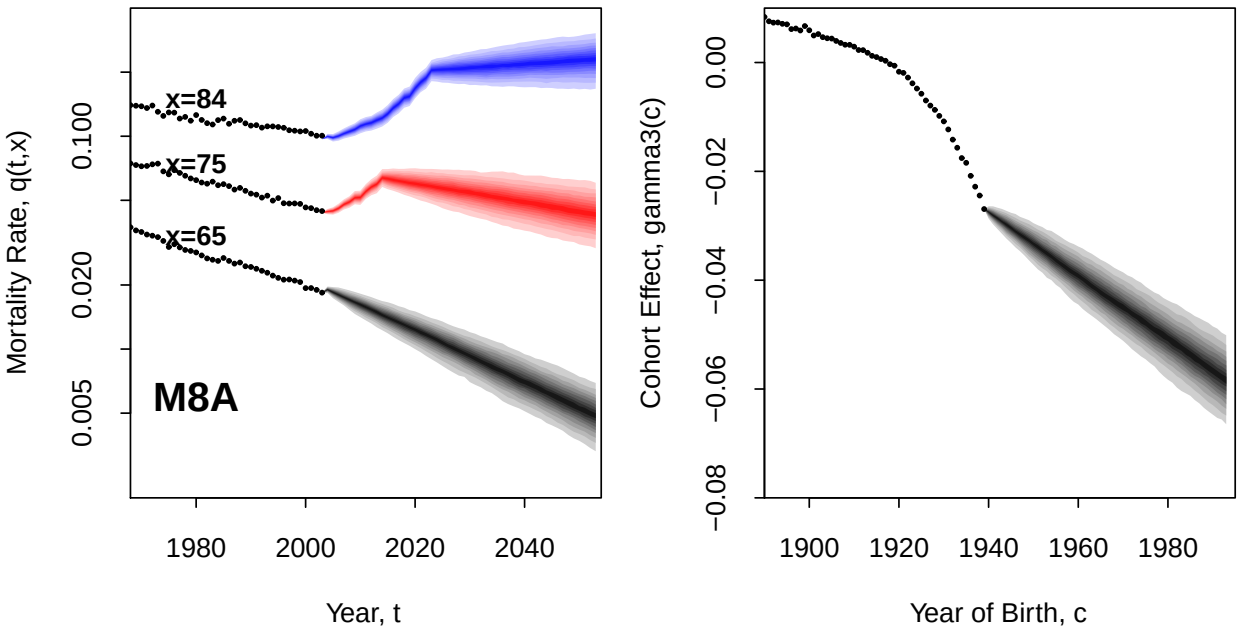

Figure 9: US, males: Model M8A. Left: Fan charts for mortality rates at ages 65, 75 and 84 with the autoregressive parameter set to $\alpha = 0.9999$. Right: Fan charts for the cohort effect, $\gamma_c^{(3)}$, under model M8A with the autoregressive parameter set to $\alpha = 0.9999$.

7 Conclusions

One of the main lessons from this investigation into forecasting with stochastic mortality models is the danger of ranking and selecting models purely on the basis of how well they fit historical data: it is quite possible for a model to give a good fit to the historical data, and still give inadequate forecasts. We propose here new qualitative criteria that focus on a model's ability to produce plausible forecasts: biological reasonableness of forecast mortality term structures, biological reasonableness of individual stochastic components of the forecasting model (for example, the cohort effect), reasonableness of forecast levels of uncertainty relative to historical levels of uncertainty; and robustness of forecasts relative to the sample period used to fit the model.

Had we only considered the quality of fit using historical data, we would have chosen model M8 for modelling EW males mortality, since it had the highest BIC amongst the 8 models we have examined (Cairns et al. (2009, Table 3)). Model M8 is a particular extension of the CBD class of models allowing for a cohort effect, and was, in fact, specifically designed to fit the historical data well. It was also designed to satisfy a range of qualitative criteria, such as ease of implementation, parsimony, and robustness of parameter estimates relative to the period of data employed. However, when the model was used for forecasting, the forecasts for US males were so implausible that M8 can be dismissed as an acceptable model for this specific dataset on this ground alone.

M2 had also been found to fit historical data well (Cairns et al. 2009). However, at least in the way that it has been implemented here, M2 lacks robustness in its forecasts. Other implementations or extensions of M2 might be more stable.

