DISCUSSION PAPER PI-0801

Mortality Density Forecasts: An Analysis of Six Stochastic Mortality Models

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

April 2008

ISSN 1367-580X

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

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

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.

Mortality Density Forecasts: An Analysis of Six Stochastic Mortality Models

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

April 2008

Abstract

We investigate the uncertainty of forecasts of future mortality generated by a number of previously proposed stochastic mortality models. We specify fully the stochastic structure of the models to enable them to generate forecasts. Mortality fan charts are then used to compare and contrast the models, with the conclusion that model risk can be significant.

The models are also assessed individually with reference to three criteria that focus on the plausibility of their forecasts: biological reasonableness of forecast mortality term structures; biological reasonableness of individual stochastic components of the forecasting model (for example, the cohort effect); and reasonableness of forecast levels of uncertainty relative to historical levels of uncertainty. In addition, we consider a fourth assessment criterion dealing with the robustness of forecasts relative to the sample period used to fit the model.

To illustrate the assessment methodology, we analyse a data set consisting of national population data for England & Wales, for Males aged between 60 and 90 years old. We note that this particular data set may favour those models designed for application to older ages, such as variants of Cairns-Blake-Dowd, and emphasise that a similar analysis should be conducted for the specific data set of interest to the reader. We draw some conclusions based on the analysis and compare to the application of the models for the same age group and gender for the United States population. Finally, we note the broader application of the approach to model selection for alternate data sets and populations.

Keywords: Stochastic mortality model, cohort effect, 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] Centre for Risk & Insurance Studies, Nottingham University Business School, Jubilee Campus, Nottingham, NG8 1BB, United Kingdom.

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

1 Introduction

A range of different stochastic mortality models have emerged over the last fifteen years: e.g., Lee and Carter (1992), Renshaw and Haberman (2006), Cairns, Blake and Dowd (2006b, hereafter denoted CBD, and 2008), Cairns et al. (2007, sections 4.6-4.8), and Delwarde, Denuit and Eilers (2007). They share a common feature in that they are all time series models with parameters that are estimated from historical mortality rates. They also have some key differences. Some models build in an assumption of smoothness in mortality rates between ages (e.g. Cairns et al, 2006, and Delwarde et al, 2007) in any given year, while others allow for roughness (e.g. Lee-Carter). In contrast, Currie et al (2004) assume smoothness in both the age and time dimensions through the use of P-splines. Some models have dynamics that are driven by just one source of randomness (e.g. Lee-Carter), while others have several sources (e.g. the model proposed by Cairns et al. 2007 – here labelled M7 – has four). Some researchers extend earlier models to allow for more-recently-recognised phenomena, such as cohort effects (e.g., Renshaw and Haberman (2006), Cairns et al. (2007, sections 4.6-4.8)).

A number of studies have sought to draw out more formal comparisons between various models. CMI (2005, 2006, 2007), for example, compared the Lee-Carter and P-splines models. Cairns et al. (2007) focused on quantitative and qualitative comparisons of the eight models listed in Table 1, 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.

They found that some models fared better under some criteria than others, but that no single model could claim superiority under all the criteria considered. One implication of this is that there remains a large number of potentially valid stochastic mortality models, despite significant conceptual differences between them. Another implication is that model choice depends on what priority the model user attaches to each of the assessment criteria.

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	$\log m(t,x) = \beta_x^{(1)} + n_a^{-1}\kappa_t^{(2)} + n_a^{-1}\gamma_{t-x}^{(3)}$
M4	$\log m(t,x) = \sum_{i,j} \theta_{ij} B_{ij}^{ay}(x,t)$
M5	$\text{logit } q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x})$
M6	$\text{logit } q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \gamma_{t-x}^{(3)}$
M7	$\text{logit } q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \kappa_t^{(3)}\left((x-\bar{x})^2 - \hat{\sigma}_x^2\right) + \gamma_{t-x}^{(4)}$
M8	$\text{logit } q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \gamma_{t-x}^{(3)}(x_c - x)$

Table 1: Formulae for the eight mortality models considered by Cairns et al. (2007): The functions $\beta_x^{(i)}$, $\kappa_t^{(i)}$, and $\gamma_{t-x}^{(i)}$ are age, period and cohort effects, respectively. The $B_{ij}^{ay}(x,t)$ are B-spline basis functions and the θ_{ij} are weights attached to each basis function. $\bar{x}$ 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.

In this study, we describe a set of procedures that can be used to explore forensically and diligently the appropriateness of the forecast models for a chosen data set. We consider additional assessment criteria that allow us to examine the *ex ante* plausibility of the forecasts generated by the stochastic mortality models, illustrating with national population data for England & Wales, and separately, the United States, for an age group consisting of 60-89 year old Males. Further work should be undertaken to look at the related, but distinct, issue of the *ex post* forecasting performance (i.e. backtesting) of stochastic mortality models (see Dowd et al., 2008a,b).

We will concentrate on just six of the models discussed by Cairns et al. (2007): 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. (2007) as the best fitting, in terms of BIC, of the eight models considered on the basis of male mortality data from England & Wales and the US for the age group under consideration. M2 is the Renshaw and Haberman (2006) extension[1] of the original Lee-Carter model

[1] We consider here, a version of the Renshaw and Haberman (2006) model, M2, discussed by

(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, although they fit the historical data less well, they provide useful benchmarks for comparison with the four models involving cohort effects M2, M3, M7 and M8. Models M4 and M6 are not considered any further in this study because of their low BIC and qualitative rankings for these dataset in Cairns et al. (2007, Table 3). Although M3 is a special case of M2, we include it here for two reasons. First, it had a relatively high BIC ranking for the US data. Second, it avoids the problem with the robustness of parameter estimates for M2 identified by Cairns et al. (2007).

There are three aspects to this study. First, we specify the stochastic structure of the models to enable them to generate forecasts of mortality rates, determine central projections and judge the uncertainty inherent in each model.

Second, we utilise the following assessment criteria to evaluate the plausibility and robustness of the mortality forecasts produced by each model:

- biological reasonableness of the forecast mortality term structures;
- biological reasonableness of individual stochastic components of each model (for example, the cohort effect);
- reasonableness of forecast levels of uncertainty relative to historical levels of uncertainty;
- robustness of forecasts with respect to the time period used to fit the model.

Third, we discuss model risk as a complement to the discussion in Cairns, Blake and Dowd (2006b) on parameter uncertainty. Our purpose is to determine whether or not the choice of model has a material impact on forecasts of key variables of interest, especially mortality rates.

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 models M1, M2, M3, M5, M7 and M8. Results for the different models using England & Wales male mortality data are compared and contrasted in Section 3. Section 4 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 5 and this is augmented in Section 6 by a sensitivity analysis of the forecasts to changes in key parameters in a fully specified stochastic model.

Cairns et al. (2007) which has problems with the stability of parameter estimates and projections for this dataset. In this study, we do not examine alternative versions of this model and note that other specifications of or extensions to this model might resolve the stability problem identified herein.

Finally, in Section 7 and in an Appendix we repeat the analysis for US male mortality data: our aim here is to draw out features of the US data that are distinct from the England & Wales data. In Section 8 we conclude.

2 Forecasting with stochastic mortality models

In this section, we take six stochastic mortality models which, on the basis of fitting to historical data, appear to be suitable candidates for forecasting future mortality for the age group under consideration (that is, 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 $q(t, x) = 1 - \exp[-m(t, x)]$. All the models considered are of the form (see M1, M2, M3, M5, M7 and M8 in Table 1):

$$\begin{aligned} \log m(t,x) &= \sum_{i=1}^{N} \beta_x^{(i)} \kappa_t^{(i)} \gamma_{t-x}^{(i)} \quad \text{(models M1, M2 and M3)}, \\ \text{or} \ \ \operatorname{logit} q(t,x) = \log \frac{q(t,x)}{1-q(t,x)} &= \sum_{i=1}^{N} \beta_x^{(i)} \kappa_t^{(i)} \gamma_{t-x}^{(i)} \quad \text{(models M5, M7 and M8)}, \end{aligned}$$

where $\beta_x^{(i)}$ is an age effect, $\kappa_t^{(i)}$ a period effect, and $\gamma_{t-x}^{(i)}$ a cohort effect (see Cairns et al., 2007).

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., 2002, Czado et al., 2005, and de Jong and Tickle, 2006).

Following Cairns, Blake and Dowd (2006b), we use a multivariate random walk with drift to drive the dynamics of the period effect. This model appears to be consistent with the data (see the plots of the $\kappa_t^{(i)}$ in Cairns et al. (2007)). 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).

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. In Figure 1 (right-hand column), we plot the fitted values of the cohort effect for M2 ($\gamma_{t-x}^{(3)}$), M3 ($\gamma_{t-x}^{(3)}$), M8 ($\gamma_{t-x}^{(3)}$) and M7 ($\gamma_{t-x}^{(4)}$), where $t - x$ is the cohort year of birth (see Cairns et al., 2007).[2] From these plots, we can see that a simple random-walk process is unlikely to be appropriate and, in the sub-sections that

[2] The left-hand plots in the figure show the corresponding age effect for each model's age-cohort component.

follow, we discuss various alternative stochastic processes that might be suitable for the different models. As with previous studies (e.g., Renshaw and Haberman, 2006, and CMI, 2007), we will assume that the cohort effect, $\gamma_{t-x}^{(i)}$, has dynamics that are independent of the period effect, $\kappa_t^{(i)}$.

The age effects, $\beta_x^{(i)}$, are either non-parametric and estimated from historical data (M1, M2 and M3), or assume some particular functional form (M5, M7 and 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 to simulate or extrapolate the age effects.

2.1 Model M1

M1 is the original Lee-Carter (1992) model. It is a two-component model with a single random process, $\kappa_t^{(2)}$, driving all the dynamics. In line with Lee and Carter (1992), and for consistency with the remaining models, we assume that $\kappa_t^{(2)}$ follows a one-dimensional random walk with drift. There is no cohort effect.

2.2 Model M2

M2 is the Renshaw and Haberman (2006) extension to the Lee-Carter model involving a cohort effect. We assume that $\kappa_t^{(2)}$ follows a one-dimensional random walk with drift. Determining the dynamics of the cohort effect (Figure 1, top right panel) is rather more difficult. The observed path of $\gamma_{t-x}^{(3)}$ in M2 has a pronounced hump shape, a path that one would be highly unlikely to observe if it followed a random walk with drift. Furthermore, the path seems relatively smooth around a trend that is gradually changing over time with more pronounced changes in trend around 1900 and 1925. It is not clear how the trend might change in the future. The curve might continue to steepen; on the other hand, it might easily become less steep. The latter possibility is consistent with the results of CMI (2007) which used a wider range of ages than Cairns et al. (2007) to fit the Renshaw and Haberman (2006) model.

2.2.1 Model M2A

To investigate further the dynamics of the cohort effect in M2, we examined a range of ARIMA(p, d, q) processes for $\gamma_{t-x}^{(3)}$ with $d = 0, 1, 2$, $p = 0, 1, 2, 3, 4$ and $q = 0, 1, 2, 3, 4$. The full set of $\gamma_{t-x}^{(3)}$ England & Wales male data run from 1881 through to 1940 with one missing observation in 1886.[3]

[3] The 1886 cohort was excluded from our analysis because it was felt that there were specific problems with the exposure data for this cohort. For further discussion, see Cairns et al. (2007).

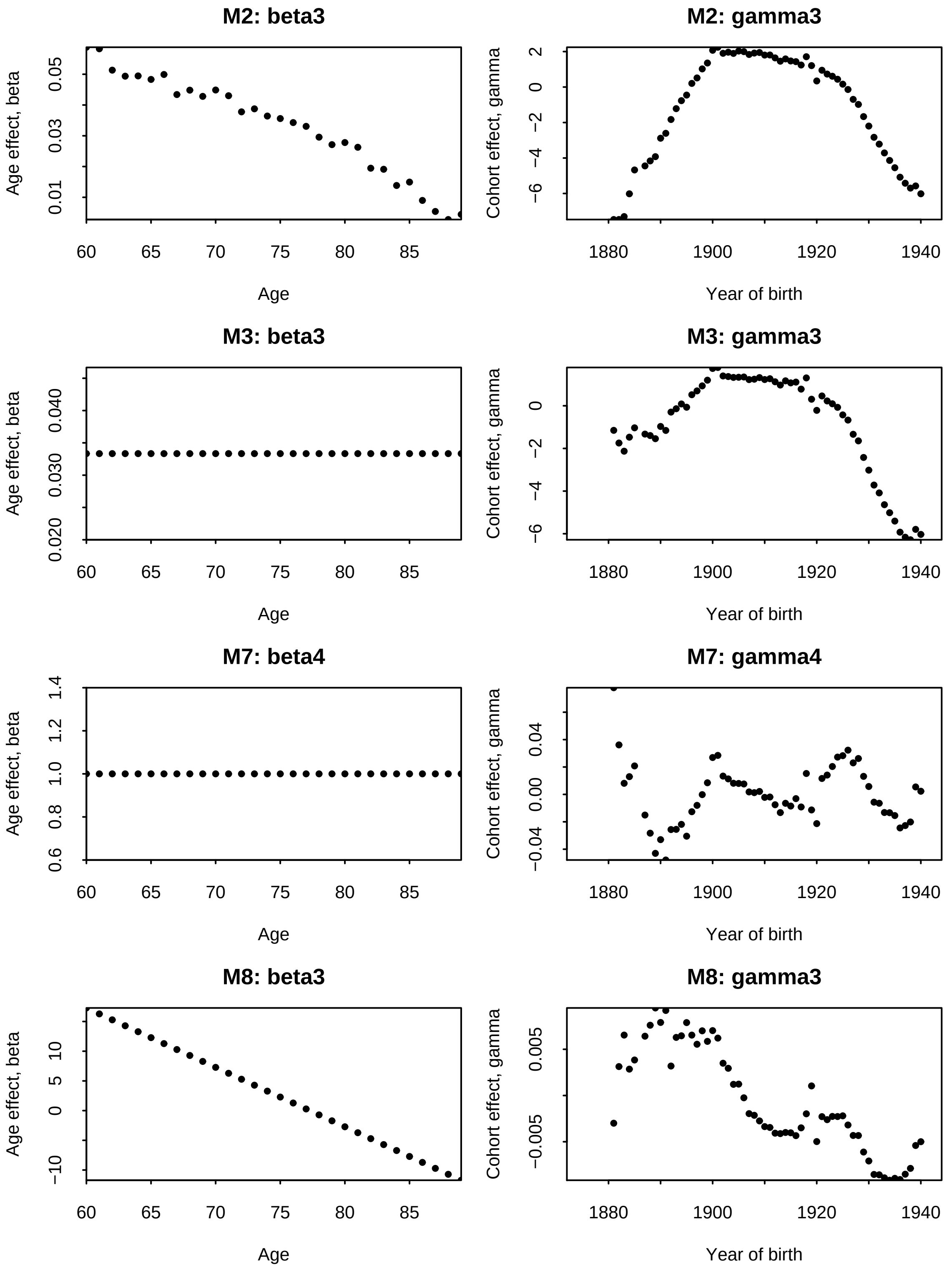


Figure 1: England & Wales, males: Fitted age (beta) and cohort (gamma) effects for models M2, M3, M7 and M8.

Differencing	Processes	BIC
	Optimal processes	
$d = 0$	ARIMA(2,0,2)	-28.8
$d = 1$	ARIMA(1,1,1)	-26.9
$d = 2$	ARIMA(0,2,1)	-24.8
	Suboptimal processes	
$d = 0$	ARIMA(1,0,0)	-42.8
$d = 1$	ARIMA(1,1,0)	-30.2
$d = 2$	ARIMA(1,2,0)	-32.6

Table 2: Bayes Information Criterion (BIC) for various ARIMA processes for $\gamma^{(3)}_{t-x}$ in model M2. The optimal processes are those over the range $p = 0, \ldots, 4$ and $q = 0, \ldots, 4$ for any given level of differencing.

For each level of differencing, $d = 0, 1, 2$, Table 2 shows the model with the highest BIC.[4] The table also shows the BIC values for selected suboptimal models.

One consequence of a second-order ($d = 2$) process is that large positive or negative values in the second differences result in changes in the trend of $\gamma^{(3)}_{t-x}$. A glance at the historical values for $\gamma^{(3)}_{t-x}$ (Figure 1) shows potential changes in trend around 1900 and 1925.

On the basis of Table 2, we chose ARIMA(0,2,1) as the process driving the cohort effect, and we denote this variant of the Renshaw-Haberman model as M2A. Thus we have the process:[5]

$$\Delta^2\gamma^{(3)}_c = \mu^{(3)} + \epsilon_c + \alpha\epsilon_{c-1} \tag{1}$$

where $\epsilon_c \sim N(0, \sigma^2)$. We have assumed that the mean level, $\mu^{(3)}$ is zero.[6] Using data from 1881 to 1940, we estimate $\hat{\alpha} = -0.7453$ and $\hat{\sigma}^2 = 0.1191$ (given $\mu^{(3)}$=0).

Forward simulation requires knowledge of the (latent) value of the residual ϵ_c for

[4] Here we calculate the BIC for the ARIMA(p, d, q) process as $\hat{l} - 0.5(p + q)\log n$ where $\hat{l}$ is the maximum log-likelihood, and n is the number of observations. p and q are the variable numbers of parameters: we have excluded other parameters such as the mean level and the standard deviation which exist in all processes.

[5] Δ is the first difference operator, so that $\Delta\gamma^{(3)}_c = \gamma^{(3)}_c - \gamma^{(3)}_{c-1}$ and $\Delta^2\gamma^{(3)}_c = \Delta\left(\Delta\gamma^{(3)}_c\right) = \gamma^{(3)}_c - 2\gamma^{(3)}_{c-1} + \gamma^{(3)}_{c-2}$.

[6] The inclusion of a non-zero mean, $\mu^{(3)}$, would add a deterministic, quadratic trend to $\gamma^{(3)}_c$, which could then be transformed into an age-period effect that is quadratic in both x and t. Quadratic effects in t seem problematic from a biological point of view, since they imply that there would be an age-period component to the model that accelerates with time. If the relevant age effect (here $\beta^{(3)}_x$) is very small then the combination of this with a quadratic period effect might not cause visible problems in projections out 25 or 50 years, say. Otherwise, we might find that the accelerating quadratic period effect dominates projections in a biologically unreasonable way.

the final cohort year of birth (here $c = t - x = 1940$) to which we have fitted $\gamma_c^{(3)}$.

2.2.2 Model M2B

As an alternative to an ARIMA(0,2,1) process, we considered an ARIMA(1,1,0) process (as employed in CMI, 2007):

$$\Delta\gamma_c^{(3)} = \alpha\Delta\gamma_{c-1}^{(3)} + \sigma\epsilon_c \tag{2}$$

where the ϵ_c are i.i.d. $\sim N(0,1)$. From Table 2, this process fits the historical data less well. However, the difference in BIC values of 5.4 is relatively modest, indicating that an ARIMA(1,1,0) is not an unreasonable choice and we denote this variant of the Renshaw-Haberman model as M2B. The table assumes that the first differences of $\gamma_c^{(3)}$ revert to a zero mean. The fit can be improved further by allowing for reversion to a non-zero mean, although this would then convert into a drift in $\gamma_c^{(3)}$ itself.

2.3 Model M3

M3 is a special case of M2 that assumes the age effects $\beta_x^{(2)}$ and $\beta_x^{(3)}$ are constant and assumed to be equal to 1/(no. of ages) in this study, and we see from Figure 1 that the fitted cohort effect, $\gamma_{t-x}^{(3)}$, is relatively close to that for M2, so we might expect to use similar stochastic models for the cohort effect.

A range of ARIMA processes were fitted to the $\gamma_c^{(3)}$ observations from 1881 to 1940 with BIC values for the optimal models and selected others at each level of differencing reported in Table 3. From this table, we see that we can repeat the conclusions of model M2 and propose the use of the following models:

- M3A: $\gamma_c^{(3)}$ is modelled as an ARIMA(0,2,1) process;
- M3B: $\gamma_c^{(3)}$ is modelled as an ARIMA(1,1,0) process.

2.4 Model M5

M5 is the original two-factor CBD model. The factors $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ are modelled as a 2-dimensional random walk with drift. There is no cohort effect.

2.5 Model M7

M7 is one extension of the CBD model (see Cairns et al., 2007) that allows for a cohort effect. The three factors $\kappa_t^{(1)}, \kappa_t^{(2)}$ and $\kappa_t^{(3)}$ are modelled as a 3-dimensional

Differencing	Processes	BIC
	Optimal processes	
$d = 0$	ARIMA(2,0,1)	-30.6
$d = 1$	ARIMA(1,1,1)	-28.2
$d = 2$	ARIMA(0,2,1)	-27.1
	Suboptimal processes	
$d = 0$	ARIMA(1,0,0)	-33.7
$d = 1$	ARIMA(1,1,0)	-29.7
$d = 2$	ARIMA(1,2,0)	-33.6

Table 3: Bayes Information Criterion (BIC) for various ARIMA processes for $\gamma_{t-x}^{(3)}$ in model M3. The optimal processes are those over the range $p = 0, \ldots, 4$ and $q = 0, \ldots, 4$, for a given level of differencing.

Differencing	Processes	BIC
	Optimal processes	
$d = 0$	ARIMA(2,0,1)	172.4
$d = 1$	ARIMA(0,1,0)	169.5
$d = 2$	ARIMA(0,2,1)	163.0
	Suboptimal models	
$d = 0$	ARIMA(1,0,0)	170.8

Table 4: Bayes Information Criterion (BIC) for various ARIMA models for $\gamma_{t-x}^{(4)}$ in model M7. Optimal models are the optimal models over the range $p = 0, \ldots, 4$ and $q = 0, \ldots, 4$, for a given level of differencing.

random walk with drift.

For England & Wales male data covering the period 1961 to 2004, estimates of the cohort effect, $\gamma_c^{(4)}$ (where $c = t - x$ is the cohort year of birth), can be found in Figure 1 (right middle panel) and in Cairns et al. (2007). We fitted a range of ARIMA(p, d, q) processes and calculated the maximum BIC for three levels of differencing $d = 0, 1, 2$.

From Table 4, we see that the ARIMA(2,0,1) model has the highest BIC with the ARIMA(1,0,0) model (i.e. AR(1)) close behind. Although the BIC already penalises the likelihood function for the number of parameters estimated, we nevertheless opt for the AR(1) process.[7] The simple form of the process driving the cohort effect

[7] The AR(1) process actually dominates when shorter runs of data than the full range cohort years of birth 1881-1940 are considered.

in M7 arises from the three identifiability constraints for M7 (Cairns et al, 2007).[8] Application of these constraints means that the fitted $\gamma_{t-x}^{(4)}$ has no discernible trend or curvature.[9] Instead, these features (trend and curvature) are transferred to the period effects when the identifiability constraints are applied.

2.6 Model M8

M8 is another extension of the CBD model (see Cairns et al., 2007) allowing for a cohort effect. Figure 1 (bottom right panel) shows an apparent downward trend in the fitted values of $\gamma_c^{(3)}$, with significant fluctuations around this trend. It is worth noting that, if we subtract the deterministic linear trend, then the detrended series looks very similar to the $\gamma_c^{(4)}$ series for M7.

We considered two possibilities for modelling the future dynamics of the cohort effect: first, that $\gamma_c^{(3)}$ has no linear trend and, second, that $\gamma_c^{(3)}$ does have a linear trend. For the first case, we fitted a range of ARIMA processes to the raw $\gamma_c^{(3)}$ values. Of these, the ARIMA(1,0,0) (i.e., AR(1)) process had the highest BIC (282.3). For the second case, we used a linear regression to detrend the $\gamma_c^{(3)}$ series before fitting a range of ARIMA processes. The ARIMA(1,0,0) (AR(1)) process again came out top, but with a slightly lower BIC value of 280.2 (due to the penalty from including the additional drift parameter).

In our simulations, we consider two possible variations:

- Model M8A: $\gamma_c^{(3)}$ is modelled as an AR(1) process with drift;
- Model M8B: $\gamma_c^{(3)}$ is modelled as an AR(1) process with no drift.

In M8A, the deterministic drift can be converted into a mixture of age-period effects (which results in adjustments to the $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ estimates) plus a quadratic age effect that is constant in time.[10] This implicit quadratic age-period effect mimics the explicit quadratic age-period effect in model M7 with the restriction that the implicit $\kappa_t^{(3)}$ in M8 is constant.

[8] For further discussion of the relationship (for all models) between identifiability constraints and the stochastic model for the period and cohort effects, see Appendix A.

[9] The estimated $\gamma_c^{(4)}$ will have no discernible linear trend or quadratic curvature; it will simply be a process that fluctuates around zero. This is because the three constraints used by Cairns et al. (2007) mean that if a quadratic function $\alpha_0 + \alpha_1 c + \alpha_2 c^2$ is fitted to the estimated $\gamma_c^{(4)}$ using least squares, the estimates for α_0, α_1 and α_2 will all be zero.

[10] If the trend is $\theta[(t-x)-(\bar{t}-\bar{x})]$ (where $\bar{t}$ is the mean calendar year) then this trend multiplied by $\beta_x^{(3)} = (x_c - x)$ can be separated out into three age-period effects ($\theta(x_c-\bar{x})(t-\bar{t})$, $-\theta(x-\bar{x})(t-\bar{t}-\bar{x}+x_c)$, and $\theta(x-\bar{x})^2$) of which the first two can be incorporated into the existing age-period effects, while the third is an age-effect that is quadratic in age but is not explicitly incorporated into M8.

2.7 Model risk

We end this section with some comments on model risk. Model risk arises in two ways in the current context. On the one hand, it is the risk that we make a decision based on one model that would be different if we had perfect information about the true model and about its parameters (but still no information about future changes in mortality). On the other hand, if we do not have this perfect information, model risk still arises if there is a range of alternative models (all of which are acceptable by our assessment criteria) that generate significantly different forecasts. The latter happens with the models considered here: so a key conclusion from our analysis is that model risk is a significant factor that needs to be considered carefully whenever projections of mortality rates are required.

3 Forecasts and model comparisons

We now proceed to compare the forecasting results for England & Wales for the nine models M1, M2A, M2B, M3A, M3B, M5, M7, M8A and M8B for our chosen dataset. Corresponding results for US males are presented and discussed in Section 7 and Appendix B. To do this, we will present fan charts of the forecasts produced by the models.[11] This will allow us to explore any distinctive visual features of each model, as well as any differences between the models. 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 help us to place a high weight on some models and to question the suitability of others for our purposes.

Cairns et al. (2007) used a range of criteria to compare and assess models and these focused on the within-sample fit of each model. In this section, we add three further criteria that focus on the plausibility of their forecasts: biological reasonableness of the projections of the future term-structure of mortality; biological reasonableness of projected period and cohort effects; and reasonableness of 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 5, 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. (2007) who considered the robustness of parameter estimates.

[11] Fan charts were first proposed for illustrating the output from stochastic mortality models by Dowd, Blake and Cairns (2007).

An examination of Figures 2 to 7 reveals the following:

- Figure 2 shows fan charts for the cohort effects for each model.[12] Amongst these, 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. However, we would suggest that the latter are not so implausible as to rule out either model at this stage.

  The differences between the fan charts for M8A and M8B reflect differences in the trend in $\gamma_c^{(3)}$ (which 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 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, but this cannot be fully established until we focus on key output variables.

- In Figure 2, M2A, M3A and M8A all incorporate a linear trend. As remarked earlier (Footnote 10), 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 an age-cohort component, with a further requirement that the cohort effect has no drift.[13]

- Figure 3 allows us to make an interesting comparison between model M1, on 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. The contrasting result for M1 occurs because it has a single stochastic period effect, $\kappa_t^{(2)}$. The widths of the fans[14] is proportional to the age effect, $\beta_x^{(2)}$, and with M1 (see Cairns

[12] M1 and M5 are not plotted since they have no cohort effect.

[13] We do not consider such an extension in this paper.

[14] Under model M1, the standard deviation of $\log m(t,x)$ is $\beta_x^{(2)}\sqrt{Var[\kappa_t^{(2)}]}$.

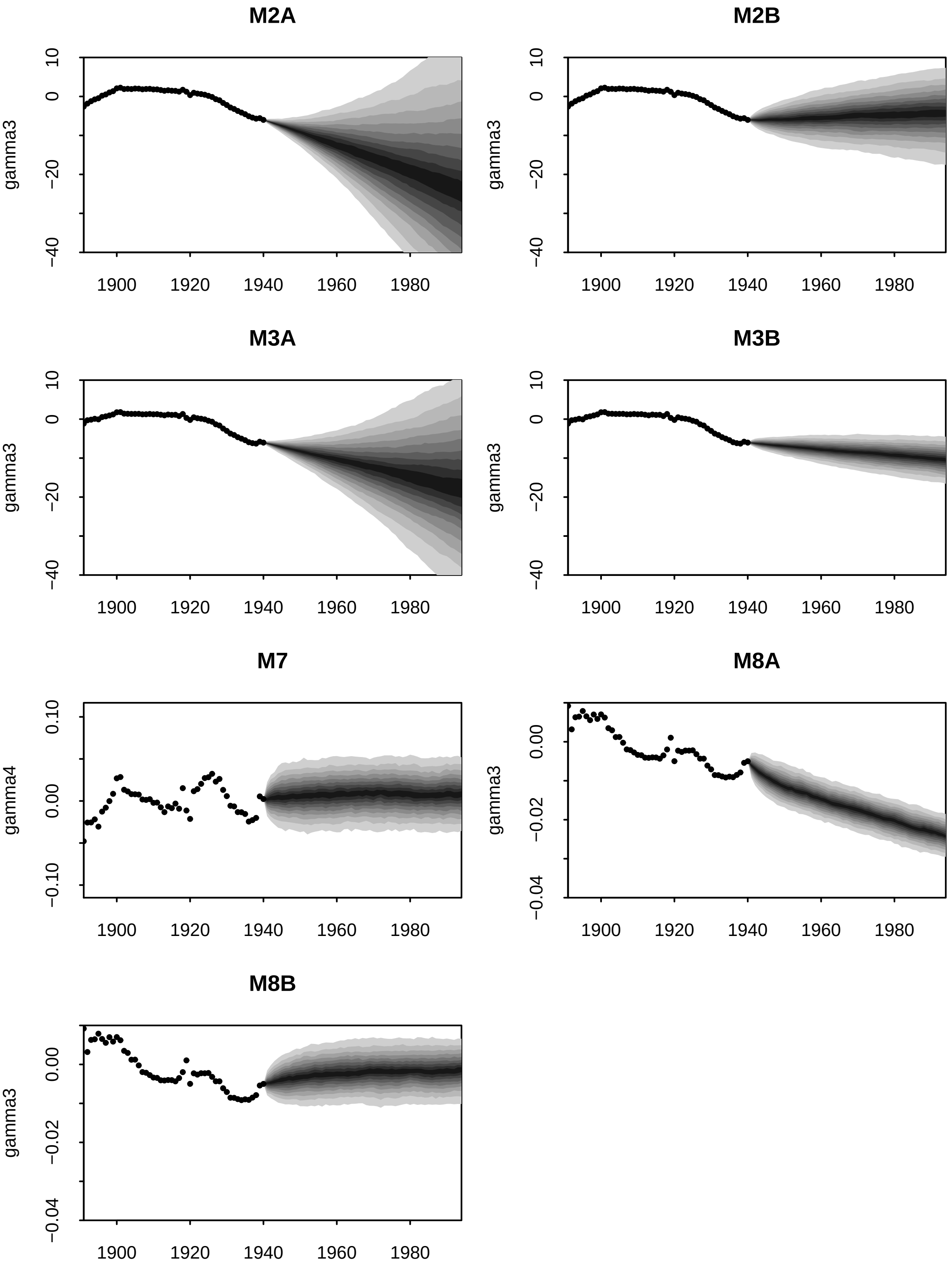


Figure 2: England & Wales, males: Fan charts for the projected cohort effect. For M1 and M5, there is no cohort effect so no fan charts have been plotted. (See Dowd, Blake and Cairns, 2007, for detailed description of how the fans are constructed.)

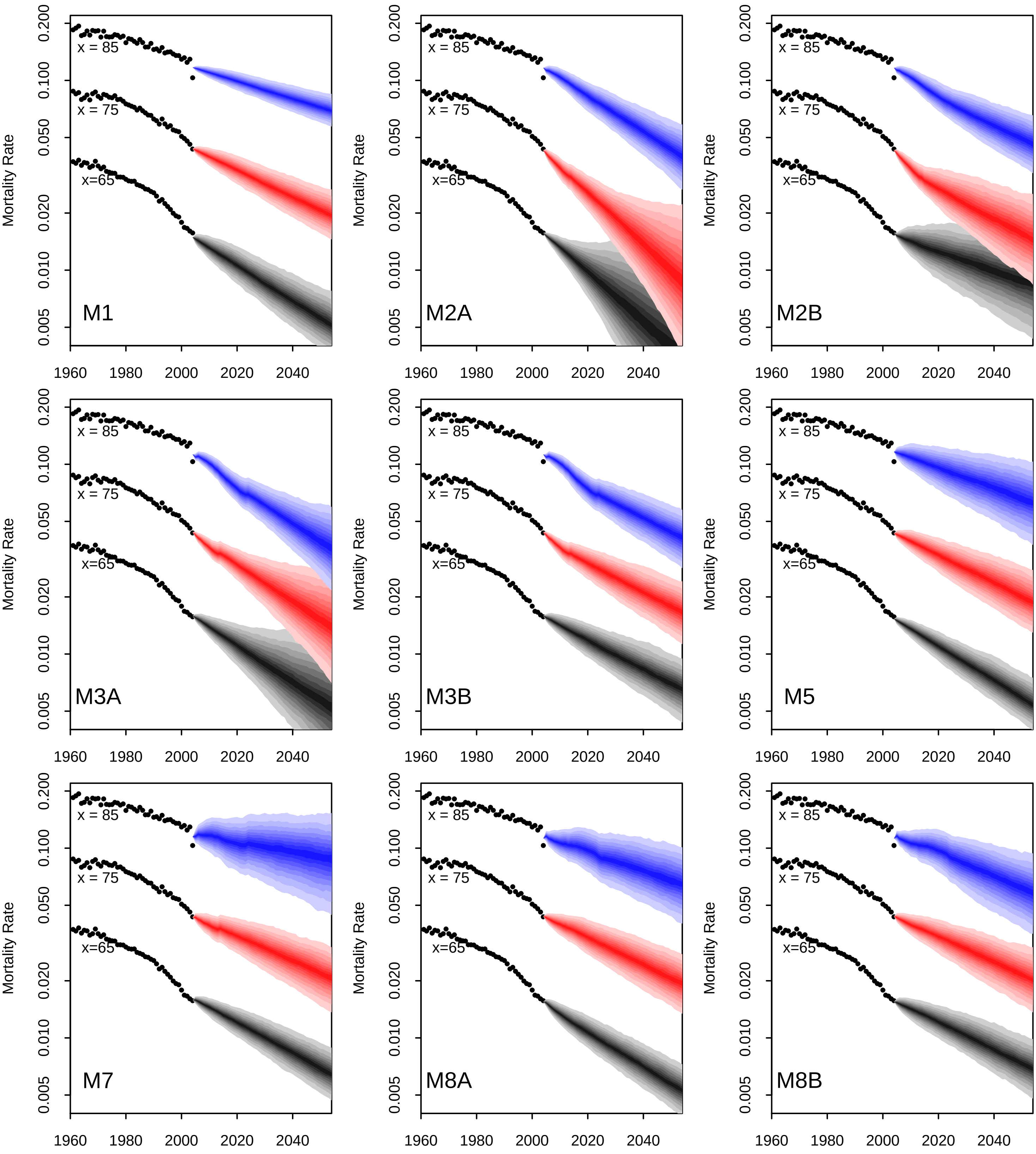


Figure 3: England & Wales, males: Mortality rates, $q(t, x)$, for models M1, M2A, M2B, M3A, M3B, M5, M7, M8A and M8B for ages $x = 65$ (grey), 75 (red), and 85 (blue). The dots show historical mortality rates for 1961 to 2004.

et al, 2007, Figure 7), $\beta_x^{(2)}$ declines with age,[15] forcing the fans at higher ages to be narrower, rather than wider. However, we note that these fan charts do not allow for parameter uncertainty, which would increase the width of the fan charts at 85.