On the basis of the additional forecast-related criteria, we found that for the datasets considered here:

- Ignoring parameter uncertainty, the Lee-Carter model, M1, produces forecasts at higher ages that are 'too precise': that is, having too little uncertainty relative to historical volatility as well as predicted uncertainty at lower ages. This problem was not evident from simply estimating the parameters of the models, but only became apparent when the models are used for forecasting.
- Model M3 performed in a satisfactory way. It produces biologically plausible results and seems to be a robust model.
- Models M5 and M7 both performed well in the forecasting experiments in this paper. Both produce biologically plausible results and seem robust.

We started in Cairns et al. (2009) with eight possible stochastic mortality models. Fitting the models to historical data and assessing the results against a set of quantitative and qualitative model-fitting criteria allowed us to reduce this number to

six. Examining the forecasts produced by these models and assessing them against a set of qualitative forecast-related criteria has enabled us to further assess their suitability for a particular dataset and forecasting application. We would recommend a similar methodology be conducted to identify suitable forecast models for other datasets of interest since results and conclusions are likely to vary by gender, age range and nationality. Finally, in addition to analysing the *ex ante* forecasting performance of stochastic mortality models, it is important to examine the related, but distinct, issue of their *ex post* forecasting performance. This issue of backtesting forecast performance is addressed in a companion piece (Dowd et al., 2010b).

Acknowledgements

The authors would like to thank an anonymous referee for his/her helpful comments.

References

Bauwens, L., and Sucarrat, G. (2008) "General to specific modelling of exchange rate volatility: A forecast evaluation", Economics Department Working Paper 08-18, Universidad Carlos III de Madrid.

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

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

Booth, H., Maindonald, J., and Smith, L. (2005) "Evaluation of the variants of the Lee-Carter method of forecasting mortality: A multi-country comparison", *New Zealand Population Review*, 31: 13-34.

Bradford-Hill, A. (1965) The environment and disease: association or causation. *Proceedings of the Royal Society of Medicine*, 58: 295-300.

Brouhns, N., Denuit, M., and Vermunt, J.K. (2002) "A Poisson log-bilinear regression approach to the construction of projected life tables", *Insurance: Mathematics and Economics,* 31: 373-393.

Butt, Z., and Haberman, S. (2010) "A comparative study of parametric mortality projection models" Actuarial Research Paper No. 196, Cass Business School.

Cairns, A.J.G., Blake, D., and Dowd, K. (2006a) "Pricing death: Frameworks for the valuation and securitization of mortality risk", *ASTIN Bulletin,* 36: 79-120.

Cairns, A.J.G., Blake, D., and Dowd, K. (2006b) "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., Coughlan, G.D., D. Epstein, and Khalaf-Allah, M. (2008a) "Mortality density forecasts: An analysis of six stochastic mortality models", Pensions Institute Discussion Paper PI-0801.

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

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A., and 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.

Campos, J., Ericsson, N. R., and Hendry, D. F. (2005) "General-to-specific modeling: An overview and selected bibliography", in J. Campos, D. F. Hendry, and N. R. Ericsson (Eds.), General-to-Specific Modeling, Volume 1. Cheltenham: Edward Elgar Publishing.

Continuous Mortality Investigation (CMI) (2005) "Projecting future mortality: Towards a proposal for a stochastic methodology", Working paper 15.

Continuous Mortality Investigation (CMI) (2006) "Stochastic projection methodologies: Further progress and P-Spline model features, example results and implications", Working paper 20.

Continuous Mortality Investigation (CMI) (2007) "Stochastic projection methodologies: Lee-Carter model features, example results and implications", Working paper 25.

Currie, I.D., Durban, M. and Eilers, P.H.C. (2004) "Smoothing and forecasting mortality rates", *Statistical Modelling,* 4: 279-298.

Czado, C., Delwarde, A., and Denuit, M. (2005) "Bayesian Poisson log-linear mortality projections", *Insurance: Mathematics and Economics* 36: 260-284.

Debonneuil, E. (2010) A simple model of mortality trends aiming at universality: Lee Carter + cohort. Quantitative Finance Papers 1003.1802, arXiv.org.