Fans for M2A, M2B and M3A similarly are wider at age 65 than age 85. We note that for these models, the cohort effect may be significant. At age 65, the cohort effect is simulated from the inception of the projections. However at age 85, this is not the case. At older ages, projections initially use the fitted values of the cohort effect (E.g., the first 20 years of projection at age 85) and this has a consequent effect in reducing variability and the width of the fan charts.

- Figure 3 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. 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.

  These plots make full use of the data from 1961 to 2004. If we extrapolate the central section of each fan backwards in time, we see that it is approximately aligned with the mortality rates at ages 65, 75 and 85 in 1961.

- Figure 4 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 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 (grey) and M7 (red) equates to a difference in the rate of improvement in the age-65 mortality rate of 0.3% per annum.[16]

  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 high degree of uncertainty at age 85 resulting from a mixture of the variances

[15] The reason why $\beta_x^{(2)}$ declines with age is that mortality rates at higher ages have been improving at a lower rate than at younger ages.

[16] Specifically, for age 65, the M5 improvement rate was 2.1% per annum, while for M7 the improvement rate was 1.8% per annum.

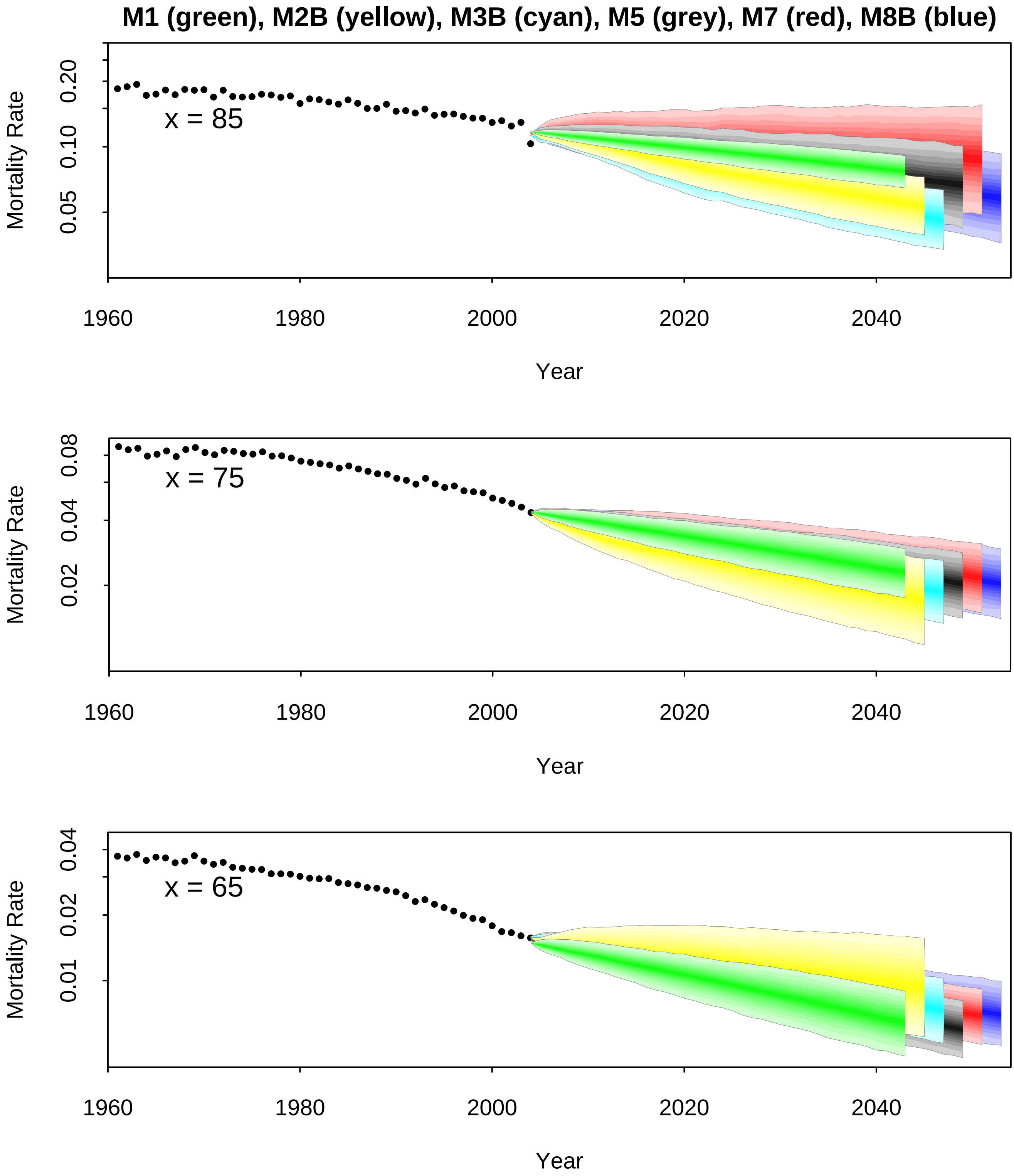


Figure 4: England & Wales, males: Mortality rates, $q(t, x)$, for models M1 (green), M2B (yellow), M3B (cyan) M5(grey), M7 (red), and M8A (blue) with fans overlaid for ages $x = 65$, 75, and 85. The dots show historical mortality rates for 1961 to 2004.

of and covariances between the $\kappa_t^{(i)}$ and $\beta_x^{(i)}$ terms. The fact that the central trend for M7 lies above that for M5 at ages 65 and 85 is due to the quadratic age effect $\beta_x^{(3)}$ in M7.

- Figure 5 shows the relative impact on forecast mortality rates at ages 65, 75 and 85 from using models M2A and M2B. In all cases, the M2A fan is wider, and more 'trumpet' shaped reflecting the greater uncertainty in the ARIMA(0,2,1) model.

  The differences between the two fans are largest at age 65. Everything else being equal, the age-65 fan will be wider because the uncertainty in $\gamma_c^{(3)}$ affects mortality rates as soon as the 1940 cohort has passed through. So at age 65 differences between the fans emerge almost immediately, whereas at age 85 they only emerge after 2025.

  Similar comments apply when we compare models M3A and M3B (Figure 6), although the impact is less severe at age 65 as the M3 age effect, $\beta_x^{(3)}$, is constant.

  For M2A and M2B, $\beta_x^{(3)}$ is higher at low ages, and so we can see that the uncertainty in the age 65 fans is relatively higher than the uncertainty in the respective fans for M3A and M3B.

- Figure 7 shows the relative impact on mortality rates at ages 65, 75 and 85 from using models M8A and M8B. The differences between the two fans are much smaller than those in Figure 5, even though the fans for $\gamma_c^{(3)}$ are very different for these two models (see Figure 2). The biggest difference is at age 65: the fans have a similar width, but the different trends equate to a difference in mortality improvement rate of about 0.6% per annum. This difference in trend is a direct consequence of the differences between the central trends of $\gamma_{t-x}^{(3)}$ in M8A and M8B. At age 65, we see that the trend with M8A (grey) is lower than that with M8B (red). In contrast, at age 85, the trend with M8A is higher. This is because $\beta_x^{(3)}$ (Figure 1, bottom left) is positive at age 65 (so lower values of $\gamma_t^{(3)}$ mean lower mortality) and negative at age 85.

In terms of considering the suitability of the models for the dataset under consideration, we can summarise as follows: The figures reveal reasonable consistency of forecasts between M1, M3B, M5, M7 and M8B, 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 less uncertain than at age 65, contrary to historical evidence. However, as noted earlier, in the case of M2, this might be due to the fact that the variability of the cohort effect is not allowed for till much later in the projections at age 85. Results for M1 are otherwise deemed to be plausible. M5 has escaped much comment in this section, but this reflects the

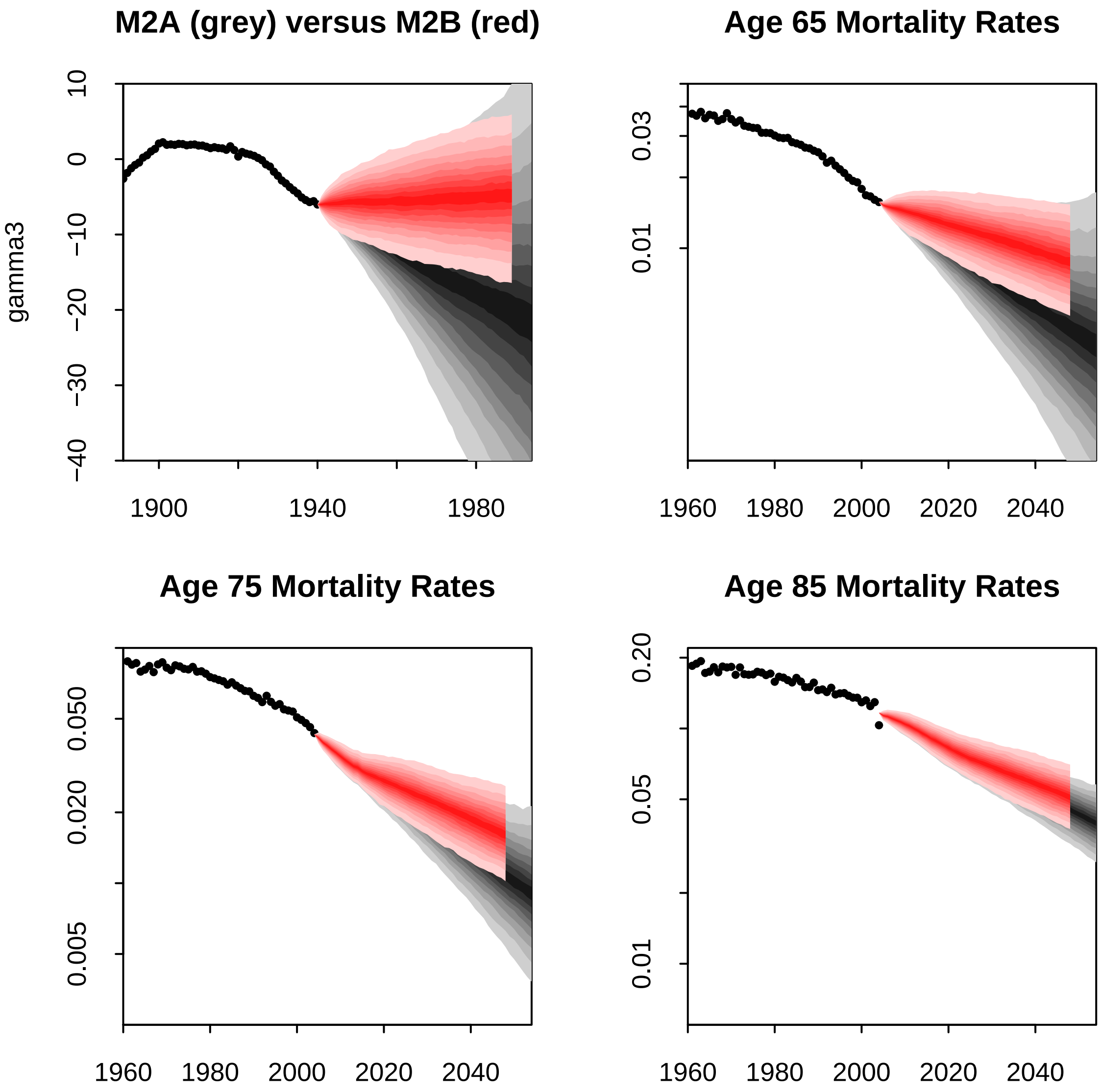


Figure 5: England & Wales, males: Fan charts comparing models M2A (grey fans) and M2B (red fans). Top left: historical (dots) and forecast (fans) values for the cohort effect, $\gamma_c^{(3)}$. Top right, bottom left and right: historical (dots) and forecast (fans) mortality rates, $q(t,x)$, for ages 65, 75 and 85.

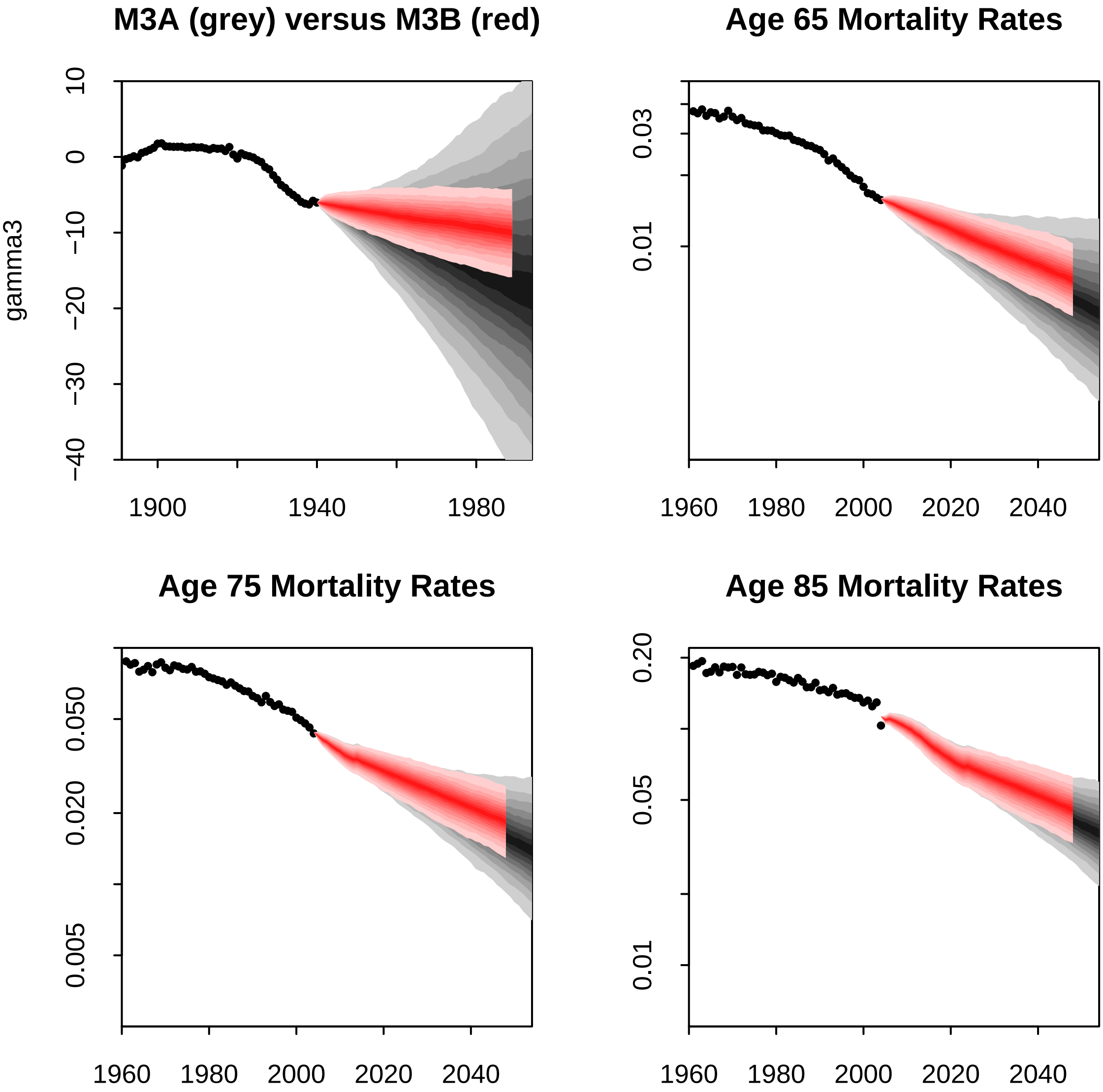


Figure 6: England & Wales, males: Fan charts comparing models M3A (grey fans) and M3B (red fans). Top left: historical (dots) and forecast (fans) values for the cohort effect, $\gamma_c^{(3)}$. Top right, bottom left and right: historical (dots) and forecast (fans) mortality rates, $q(t, x)$, for ages 65, 75 and 85.

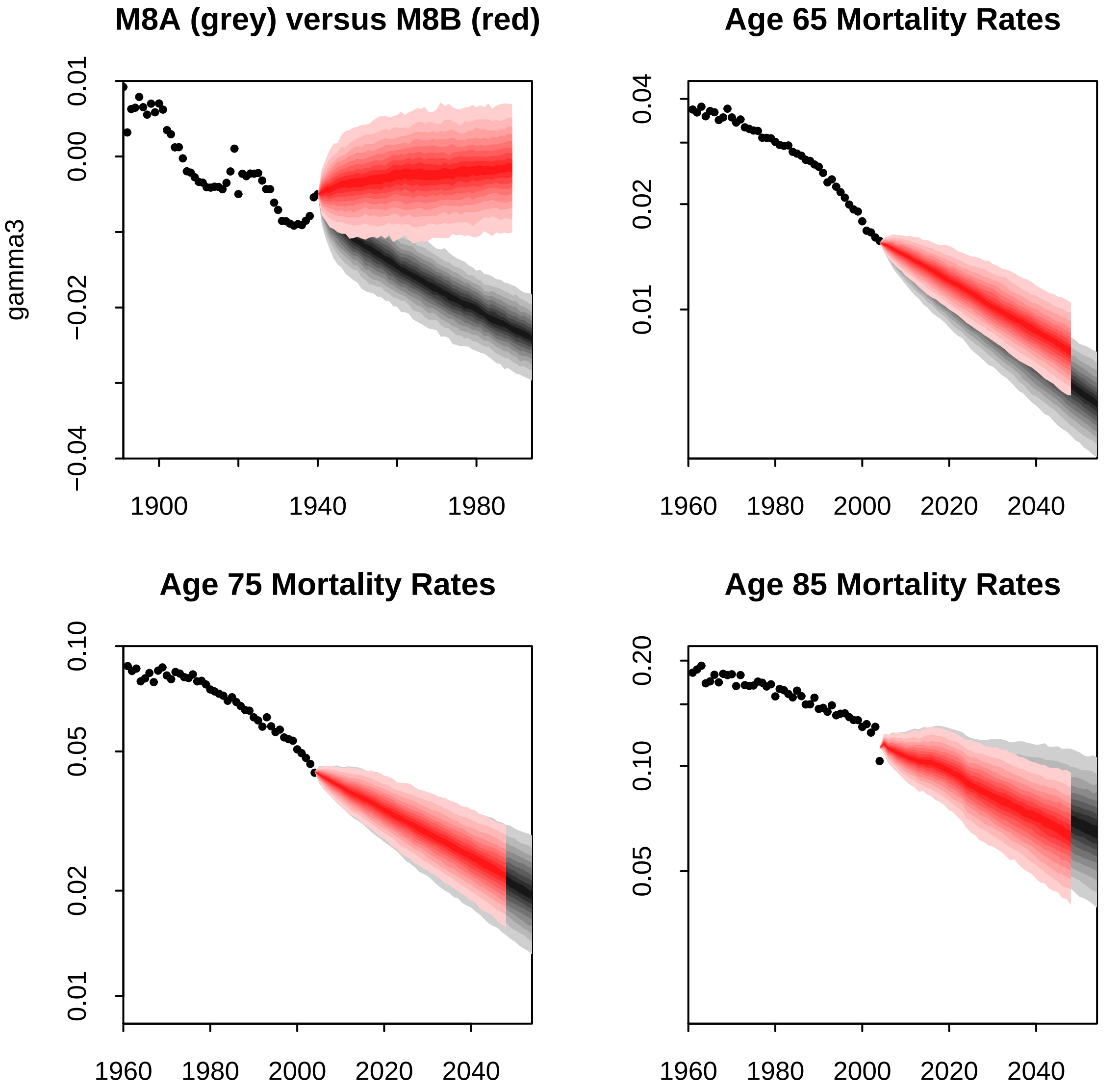


Figure 7: England & Wales, males: Fan charts comparing models M8A (grey fans) and M8B (red fans). Top left: historical (dots) and forecast (fans) values for the cohort effect, $\gamma_c^{(3)}$. Top right, bottom left and right: historical (dots) and forecast (fans) mortality rates, $q(t,x)$, for ages 65, 75 and 85.

fact that its forecasts have, so far, passed the plausibility test. M3, M7 and M8, have attracted more comment, but the same conclusion can be made, namely that they too have, so far, passed the plausibility test.

4 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 “derived” 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). These provide additional illustrations of possible model risk.

Figure 8 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 the cohort effect, $\gamma_c^{(3)}$, for model M2 for this group of males has already been estimated from the historical data. Consequently, the choice of forecasting model for $\gamma_c^{(3)}$ has no impact on $S(t, 65)$: as a consequence, models M2A and M2B produce identical results. The same applies to M3 and M8. For younger cohorts (see, for example, our second example for age 60 below), however, we would see a difference between M2A and M2B, between M3A and M3B, and between M8A and M8B.

The fans for M1, M2B, M3B, M5, M7 and M8B are superimposed in Figure 9 to aid comparison. This reveals some differences between the trends and more significant differences between the dispersions. Again, therefore, model risk cannot be ignored: with this particular application, it manifests itself in terms of different survivor index trends.

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 10. We can see that there are some moderate differences between the models. (see Table 5).

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

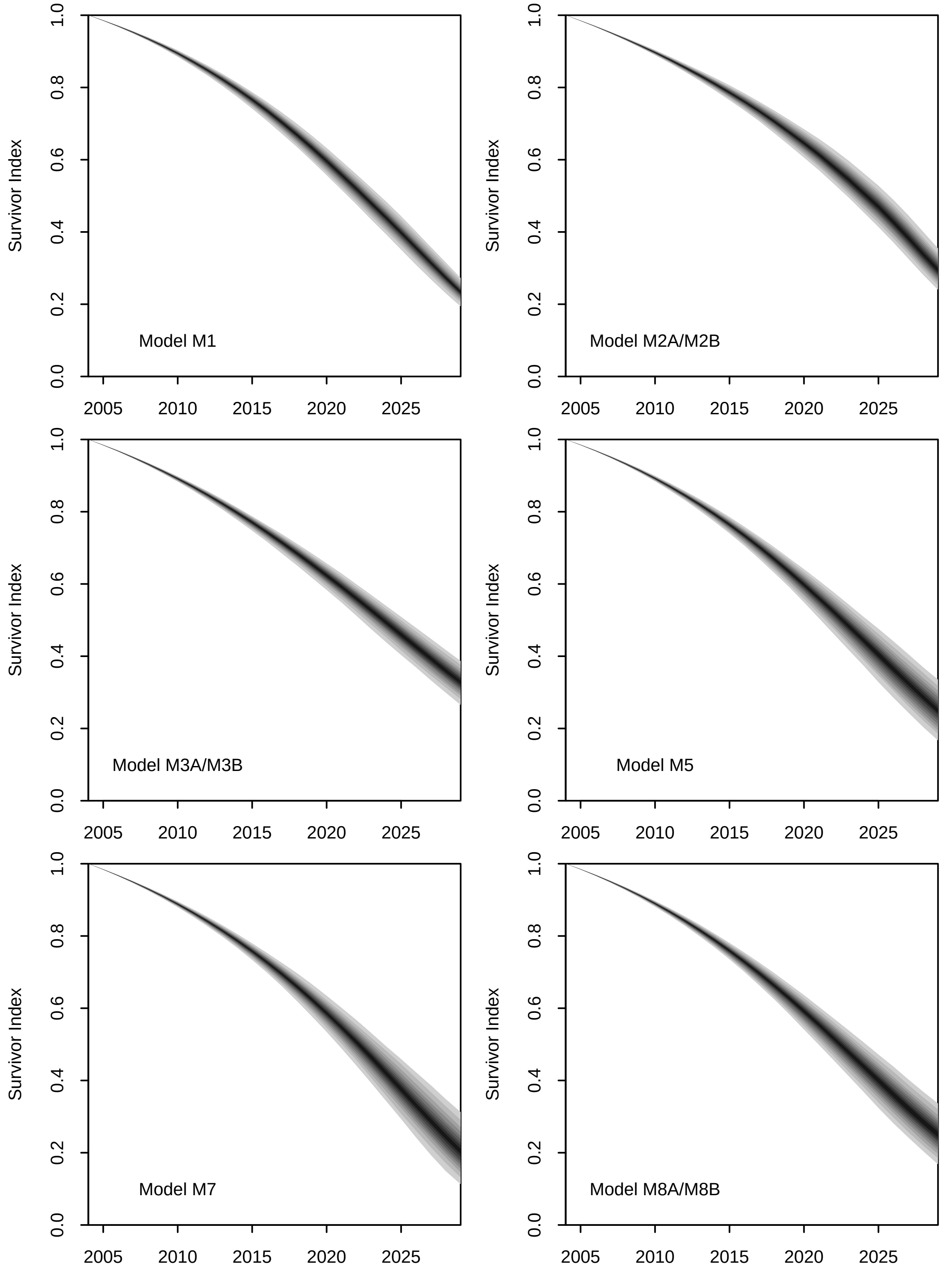


Figure 8: England & Wales, males: Fan charts for the survivor index $S(t, 65)$ for the cohort aged 65 at the start of 2005, for models M1, M2A/M2B, M3A/M3B, M5, M7, and M8A/M8B.

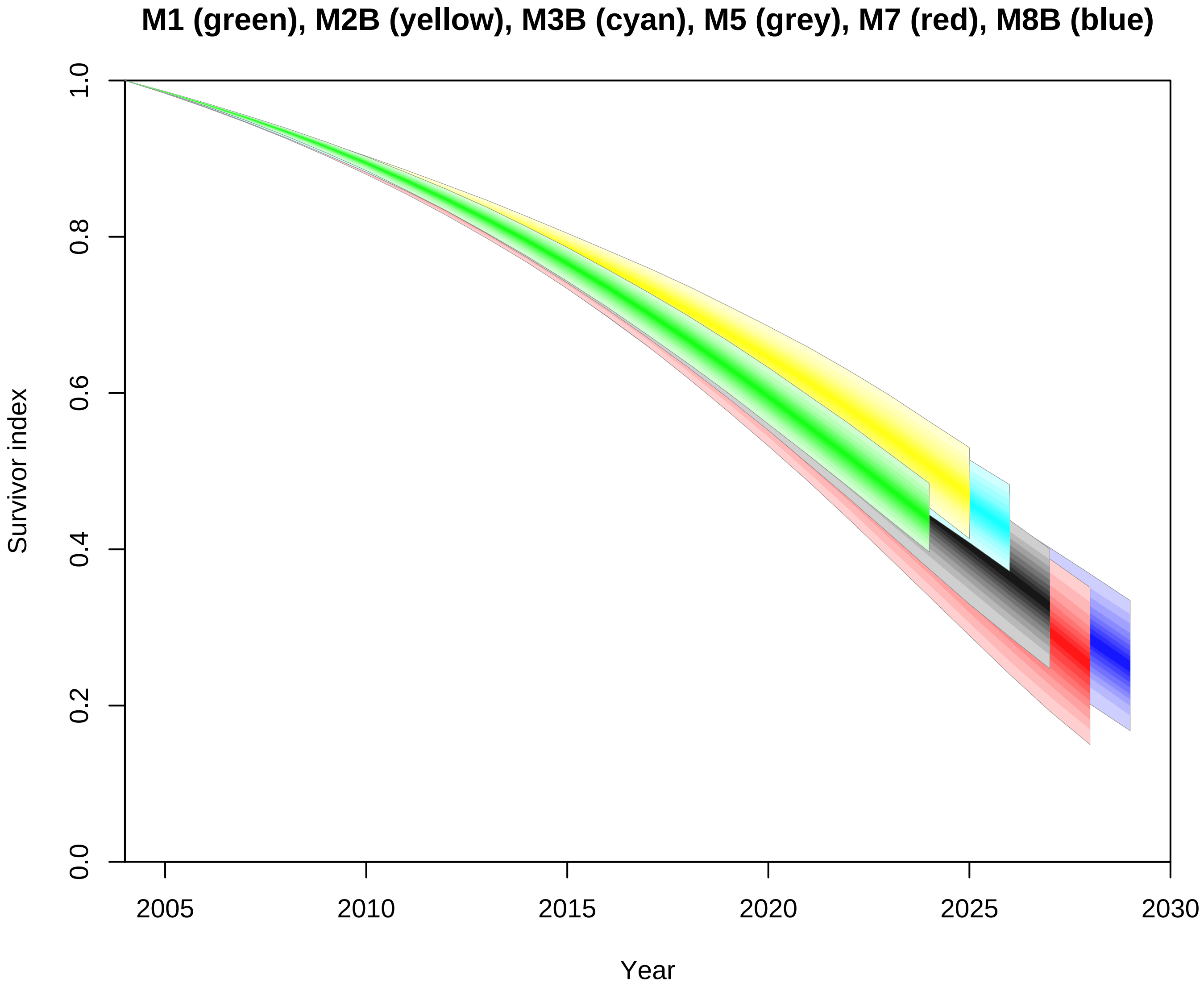


Figure 9: England & Wales, males: Fan charts for the survivor index $S(t, 65)$ for the cohort aged 65 at the start of 2005, for models M1 (green), M2B (yellow), M3B (cyan), M5(grey), M7(red) and M8B (blue).

2005:

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

In this case the cohort effect needs to be simulated for the underlying cohort and so differences between M2A and M2B, M3A and M3B, and M8A and M8B emerge (see Figure 11, and Table 6). 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.

			Coefficient
Model	Mean	St. Dev.	of variation
M1	11.393	0.201	1.76%
M2A/M2B	11.796	0.217	1.83%
M3A/M3B	11.673	0.210	1.80%
M5	11.415	0.255	2.23%
M7	11.264	0.279	2.48%
M8A/M8B	11.357	0.259	2.28%

Table 5: England & Wales, males: Mean, standard deviation and coefficient of variation (the standard deviation divided by the mean) of the random present value $P = \sum_{t=1}^{25} v^t S(t, 65)$.

			Coefficient
Model	Mean	St. Dev.	of variation
M1	13.428	0.222	1.65 %
M2A	13.804	0.260	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	13.201	0.304	2.30 %
M8A	13.393	0.272	2.03 %
M8B	13.312	0.276	2.07 %

Table 6: England & Wales, males: Mean, standard deviation and coefficient of variation (the standard deviation divided by the mean) of the random present value $P = \sum_{t=1}^{30} v^t S(t, 60)$.

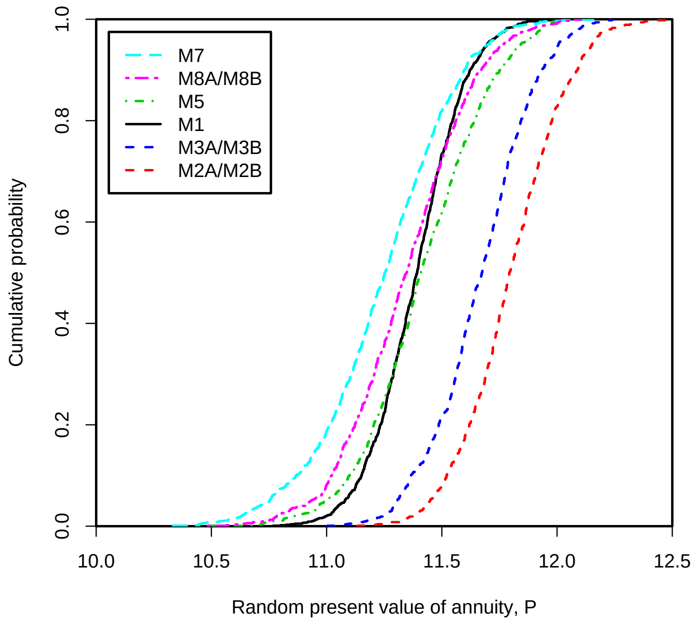


Figure 10: England & Wales, males: Random present value of an annuity payable annually in arrears for a maximum of 25 years to a male aged 65 at the start of 2005, assuming a rate of interest of 4% per annum. The legend follows the order from left to right at probability 0.2.

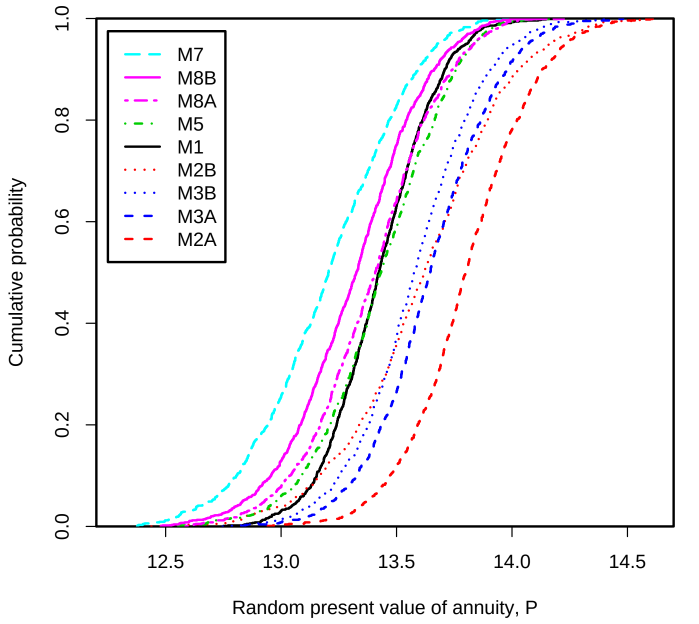


Figure 11: England & Wales, males: Random present value of an annuity payable annually in arrears for a maximum of 30 years to a male aged 60 at the start of 2005, assuming a rate of interest of 4% per annum. The legend follows the order from left to right at probability 0.2.

5 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 constructing the simulation model. For each model, we compare three sets of simulations in Figures 12 to 18:

- (Grey fans) (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).
- (Blue fans) (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).
- (Red fans) (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).

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 red and the blue fans, implying that the red and blue fans should be the same. The fact that the period and cohort effects have to be estimated means that the red and blue fans will be affected by estimation errors, but if a model is robust then we would expect the red and blue fans to have similar median trajectories and similar spreads.

From the results in Figures 12 to 18, we can make the following remarks:

- 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.[17] For example, if the central projections in the grey fans are extrapolated backwards from 2004, then the extrapolation starts off below the dots but then reconnects around about 1961. For the red and blue 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,[18] 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.

[17] Recall that for a pure random walk process, the median forecast is a straight line extrapolation of the line connecting the first and the last observations.

[18] These comments apply whether or not this change in trend is genuine, or just the result of statistical variation.

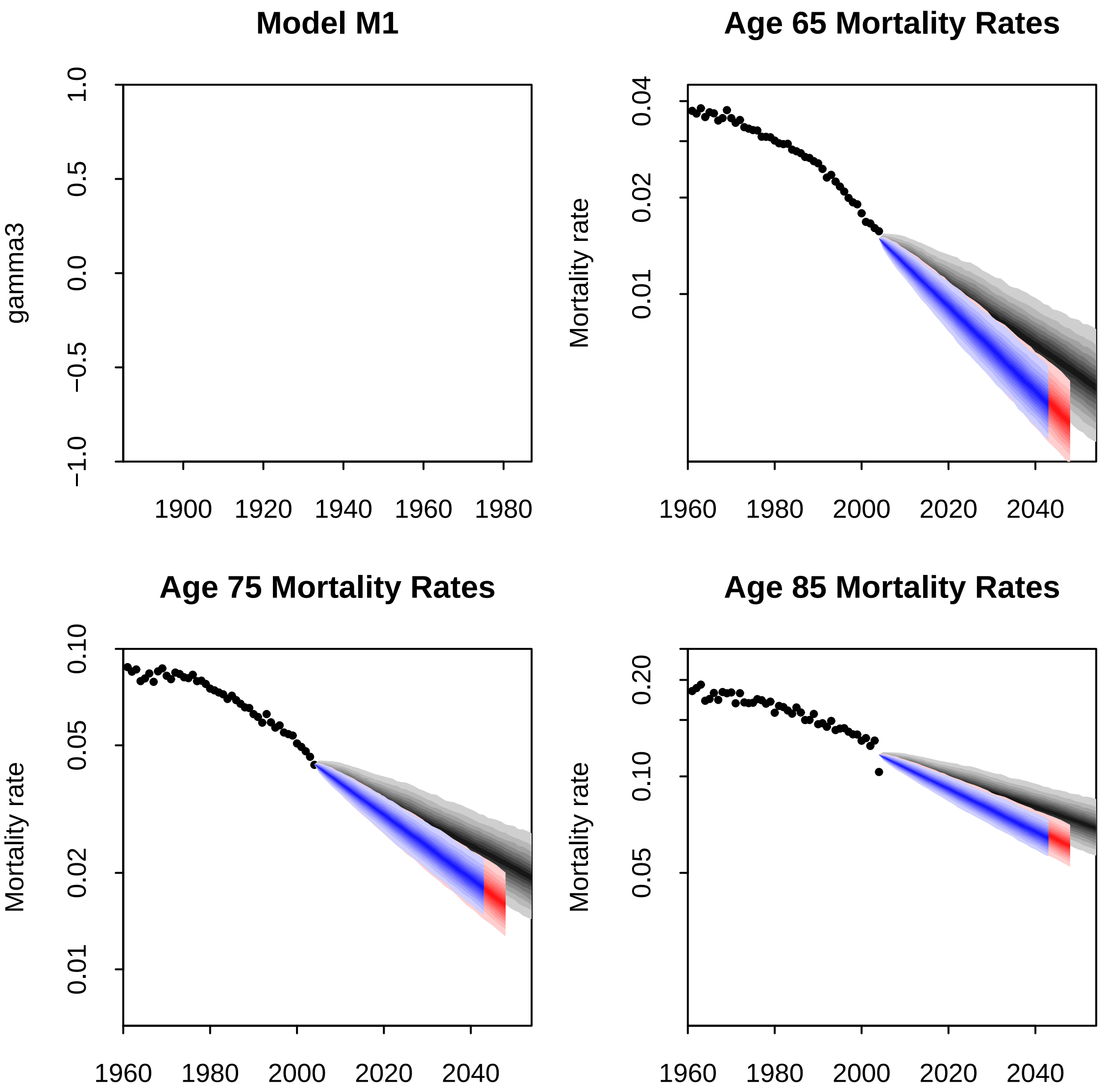


Figure 12: England & Wales, males: Model M1. Cohort effect (absent for this model) and mortality rates for ages 65, 75 and 85. Dots and grey fans: historical data from 1961 to 2004 used to estimate the historical $\kappa_t^{(2)}$; forecasting model uses the 44 $\kappa_t^{(2)}$ values. Dots and red fans: historical data from 1961 to 2004 used to estimate the historical $\kappa_t^{(2)}$; forecasting model uses the 24 most recent $\kappa_t^{(2)}$ values. Blue fans: historical data from 1981 to 2004 used to estimate the historical $\kappa_t^{(2)}$; forecasting model uses the full 24 $\kappa_t^{(2)}$ values.