De Jong, P., and Tickle, L. (2006) "Extending the Lee-Carter model of mortality projection", *Mathematical Population Studies,* 13:1-18.

Delwarde, A., Denuit, M., and Eilers, P. (2007) "Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: A penalised log-likelihood approach", *Statistical Modelling*, 7: 29-48.

Dowd, K., Blake, D., Cairns, A.J.G., Coughlan, G.D., Epstein, D., and Khalaf-Allah, M. (2010a) "Evaluating the goodness of fit of stochastic mortality models", *Insurance: Mathematics and Economics*, 47: 255-265.

Dowd, K., Blake, D., Cairns, A.J.G., Coughlan, G.D., Epstein, D., and Khalaf-Allah, M. (2010b) "Backtesting stochastic mortality models: An ex-post evaluation of multi-period-ahead density forecasts", *North American Actuarial Journal*, 14: 281-298.

Hyndman, R.J., and Ullah, M.S. (2007) "Robust forecasting of mortality and fertility rates: A functional data approach", *Computational Statistics and Data Analysis*, 51: 4942-4956.

Lee, R.D., and Carter, L.R. (1992) "Modeling and forecasting U.S. mortality", *Journal of the American Statistical Association,* 87: 659-675.

Lee, R., and Miller, T. (2001) "Evaluating the performance of the Lee-Carter method for forecasting mortality", *Demography*, 38: 537-549.

Li, J.S.-H., Hardy, M.R., and Tan, K.S. (2009) "Uncertainty in model forecasting: An extension to the classic Lee-Carter approach", *ASTIN Bulletin,* 39: 137-164.

McNown, R.F., and Rogers, A. (1989) "Forecasting mortality: A parametrized time series approach", *Demography*, 26: 645-660.

Plat, R. (2009) On stochastic mortality modelling. *Insurance: Mathematics and Economics*, 45: 393-404.

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

Willets, R.C. (2004) "The Cohort Effect: Insights and Explanations". *British Actuarial Journal*, 10: 833-877.

A Historical parameter estimation and identifiability constraints

A.1 Identifiability constraints

By way of example, consider M1. Suppose we multiply all of the historical $\kappa_t^{(2)}$'s by a constant b, and simultaneously divide all of the $\beta_x^{(2)}$'s by b. Rescaling the parameters of M1 in this way has no impact on the fitted $\hat{q}(t,x)$. Furthermore, when we refit the random walk model, the simulated future $\kappa_t^{(2)}$'s are obviously different, but the simulated future $q(t,x)$'s are not. In addition to rescaling of the $\kappa_t^{(2)}$'s, they can also be shifted with the same consequences.

The possibility of rescaling and shifting the parameters with no effect on the $q(t,x)$ means that we have an identifiability problem, which we tackle by introducing identifiability constraints. Besides M1, identifiability constraints are also required for M2, M3, M7 and M8.

The identifiability constraints we have used in this paper are as follows:

- M1
$$\sum_t \kappa_t^{(2)} = 0, \quad \text{and} \quad \sum_x \beta_x^{(2)} = 1.$$

- M2
$$\sum_t \kappa_t^{(2)} = 0, \quad \sum_x \beta_x^{(2)} = 1, \quad \sum_{x,t} \gamma_{t-x}^{(3)} = 0, \quad \text{and} \quad \sum_x \beta_x^{(3)} = 1.$$

- M3
$$\sum_t \kappa_t^{(2)} = 0, \quad \sum_{x,t} \gamma_{t-x}^{(3)} = 0,$$

  and a third constraint that tilts the $\beta_x^{(1)}$'s to keep the overall shape of the fitted $\beta_x^{(1)}$'s as close as possible to the historical average of the log death rate at age x.

- M7
$$\sum_{c=c_0}^{c_1} \gamma_c^{(4)} = 0, \quad \sum_{c=c_0}^{c_1} c\gamma_c^{(4)} = 0, \quad \text{and} \quad \sum_{c=c_0}^{c_1} c^2\gamma_c^{(4)} = 0$$

  where c_0 and c_1 are first and last years of birth that we fit the cohort effect to.

  The first two constraints are designed to facilitate the fitting of a mean-reverting model to $\gamma_c^{(4)}$, since the resulting estimates have no linear drift over the full historical period.