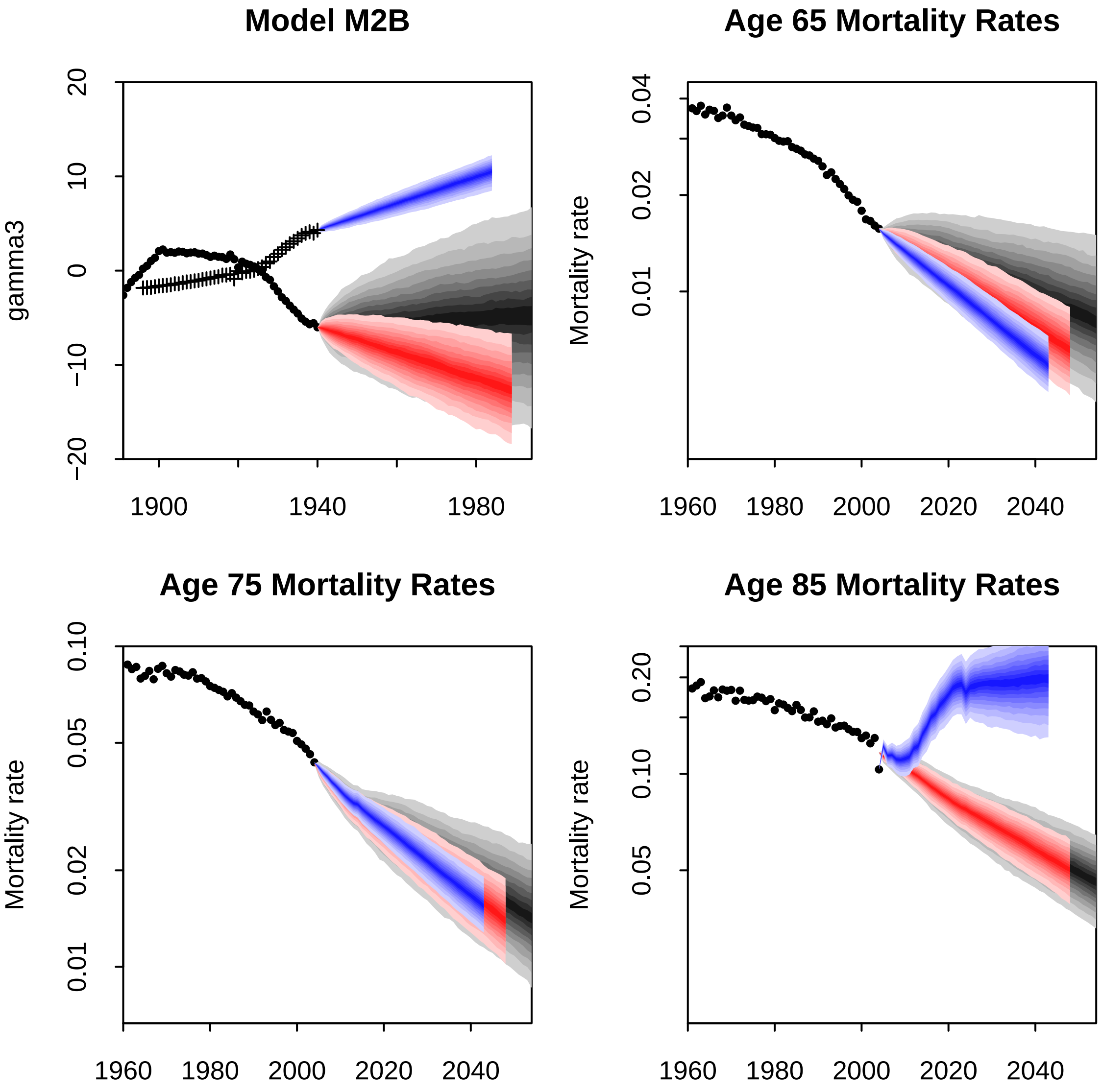


Figure 13: England & Wales, males: Model M2B. Cohort effect and mortality rates for ages 65, 75 and 85. Dots and grey fans: historical data from 1961 to 2004 used to estimate the historical $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 44 $\kappa_t^{(2)}$ values and the 60 $\gamma_c^{(3)}$ values. Dots and red fans: 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 blue fans: 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.

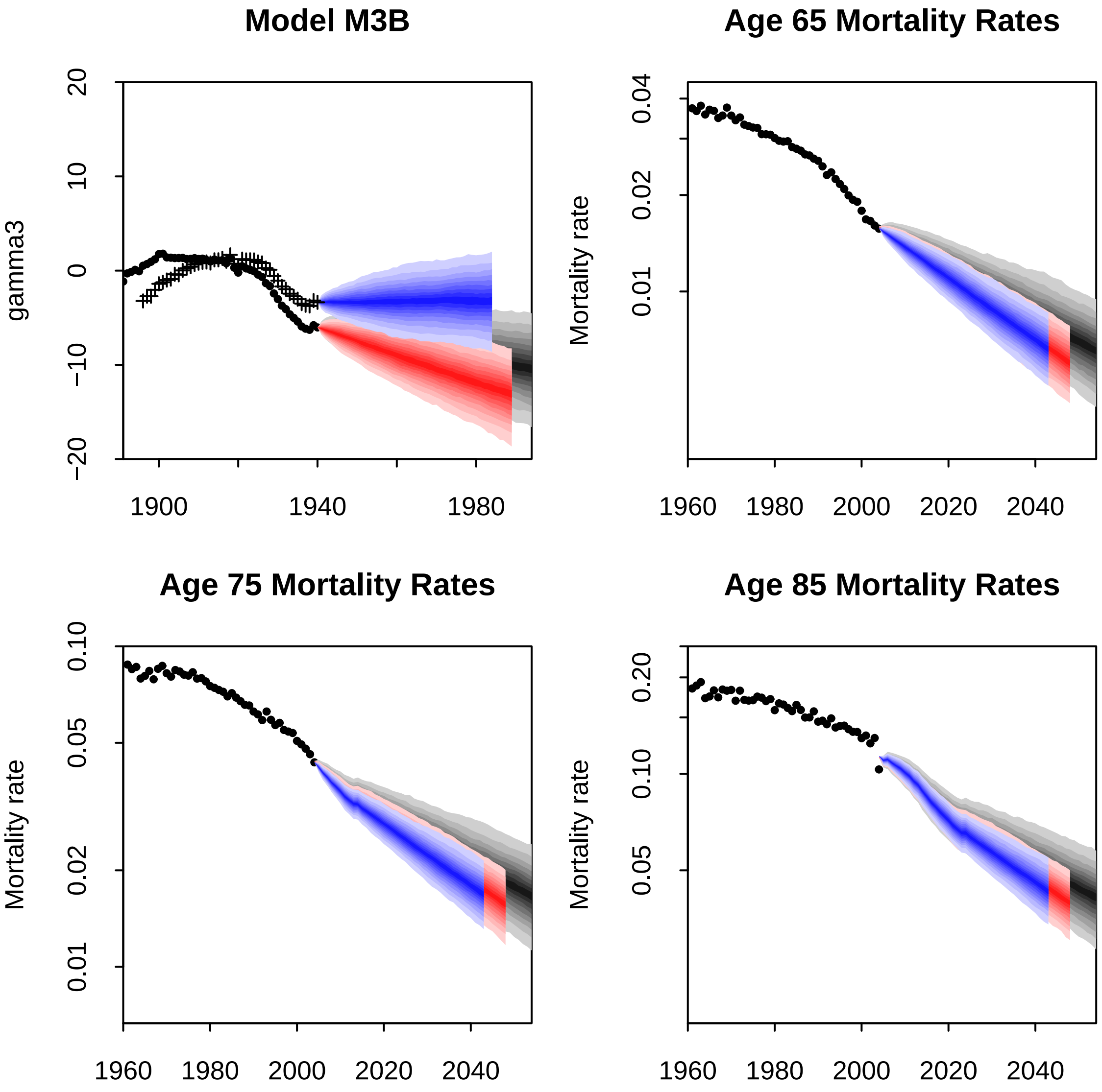


Figure 14: England & Wales, males: Model M3B. Cohort effect and mortality rates for ages 65, 75 and 85. Dots and grey fans: historical data from 1961 to 2004 used to estimate the historical $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 44 $\kappa_t^{(2)}$ values and the 60 $\gamma_c^{(3)}$ values. Dots and red fans: 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 blue fans: 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.

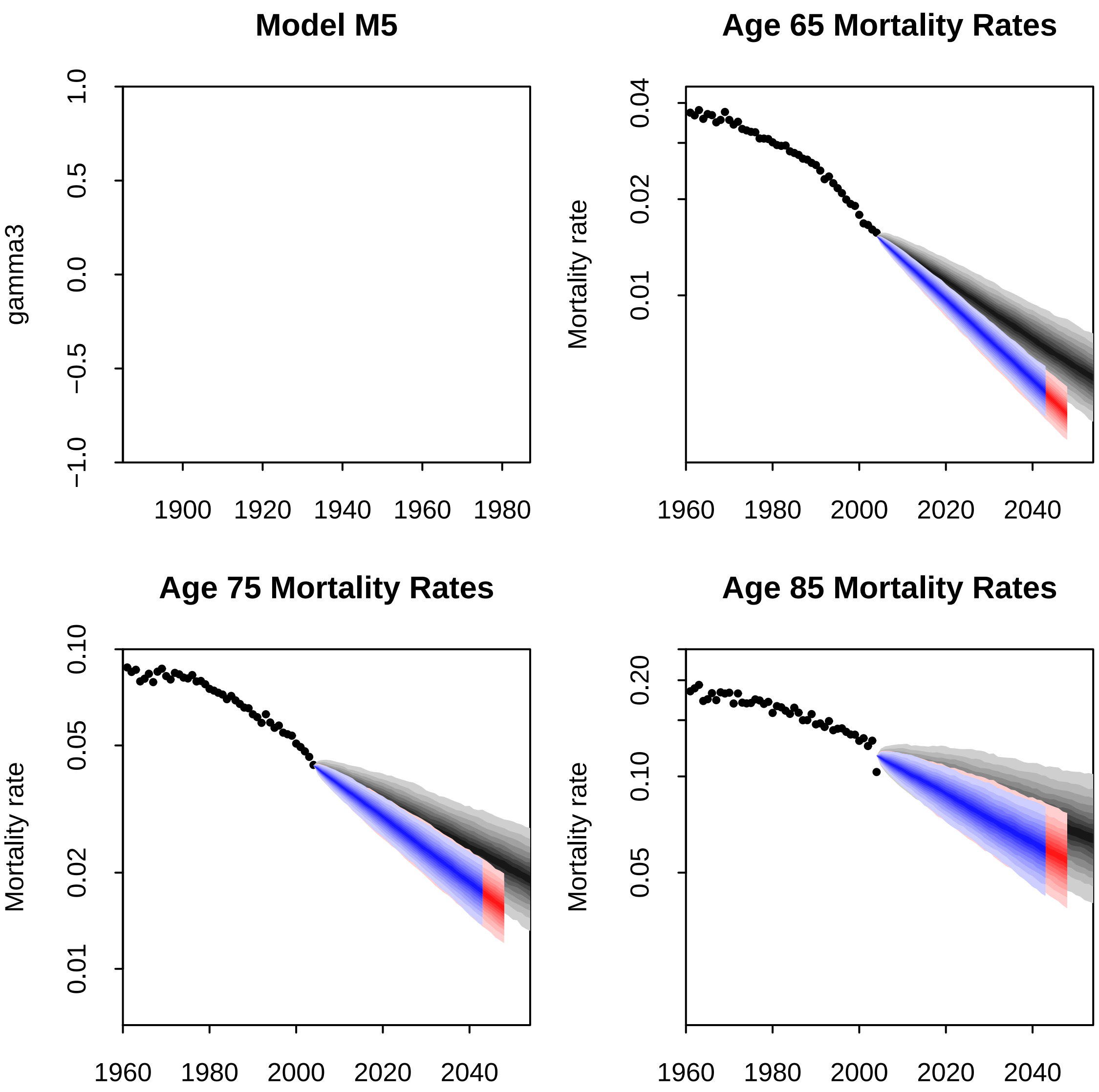


Figure 15: England & Wales, males: Model M5. Cohort effect (absent for M5) and mortality rates for ages 65, 75 and 85. Dots and grey fans: historical data from 1961 to 2004 used to estimate the historical $\kappa_t^{(i)}$; forecasting model uses the 44 $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ values. Dots and red fans: historical data from 1961 to 2004 used to estimate the historical $\kappa_t^{(i)}$; forecasting model uses the 24 most-recent $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ values. Blue fans: historical data from 1981 to 2004 used to estimate the historical $\kappa_t^{(i)}$; forecasting model uses the full 24 $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ values.

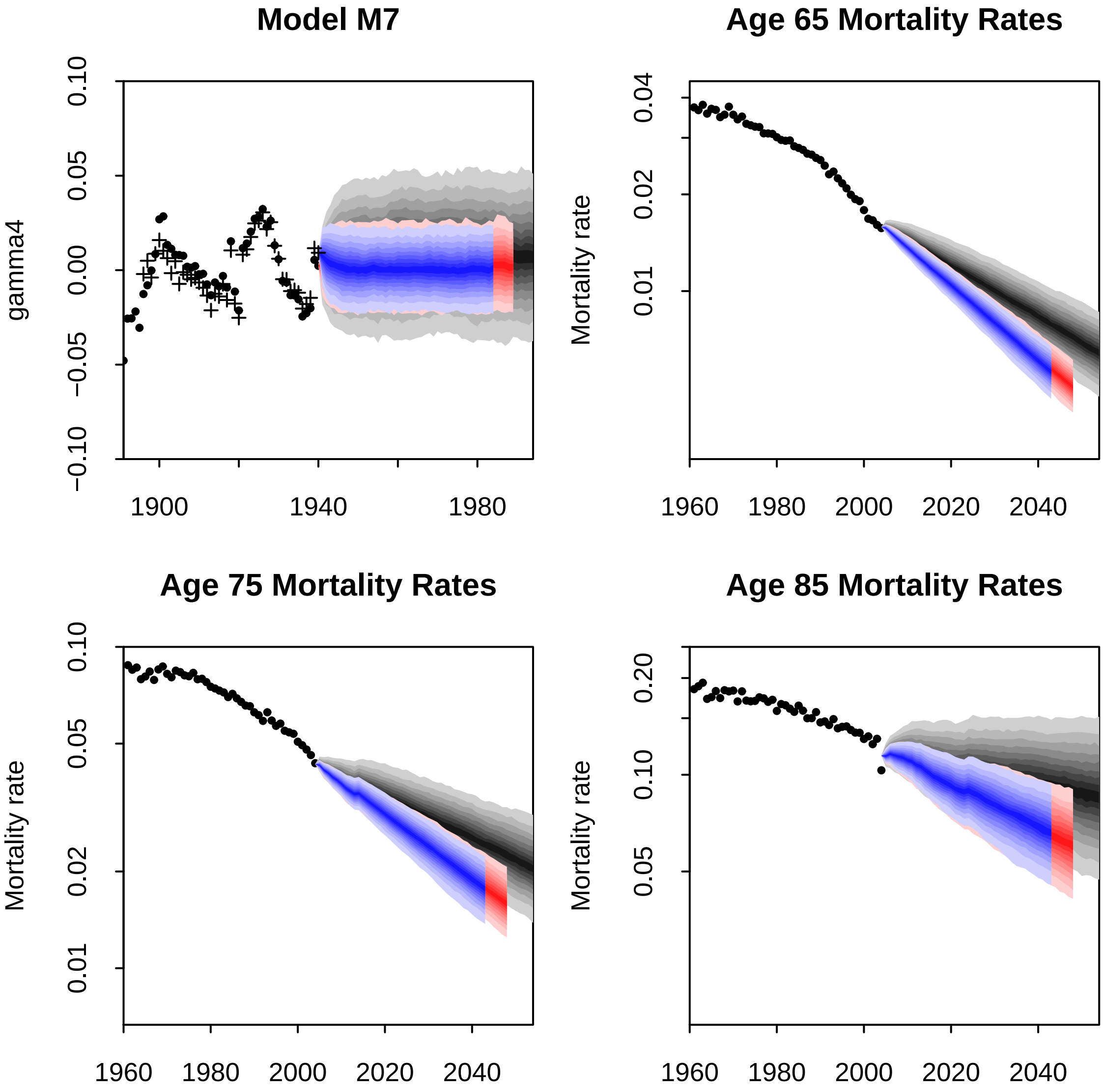


Figure 16: England & Wales, males: Model M7. Cohort effect and mortality rates for ages 65, 75 and 85. Dots and grey fans: historical data from 1961 to 2004 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 44 $\kappa_t^{(i)}$ values and 60 $\gamma_c^{(4)}$ values. Dots and red fans: historical data from 1961 to 2004 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 24 most-recent $\kappa_t^{(i)}$ values and the 45 most-recent $\gamma_c^{(4)}$ values. Crosses and blue fans: historical data from 1981 to 2004 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 24 fitted $\kappa_t^{(i)}$ values and the full 45 fitted $\gamma_c^{(4)}$ values.

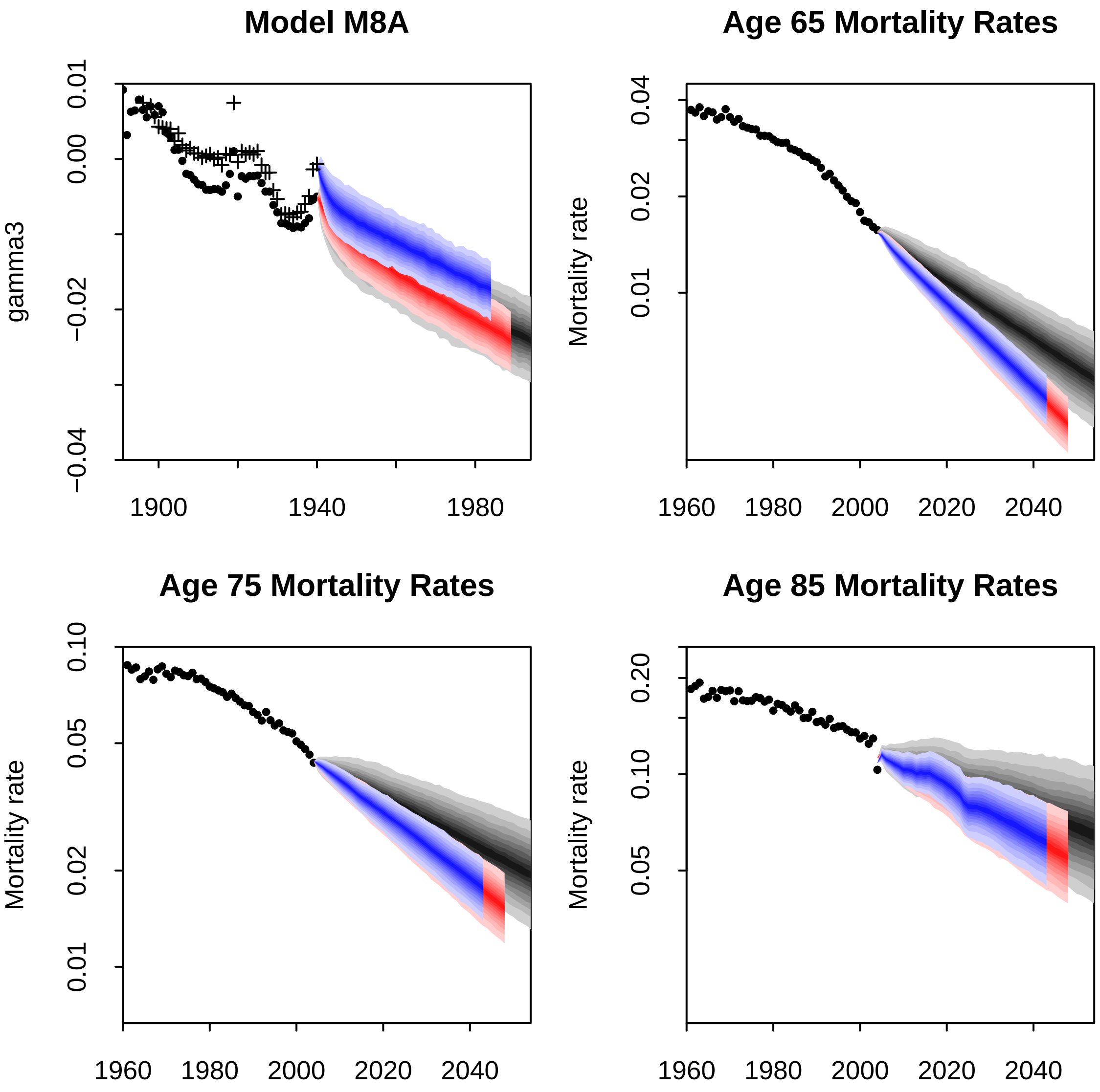


Figure 17: England & Wales, males: Model M8A. Cohort effect and mortality rates for ages 65, 75 and 85. Dots and grey fans: historical data from 1961 to 2004 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 44 $\kappa_t^{(i)}$ values and 60 $\gamma_c^{(3)}$ values. Dots and red fans: historical data from 1961 to 2004 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 24 most-recent $\kappa_t^{(i)}$ values and the 45 most-recent $\gamma_c^{(3)}$ values. Crosses and blue fans: historical data from 1981 to 2004 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 24 fitted $\kappa_t^{(i)}$ values and the full 45 fitted $\gamma_c^{(3)}$ values.

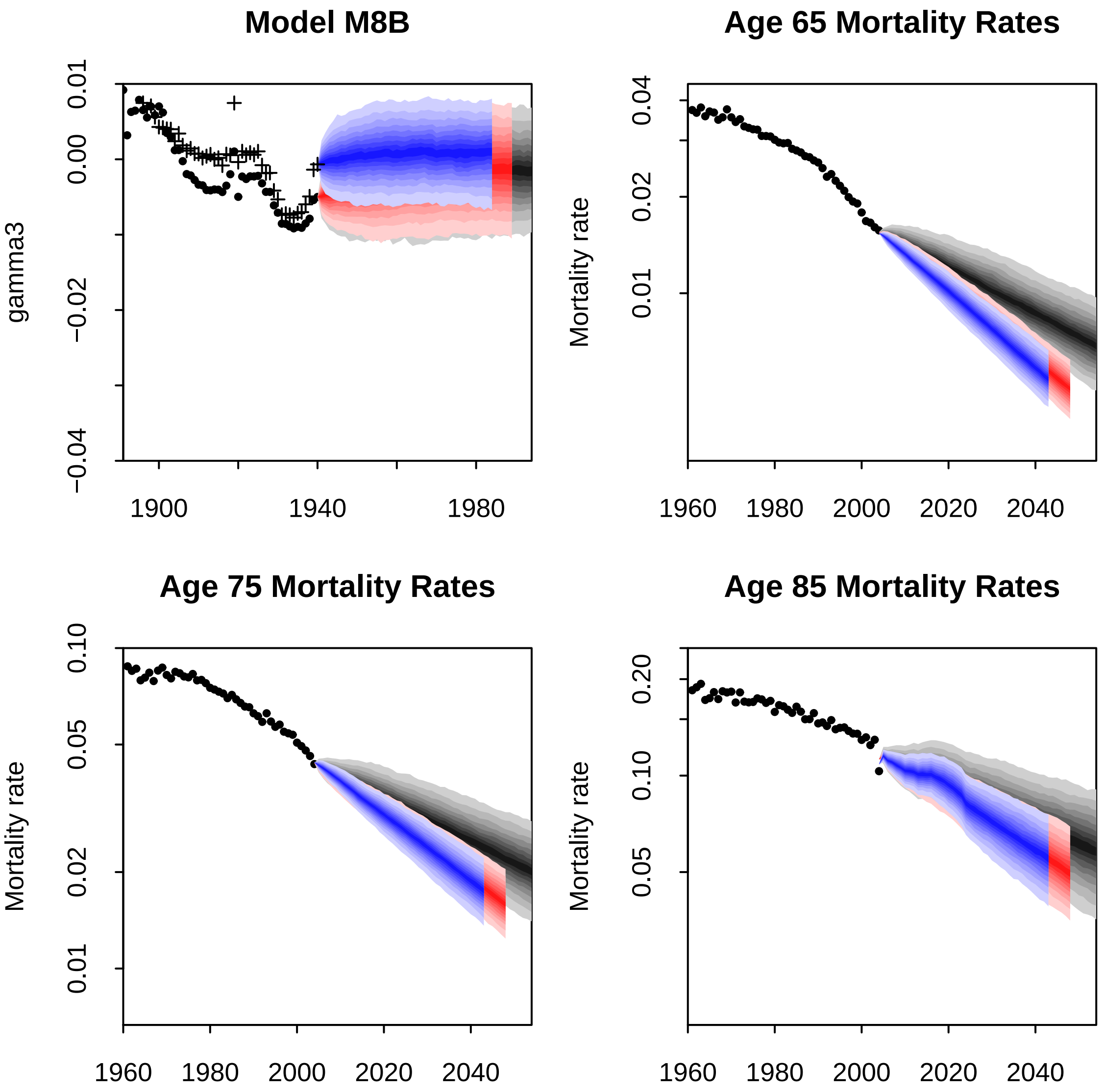


Figure 18: England & Wales, males: Model M8B. Cohort effect and mortality rates for ages 65, 75 and 85. Dots and grey fans: historical data from 1961 to 2004 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 44 $\kappa_t^{(i)}$ values and 60 $\gamma_c^{(3)}$ values. Dots and red fans: historical data from 1961 to 2004 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 24 most-recent $\kappa_t^{(i)}$ values and the 45 most-recent $\gamma_c^{(3)}$ values. Crosses and blue fans: historical data from 1981 to 2004 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 24 fitted $\kappa_t^{(i)}$ values and the full 45 fitted $\gamma_c^{(3)}$ values.

- In most cases, the grey fans are wider, reflecting the greater volatility in mortality rates that can be seen in the years 1961 to 1980, and which are not directly relevant in the red and blue fans.[19]

- For M2B (Figure 13), there are similar differences between the red or blue fans, on the one hand, and the grey fan, on the other. However, we can also see very significant differences between the red and blue fans, most obviously at age 85 where there is a clear problem with the blue fan. The explanation for the implausible shape of the blue 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 Figure 1, top left). When we use data from 1981 to 2004 (see Cairns et al., 2007, Figure 14), 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 14 in Cairns et al. (2007) 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 improving 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 13, blue 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 improve at high ages within the forecasting period 2004-2024 (Figure 13, red fans). Thus, the finding in Cairns et al. (2007), 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 outputs in 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: a sure sign that the likelihood function has multiple maxima. It therefore seems likely that the blue fan relates to one maximum and the red fan to another.

  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

[19] 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 red and blue fans draw on estimates of the period effects that cover the less-volatile years.

data is used. Nonetheless, it is possible that other implementations of M2 are less unstable.

- For M7 (Figure 16), the fans look stable. In particular, the red and blue fans are very similar in terms of trajectory and spread. The greater spread of the grey fans reflects a greater volatility in the $\kappa_t^{(i)}$ prior to 1981. Cairns et al. (2007, Figure 15) 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 grey mortality fans also have a different mean trajectory from the red and blue 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 England & Wales males 1961-2004 and 1981-2004 datasets.

In summary, for the dataset it appears that M1, M3, M5, M7 and M8 are all reasonably robust relative to the historical data used. M2B forecasts, in contrast, look to be unstable.

6 Sensitivity analysis

It is also important to perform a sensitivity analysis. Here we illustrate with M7, and discuss how sensitive outputs are to changes in key parameters of the models driving $\kappa_t^{(1)}$, $\kappa_t^{(2)}$, $\kappa_t^{(3)}$ and $\gamma_t^{(4)}$.

Why is such a sensitivity analysis important? An individual modeller might have their own subjective opinion about specific model parameters. A sensitivity analysis sheds light on what the impact of this might be. Alternatively, if the subjective opinion concerns mortality improvement rates at specific ages, then a sensitivity analysis will help the modeller to choose the right drift parameters for the period effects.

In Figure 19, we plot pairs of fans that show how projected mortality rates change if we vary key parameters in the forecasting model. In each case, one parameter is varied while others remain fixed and equal to their unconditional maximum likelihood estimates.

From these plots, we can see that the drifts in the random-walk model for the $\kappa_t^{(i)}$ have a critical impact on mortality rate dynamics:

- A decrease of 0.01 in the drift of $\kappa_t^{(1)}$ (top left) means that the $q(t,x)$ improvement rate changes by about 1% at all ages.

- A decrease of 0.001 in the drift of $\kappa_t^{(2)}$ (top right) converts into different changes in the $q(t,x)$ improvement rates at different ages. At age $x = 74.5$, there is no impact. At age 64.5, there is a $10 \times 0.001 = 0.01$ decrease in the improvement rate (that is, $\beta_x^{(2)}$ times the amount of the change in the drift). At age 84.5, there is a $10 \times 0.001 = 0.01$ increase in the improvement rate. In other words, the impact on the improvements is linear in age.

- A decrease of 0.0001 in the drift of $\kappa_t^{(3)}$ (centre left) converts into different changes in the $q(t,x)$ improvement rates at different ages. At age $x = 74.5$, there is little impact. At age 64.5, there is a 0.01 increase in the improvement rate. At age 84.5, there is also a 0.01 increase in the improvement rate. In other words, the impact on the improvements is quadratic in age.

- The AR(1) model for $\gamma_c^{(4)}$ is

  $$\gamma_{c+1}^{(4)} = \mu_\gamma + \alpha_\gamma(\gamma_c^{(4)} - \mu_\gamma) + \sigma_\gamma \epsilon_{c+1}^{(4)}$$

  where the $\epsilon_c^{(4)}$ are i.i.d. standard normal innovations that are assumed to be independent of the random-walk model for the period effects. In Figure 19 (centre right), we show the impact of changing the mean-reversion level, μ_γ, of $\gamma_c^{(4)}$. We found that setting the mean reversion level to anything reasonable and consistent with historical data (Figure 1) had a negligible effect. Therefore what is plotted here is the result of making a very substantial change in the mean-reversion level. Changing the mean-reversion level has the longer run effect of shifting the fans up or down.

- Figure 19 bottom left shows what happens when we increase the volatility, σ_γ, of $\gamma_c^{(4)}$. Again, the increase has to be very substantial to see any significant change. We can see that changing the volatility widens the fans but does not change the trend.

- Changes to the mean-reversion parameter, α_γ, only have a visible impact (Figure 19, bottom right) if there is also high volatility. The grey fan has a high volatility only, while the red fan combines high volatility with a lower mean reversion parameter. Even here, differences are difficult to detect, but, at age 65, the red fan can (just) be seen to expand at a slower rate with narrower upper and lower bounds.

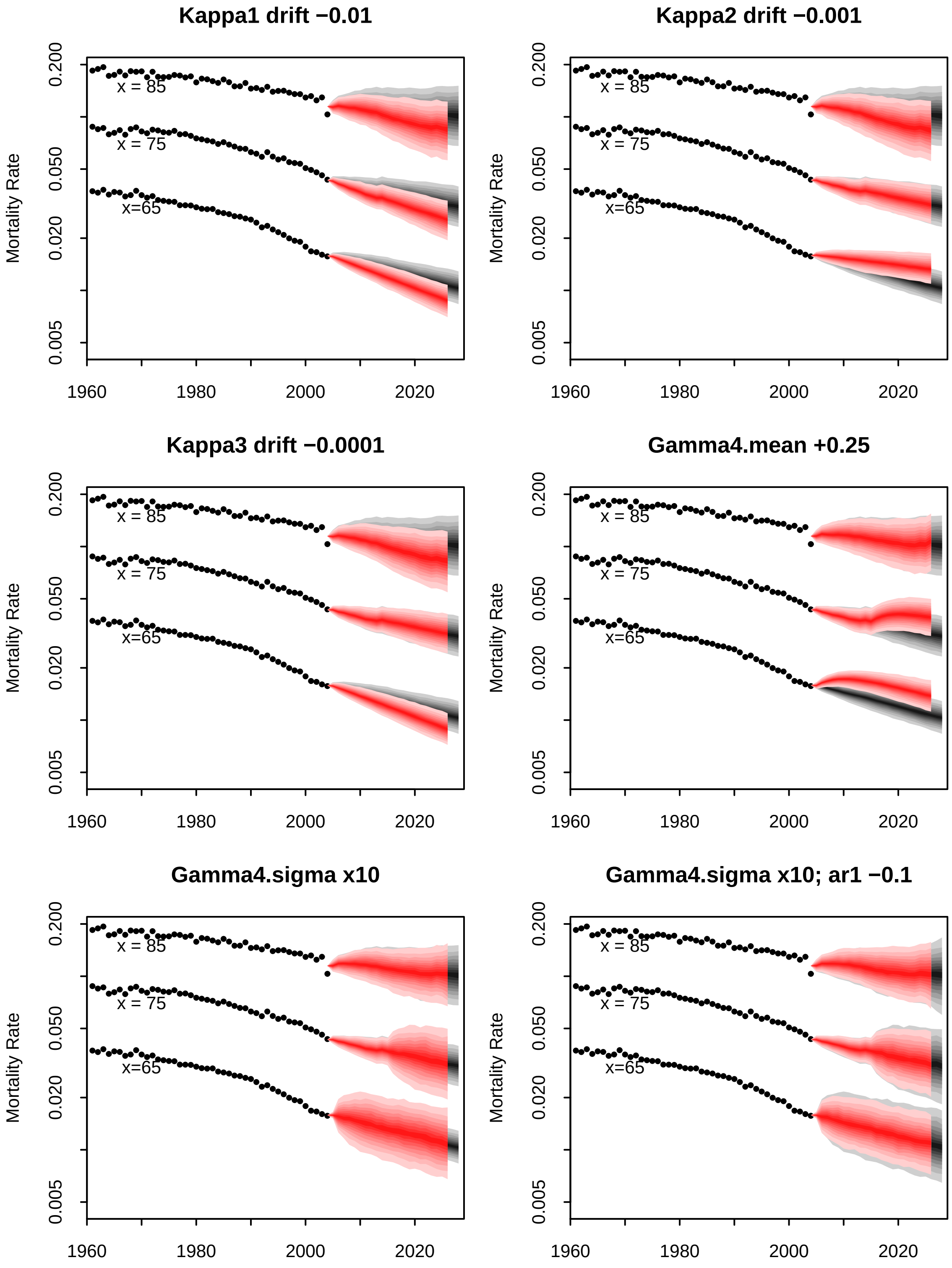


Figure 19: England & Wales, males: Sensitivity of mortality rates to changes in forecasting model parameters. Central case (lower grey fans) $\mu = (-0.016, 0.00055, 0.000027)$, $\mu_\gamma = 0.00806$, $\alpha_\gamma = 0.8824$ and $\sigma_\gamma = 0.0127$. Projections (red fans) based on central case with modifications to specified parameter values. Top left: $\mu_1 = -0.026$. Top right: $\mu_2 = -0.00045$. Middle left: $\mu_3 = -0.000073$. Middle right: $\mu_\gamma = 0.25806$. Bottom left: $\sigma_\gamma = 0.127$. Bottom right: $\sigma_\gamma = 0.127$ and $\alpha_\gamma = 0.7824$.