- M8

$$\sum_{x,t} \gamma^{(3)}_{t-x} = 0.$$

For further details, see Cairns et al. (2009).

A.2 Parameter estimation

We provide here some brief remarks on how the $\beta^{(i)}_t$'s, $\kappa^{(i)}_t$'s and $\gamma^{(i)}_c$'s are estimated. The approach to estimation is the same for each model. By way of example, therefore, consider model M3: $\log m(t,x) = \beta^{(1)}_x + n_a^{-1}\kappa^{(2)}_t + n_a^{-1}\gamma^{(3)}_{t-x}$. Our objective is to maximise the Poisson likelihood of the historical data over all of the age, period and cohort effects (see, for example, Cairns et al., 2009). (For an alternative to the Poisson model, see Li et al. (2009).)

We use an iterative scheme which proceeds as follows. Within each iteration:

- Update each of the $\beta^{(1)}_x$'s in turn, using a single-step of a Newton-Raphson algorithm using the first and second partial derivatives of the likelihood with respect to $\beta^{(1)}_x$. For a given x, this amounts to increasing the likelihood over age x cells only. The likelihood for all other ages is unaffected.
- Update each of the $\kappa^{(2)}_t$'s in turn, using a single-step of a Newton-Raphson algorithm. For a given t, this amounts to increasing the likelihood over calendar year t cells only. The likelihood for all other calendar years is unaffected.
- Update each of the $\gamma^{(3)}_c$'s in turn, using a single-step of a Newton-Raphson algorithm. For a given c, this amounts to increasing the likelihood over cells, (t,x), that have a common year of birth, $t-x=c$, only. The likelihood for all other cohort years of birth is unaffected.
- Apply the identifiability constraints (Section A.1).

The iterative scheme is repeated until the log-likelihood converges to within a specified degree of tolerance.

For further details of the Newton-Raphson updating, see, for example, Brouhns et al. (2002) or Renshaw and Haberman (2006).

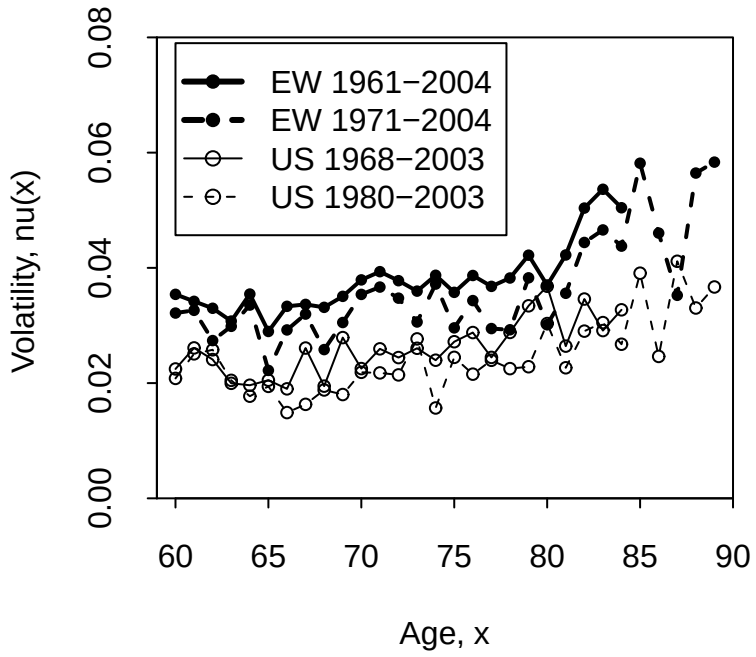

Figure 10: Historical empirical volatilities of death rates.

B Volatility of death rates

Here we calculate empirical volatilities for historical mortality rates. Let $m(t, x)$ be the crude death rates for year t and age x, and define $\delta(t, x) = \log m(t, x) - \log m(t-1, x)$. We define the volatility at age x, $\nu(x)$, to be the empirical standard deviation of $\delta(2, x), \ldots, \delta(T, x)$. Volatilities were calculated for ages 60 to 84/89 for different ranges of years and countries, and are plotted in Figure 10. It can be seen that in all four cases plotted, the volatility function is reasonably flat across most ages, but rising slightly at the higher ages.