7 Results for US males

A full discussion of forecasting results for US males is contained in Appendix B, where we compare and contrast the US and England & Wales results. Our general aim in Appendix B and this section 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. In this section, we focus on model M8 which generates such different results compared with England & Wales males data that we question the validity of M8 for this dataset .

Until now, forecasting results for M8A and M8B have appeared to be satisfactory using England & Wales data. However, Cairns et al. (2007) 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 20, bottom) causes the mortality improvements at ages 75 and 84[20] 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.[21]

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 20, 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 20, bottom). It can be seen from the upper plot in Figure 20 that the fitted $\gamma_c^{(3)}$ process seems to be far more influential when we project US mortality rates than when we project England & Wales mortality rates (Figure 7). Figure 21 shows that the model still produces strange (albeit robust) projections when we vary the sample period used to fit the model.

[20] We use age 84 as rates at age 85 were not available prior to 1980.

[21] A related point concerning M2A and M8A was discussed in subsection 3. There, though, the lack of a second or third age-period component had less serious consequences.

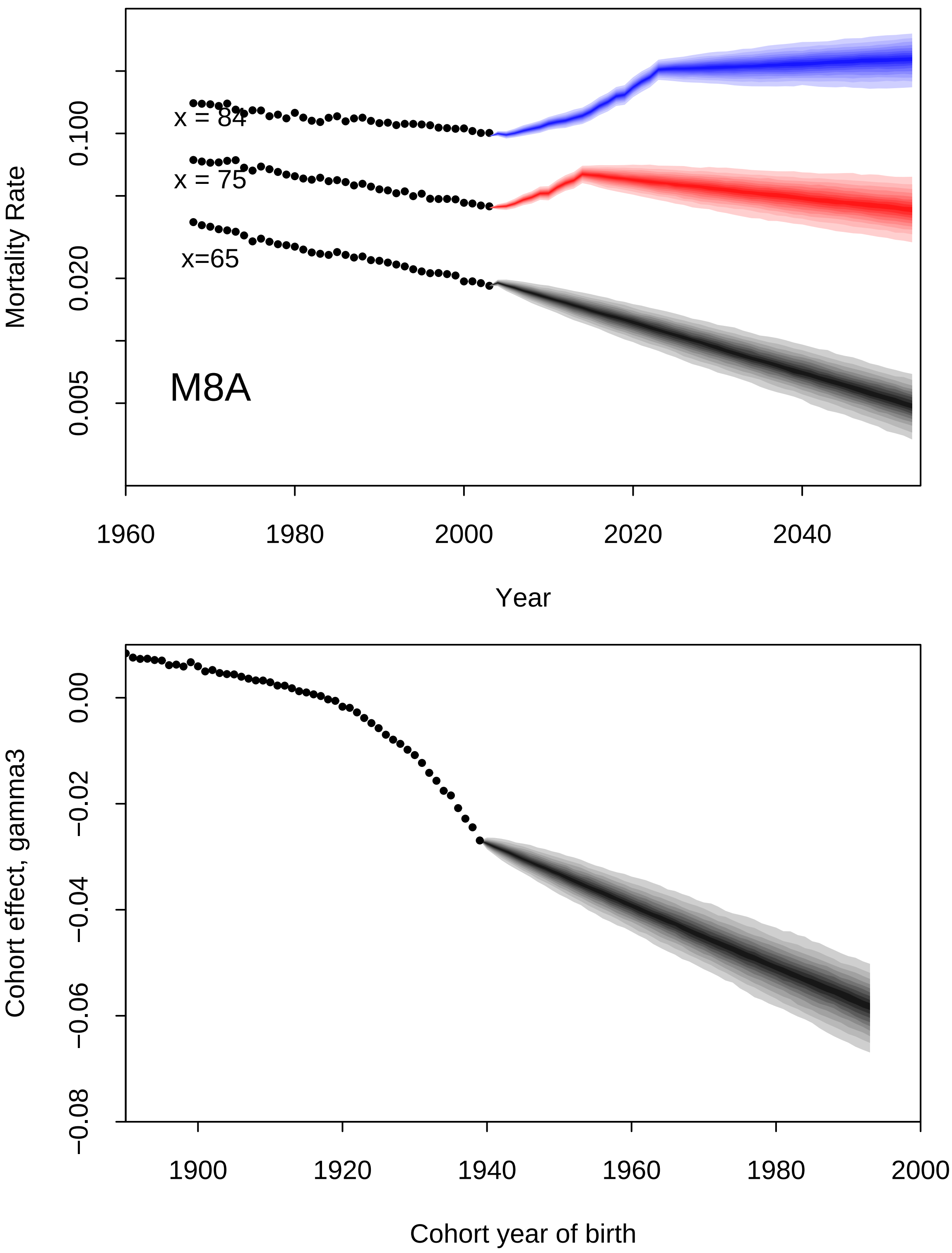


Figure 20: US, males: Top: Fan charts for mortality rates at ages 65, 75 and 84 model M8A with the autoregressive parameter set to $\alpha = 0.9999$. Bottom: Fan charts for the cohort effect, $\gamma_c^{(3)}$, under model M8A with the autoregressive parameter set to $\alpha = 0.9999$.

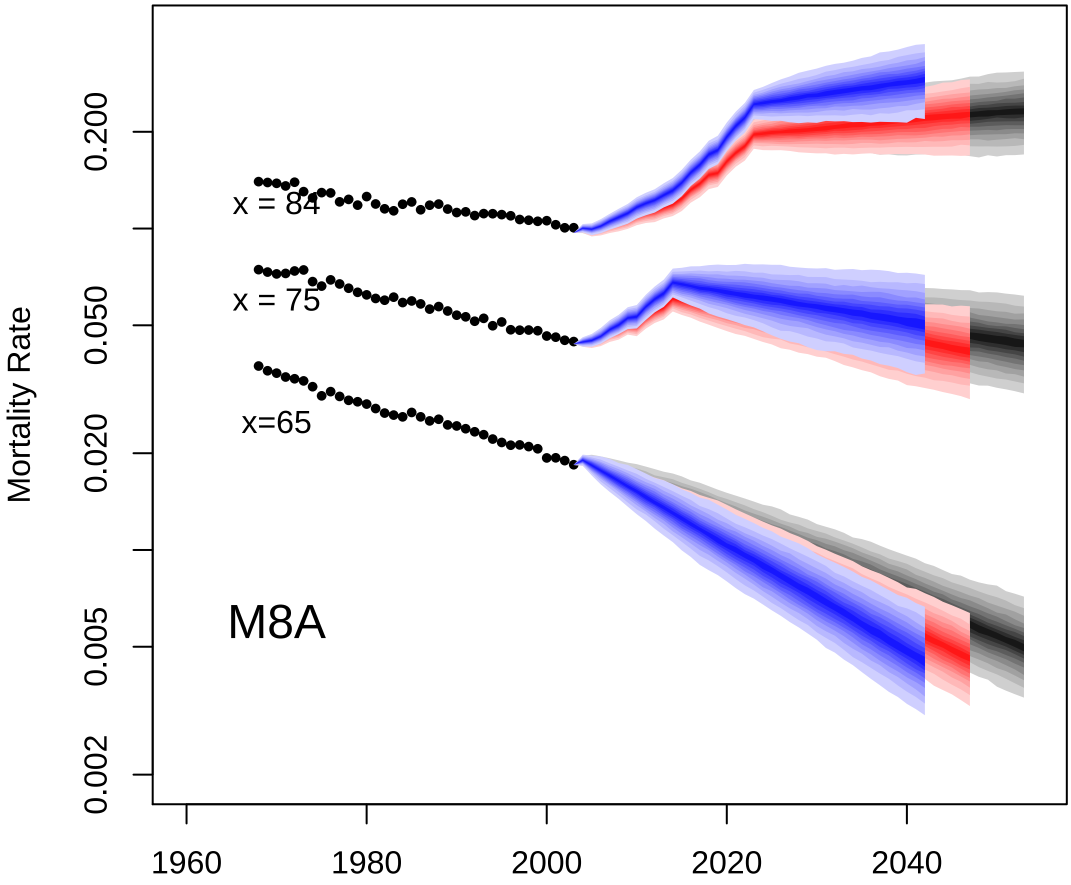


Figure 21: US, males: Model M8A. Cohort effect and mortality rates for ages 65, 75 and 84. Dots and grey fans: historical data from 1968 to 2003 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 36 $\kappa_t^{(i)}$ values and 52 $\gamma_c^{(3)}$ values. Dots and red fans: historical data from 1968 to 2003 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 24 most-recent $\kappa_t^{(i)}$ values and the 45 most-recent $\gamma_c^{(3)}$ values. Blue fans: historical data from 1980 to 2003 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 24 fitted $\kappa_t^{(i)}$ values and the full 45 fitted $\gamma_c^{(3)}$ values.

8 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. 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 England & Wales males mortality, since it had the highest BIC amongst the 8 models we have examined (Cairns et al. (2007, Table 3)). Model M8 is a particular extension of the CBD class of models allowing for a cohort effect. It was 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 data set on this ground alone.

M2 had also been found to fit historical data well (Cairns et al. 2007). 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'. 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 produce biologically plausible results and seems to be a robust model. However, the model's dependence on a single stochastic period effect means that annual changes in mortality rates are all perfectly correlated across ages which may or may not be considered appropriate.
- 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. (2007) 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 data sets of interest since results and conclusions are likely to vary by gender, age range and nationality.

9 References

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

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.

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., and Dowd, K. (2008) "Measurement, modelling and management of mortality risk: A review", To appear in *Scandinavian Actuarial Journal.*

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A., and Balevich, I. (2007) "A quantitative comparison of stochastic mortality models using data from England & Wales and the United States", Pensions Institute Discussion Paper PI-0701, March.

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.

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., and Cairns, A.J.G. (2007) “Facing up to uncertain life expectancy: The longevity fan charts”, Pensions Institute Discussion Paper PI-0703, November.

Dowd, K., Cairns, A.J.G., Blake, D., Coughlan, G.D., Epstein, D., and Khalaf-Allah, M.(2008a) “Backtesting Stochastic Mortality Models: An Ex-Post Evaluation of Multi-Period-Ahead Density Forecasts”, Forthcoming, Pensions Institute Discussion Paper PI-0802.

Dowd, K., Cairns, A.J.G., Blake, D., Coughlan, G.D., Epstein, D., and Khalaf-Allah, M.(2008b) “Evaluating the Goodness of Fit of Stochastic Mortality Models ”, Forthcoming, Pensions Institute Discussion Paper PI-0803.

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

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.

A Identifiability constraints

Some of the models analysed in this paper (following Cairns et al., 2007) involve the use of identifiability constraints to ensure uniqueness of parameter estimates. It is important to know, therefore, what impact, if any, these constraints might have on forecasts. For example, one might ask: *if a different set of constraints had been applied, would forecasts of mortality rates be different?*

When each model is fitted to the historical data, we obtain estimates of the underlying mortality rates, which we shall denote by $\hat{q}(t,x)$, which are functions of the age, period and cohort effects (the $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_{t-x}^{(i)}$) estimated using maximum likelihood. *By construction, a change in the constraints has no effect on the fitted* $\hat{q}(t,x)$.

Stochastic models are then developed to simulate forward the period and cohort effects. These, in turn, are used to calculate the simulated underlying mortality rates $q(t,x)$ for future years.

A.1 Model M1

The constraints are:

$$\begin{aligned} \sum_t \kappa_t^{(2)} &= 0 \\ \text{and} \quad \sum_x \beta_x^{(2)} &= 1. \end{aligned}$$

If different constraints were applied (e.g. $\kappa_{2004}^{(2)} = 0$) then we can make the following remarks:

- The estimated $\beta_x^{(i)}$ and $\kappa_t^{(i)}$ will change and the maximum-likelihood estimates of the parameters of the random walk will change.
- The joint[22] distribution of the forecast $q(t,x)$ will not be affected by the change.

The reason for this is that if the random-walk model is the "right" model under the original constraints, then it is still the "right" model under the revised constraints (albeit with different parameter values).

[22] The joint distribution refers to the probability distribution linking mortality rates at different ages and in different years, as well as the marginal distribution at individual ages and in individual years.

A.2 Model M2B (ARIMA(1,1,0) model for the cohort effect)

The constraints are:

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

If different constraints were applied to $\beta_x^{(2)}$ and $\kappa_t^{(2)}$ then:

- the estimated $\beta_x^{(1)}$, $\beta_x^{(2)}$ and $\kappa_t^{(2)}$ will change and the maximum-likelihood estimates of the parameters of the random-walk model for $\kappa_t^{(2)}$ will change; and
- the joint distribution of the forecast $q(t, x)$ will not be affected by the change.

If different constraints were applied to $\beta_x^{(3)}$ and $\gamma_{t-x}^{(3)}$ then:

- the estimated $\beta_x^{(1)}$, $\beta_x^{(3)}$ and $\gamma_{t-x}^{(3)}$ will change;
- the maximum-likelihood estimates of the parameters of the ARIMA(1,1,0) model for $\gamma_{t-x}^{(3)}$ will *not* change; and
- the joint distribution of the forecast $q(t, x)$ will not be affected by the change.

A.3 Model M5

There are no identifiability constraints.

A.4 Model M7

The constraints are (see Cairns, Blake and Dowd, 2008)

$$\begin{aligned}\sum_{c=c_0}^{c_1} \gamma_c^{(4)} &= 0 \\ \sum_{c=c_0}^{c_1} c\gamma_c^{(4)} &= 0 \\ \sum_{c=c_0}^{c_1} c^2\gamma_c^{(4)} &= 0\end{aligned}$$

where c_0 and c_1 are first and last years of birth that we fit the cohort effect to. The impact of this choice is that if we fit a quadratic function to the estimated $\gamma_c^{(4)}$ then the least squares fit is in fact constant and zero. In practice, this means that $\gamma_c^{(4)}$ will be fluctuating around zero with no discernible linear or quadratic trends.

In more general terms, we have three constraints which can be specified and which might be different from the current version. Different constraints will affect the level, slope and curvature of the fitted $\gamma_c^{(4)}$.

Suppose we make a change to the level of the fitted $\gamma_c^{(4)}$ but not to the slope or curvature. We can continue to fit a random-walk model to the three period effects and an AR(1) model with a non-zero mean-reversion level to the cohort effect. If we do this then:

- the joint distribution of the forecast $q(t, x)$ will not be affected by the change.

If a change in the constraints affects the slope but not the curvature of the fitted $\gamma_c^{(4)}$ then

- the joint distribution of the forecast $q(t, x)$ will not be affected by the change, *provided* we continue to use a random-walk model for the period effects and (more importantly) our stochastic model for $\gamma_c^{(4)}$ is an AR(1) model around a deterministic, linear trend.

In other words, we need to modify our stochastic model for $\gamma_c^{(4)}$.

If a change in the constraints affects the curvature of $\gamma_{t-x}^{(4)}$ then:

- the joint distribution of the forecast $q(t, x)$ *will* be affected by the change.

The change in future dynamics relates to the impact of the change of constraint on the period effects. Specifically, a change in the curvature of $\gamma_{t-x}^{(4)}$ also changes the curvature of the fitted period effect $\kappa_t^{(1)}$. If a random-walk model for $\kappa_t^{(1)}$ was the "right" model under the original constraints, the "right" model under the new constraints will not be a random-walk model.

A.5 Model M8

There is one constraint:

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

Thus a change in the constraint will not have an impact on fitted or projected mortality rates provided M8A continues to use an AR(1) model around a deterministic linear trend, and provided M8B continues to use an AR(1) model around a non-zero mean.

B Figures for the US males data, 1968 to 2003

In this Appendix, we analyse the US males data from 1968 to 2003, and provide similar figures and discussion to that for England & Wales considered in the main text:

- In Figure 22, by way of reminder, we repeat certain outputs from the anlysis in Cairns et al. (2007): the cohort effect for models M2, M3, M7 and M8, with the corresponding age effects. For M2, the age effect is a non-parametric function estimated at each age; for M8, the age affect is assumed to be linear in age; for M3, the age effect for component 3 is assumed to be constant; and for M7, the age effect for component 4 is assumed to be constant.
  - M7: $\gamma_c^{(4)}$ is more random here than we saw for the England & Wales data (Figure 1). Nevertheless, there is a discernible pattern that might be consistent with $\gamma_c^{(4)}$ being an AR(1) process, although other models might also be suitable. This relative randomness in $\gamma_c^{(4)}$ supports the view that the US has a much less significant cohort effect than England and Wales for this gender and age range.
  - M2: Unlike the England & Wales data, $\gamma_c^{(3)}$ exhibits a strong linear trend (Figure 22).
  - M3: Unlike the England & Wales data, the shape of $\gamma_c^{(3)}$ is quite different from the M2 cohort effect. However, the shape of $\gamma_c^{(3)}$ is relatively similar to the England & Wales cohort effect (Figure 1).
  - M8: In contrast with the England & Wales data, the magnitude of $\beta_x^{(3)}$ is increasing with age.
- In Figure 23, we have plotted fan charts for the cohort effect for models M2A, M2B, M7, M8A and M8B using the same stochastic models in each case as for the England & Wales data.
  - M7: the fitted AR(1) mean reversion value is very high, implying that the fans very quickly reach their stationary limits.
  - M2A and M2B: in contrast with the England & Wales plots, the fans for $\gamma_c^{(3)}$ seem more acceptable in terms of their spread. The two fans look similar, having a similar width and rate of expansion. We might speculate that the models are almost identical since an ARIMA(1,1,0) model (M2B – equation 2) with $\alpha = 1$ is identical to an ARIMA(0,2,1) model (M2A – equation 2) with $\alpha = 0$. However, a detailed look at the M2A and M2B output reveals parameter estimates that are quite far from this situation: that is, the fitted M2A and M2B are structurally quite different even though their outputs are quite similar.

  - M3A and M3B: here, the fans mimic the pattern that emerged when we analysed the England & Wales data. M3A, in particular, has fans that widen out at a fast rate without limit, making the results produced by M3A seem potentially unreasonable for this dataset.
  - M8A and M8B: these reveal substantial differences. M8A is mean reverting around an obvious linear trend. M8B involves mean reversion to a constant level rather than to a linear trend. This is not immediately obvious from the output, but is explained by the fact that the mean-reversion level is around -6, with a very slow mean-reversion rate.

    Since M8B seems to produces unreasonable results for the dataset under consideration, it was decided not to include it in the remaining plots.

- In Figures 24 and 25, we have plotted fan charts under each model for mortality rates at ages 65, 75 and 84.[23] Similar comments to the England & Wales data apply here, except for M8A. The problems with M8A are discussed in Section 7.

  The overlaid fans reveal bigger differences between models than the England & Wales data: the variation in the width of the fans and in their central trends seem to vary more between models.

- Figure 26 indicates that there is relatively little difference between M2A and M2B when we look at projected mortality rates.

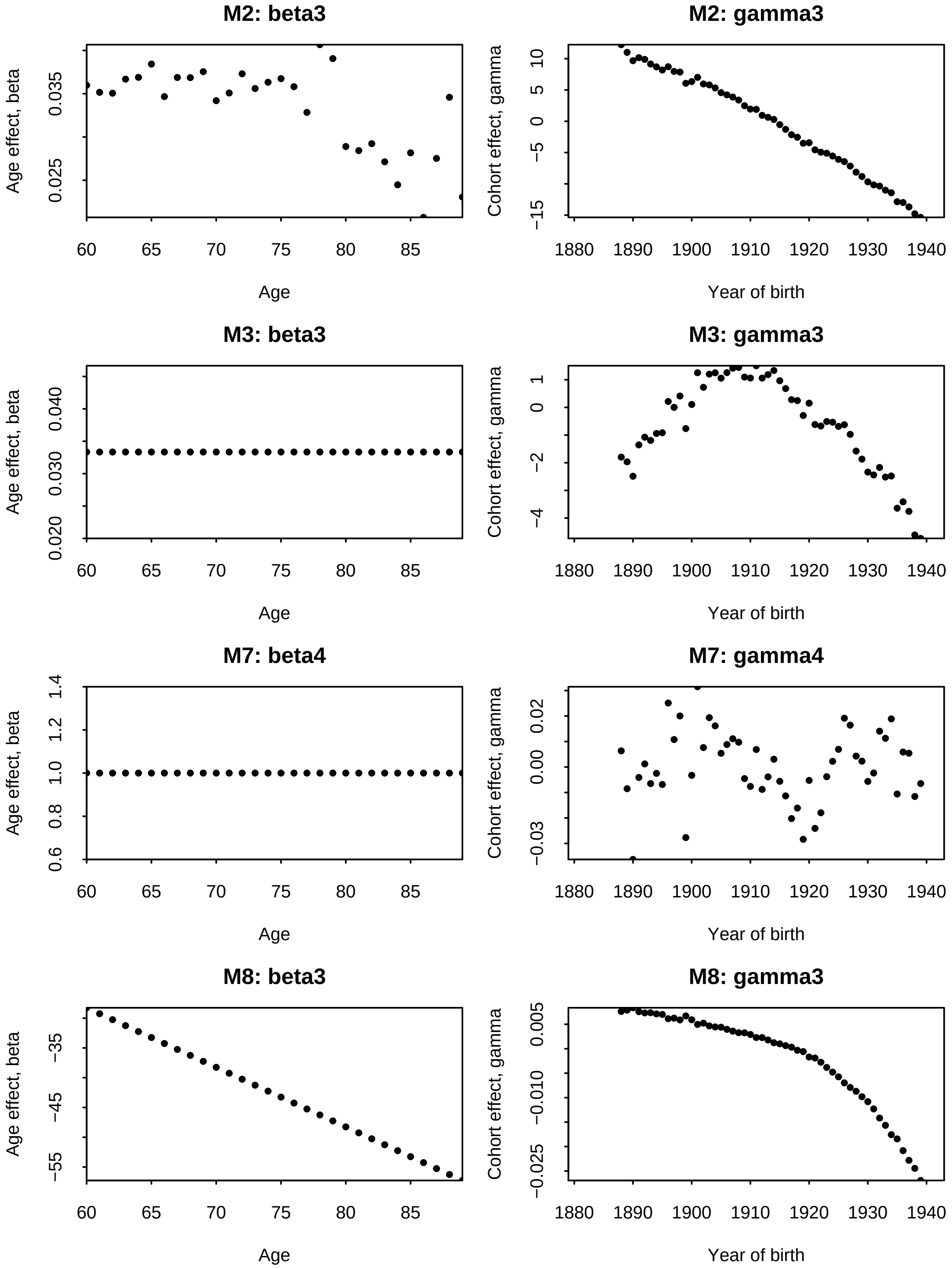
  In contrast, Figure 27 reveals that the choice between M3A and M3B has a significant impact on forecast rates of mortality. The results for M3A reveal what might be considered to be substantial uncertainty in future mortality rates.

- In Figures 28 to 35, we look at the robustness of model projections relative to the sample period used when models are fitted to the historical data and to the number of fitted $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$ used to estimate the forecasting model parameters. We find that our conclusions are similar to those for the England & Wales data. Mortality projections under models M1, M3, M5 and M7 look robust, while those under M2A and M2B are less so. For M8A, although the projections look very peculiar for ages 75 and 85, the results look reasonably robust, albeit not as robust as M1, M3, M5 and M7.

- In Figures 36 and 37, we have plotted fans for the survivor index for a cohort of US males aged 65 in 2003 under M1, M2B, M3B, M5, M7 and M8A. Note that, in all cases, the model for the cohort effect is irrelevant here since a single value of $\gamma_c^{(3)}$ required for this cohort has already been estimated from the historical data. From these models, M2A, and M3B are slightly high compared with

[23] We use age 84 here as historical rates at age 85 and above are not available for 1968 to 1979.

the rest, and M1 is narrower. But M8A the fan chart reflects the fact that mortality rates at higher ages under M8A rapidly depart from their historical trends (Figure 35).


M2: beta3

M2: gamma3

M3: beta3

M3: gamma3

M7: beta4

M7: gamma4

M8: beta3

M8: gamma3

Figure 22: US males: Fitted cohort effects for models M2, M3, M7 and M8.

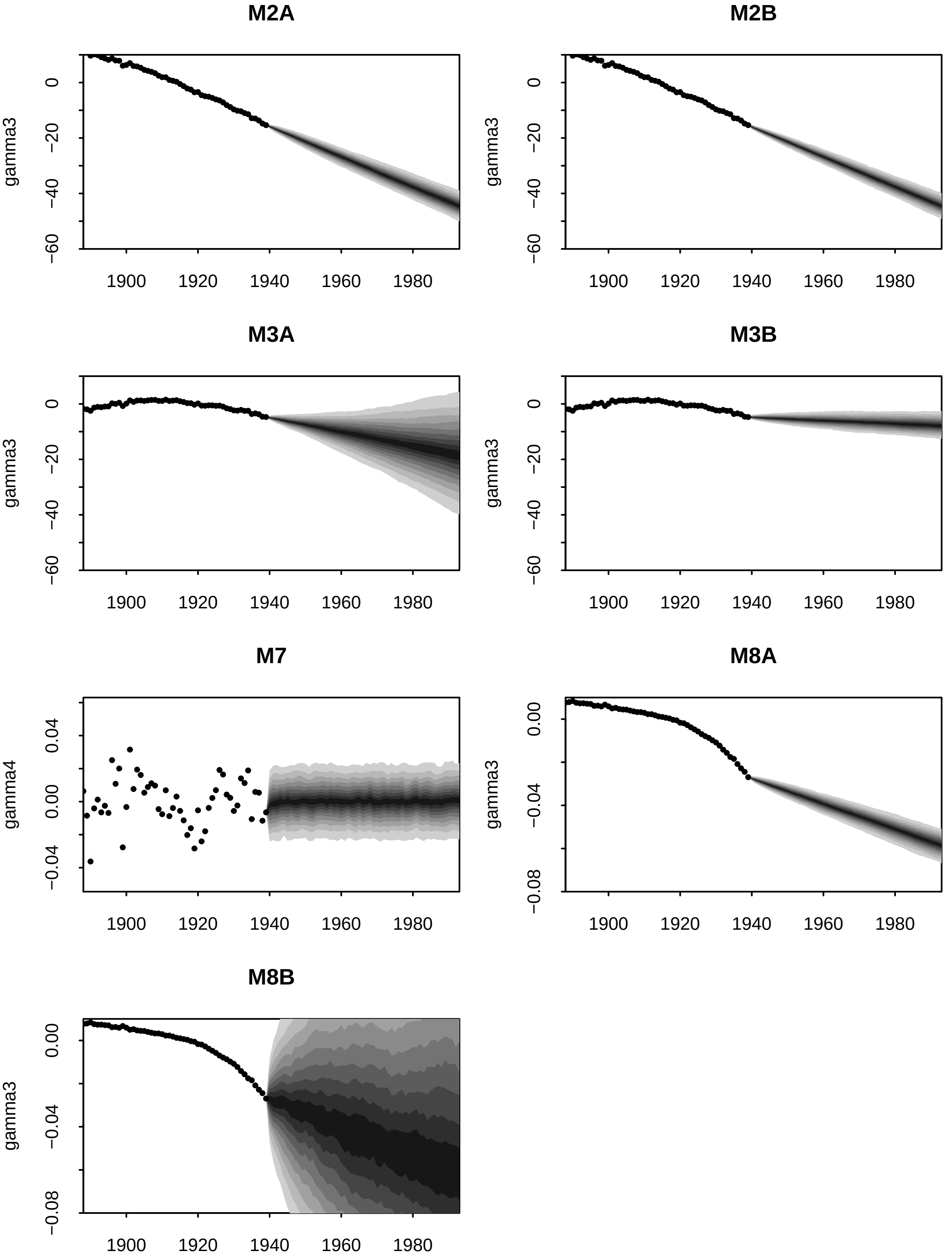


Figure 23: US males: Fan charts for the projected cohort effect. For M1 and M5, there is no cohort effect so no fan charts have been plotted.

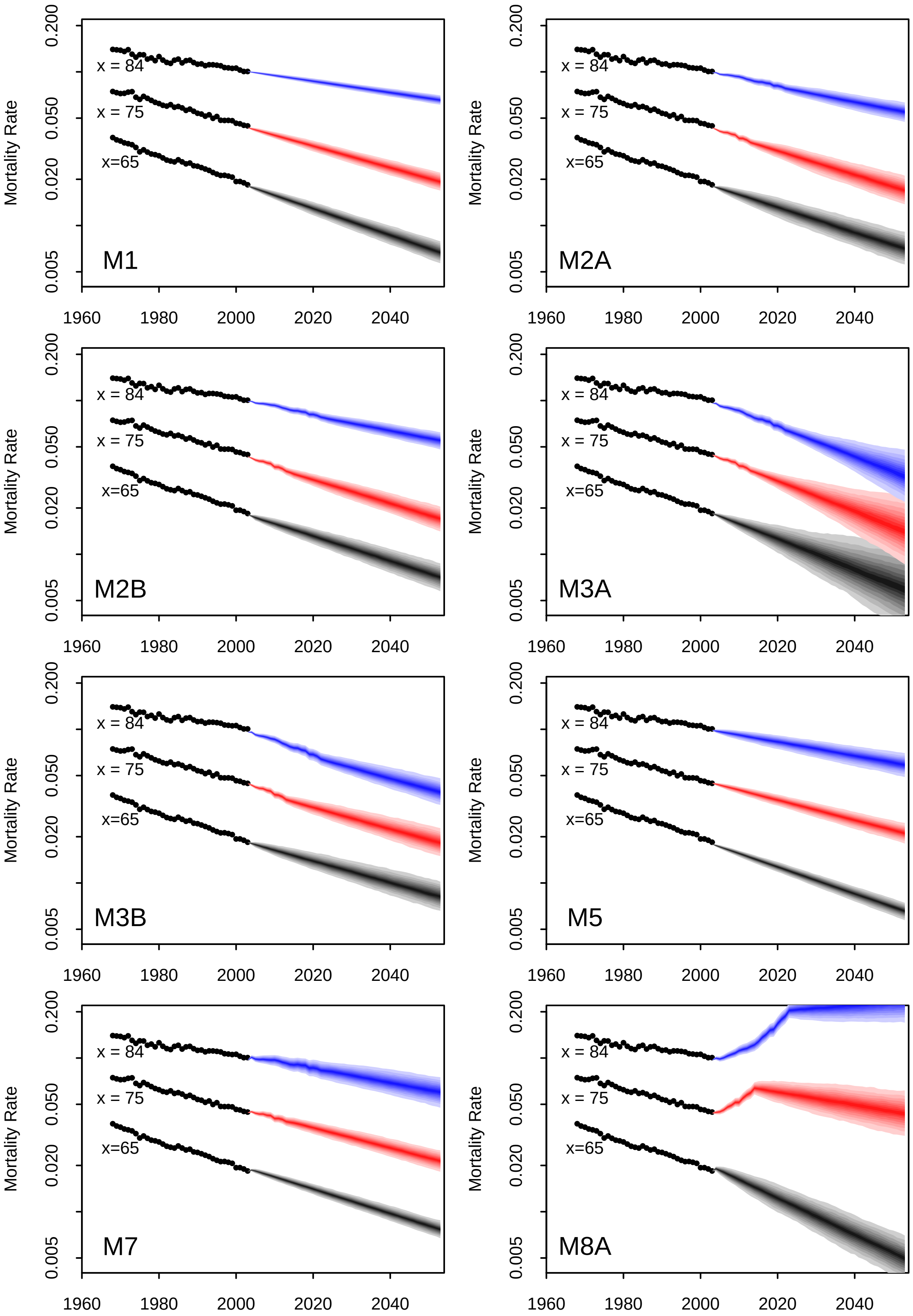


Figure 24: US males: Mortality rates, $q(t, x)$, for models M1, M2A, M2B, M5, M7 and M8A for ages $x = 65$ (grey), 75 (red), and 84 (blue). The dots show historical mortality rates for 1968 to 2003.

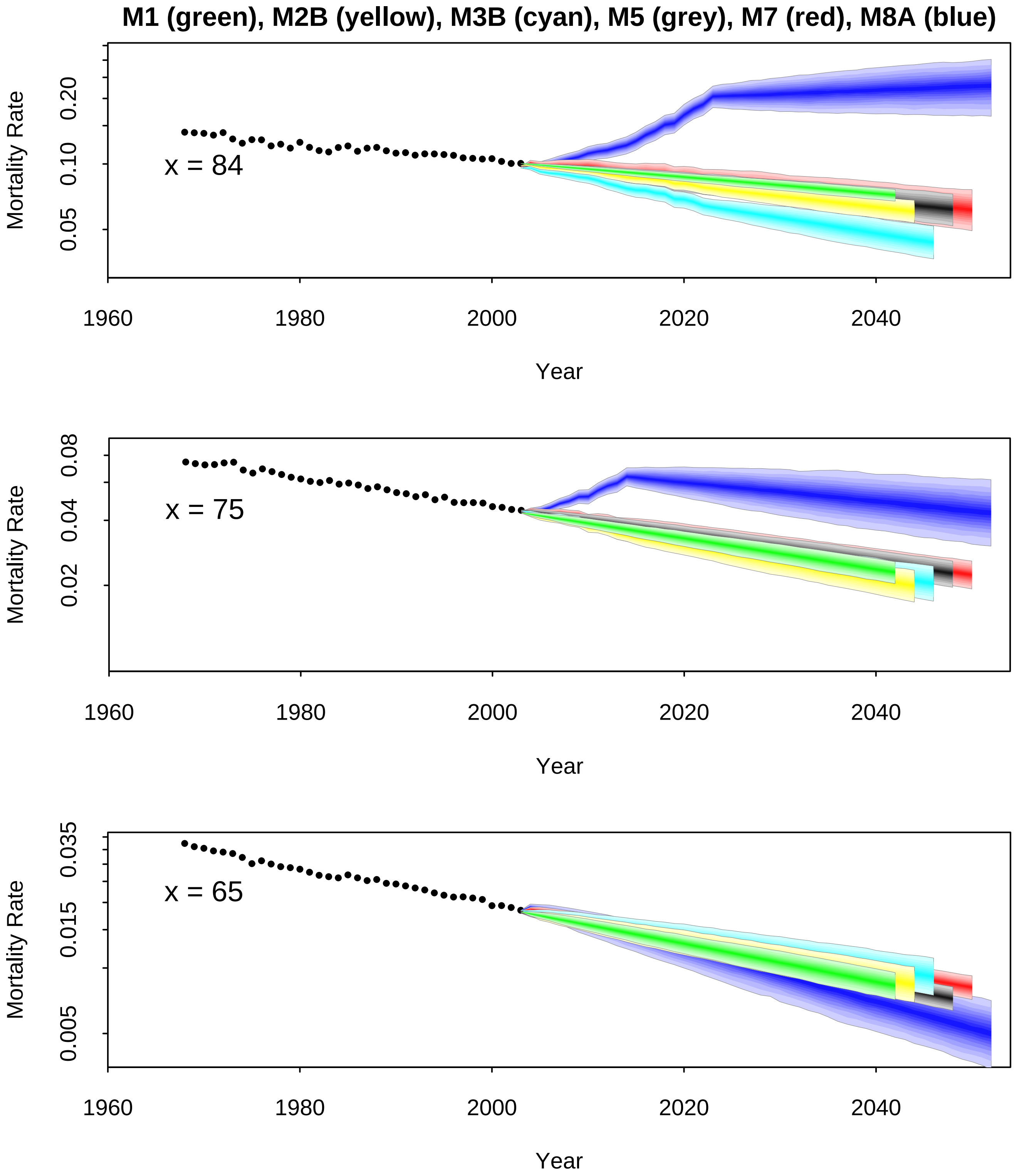


Figure 25: US, males: Mortality rates, $q(t, x)$, for models M1 (green), M2B (yellow), M3B (cyan), M5(grey), M7 (red), and M8A (blue) with fans overlaid for ages $x = 65$, 75, and 84. The dots show historical mortality rates for 1968 to 2003.

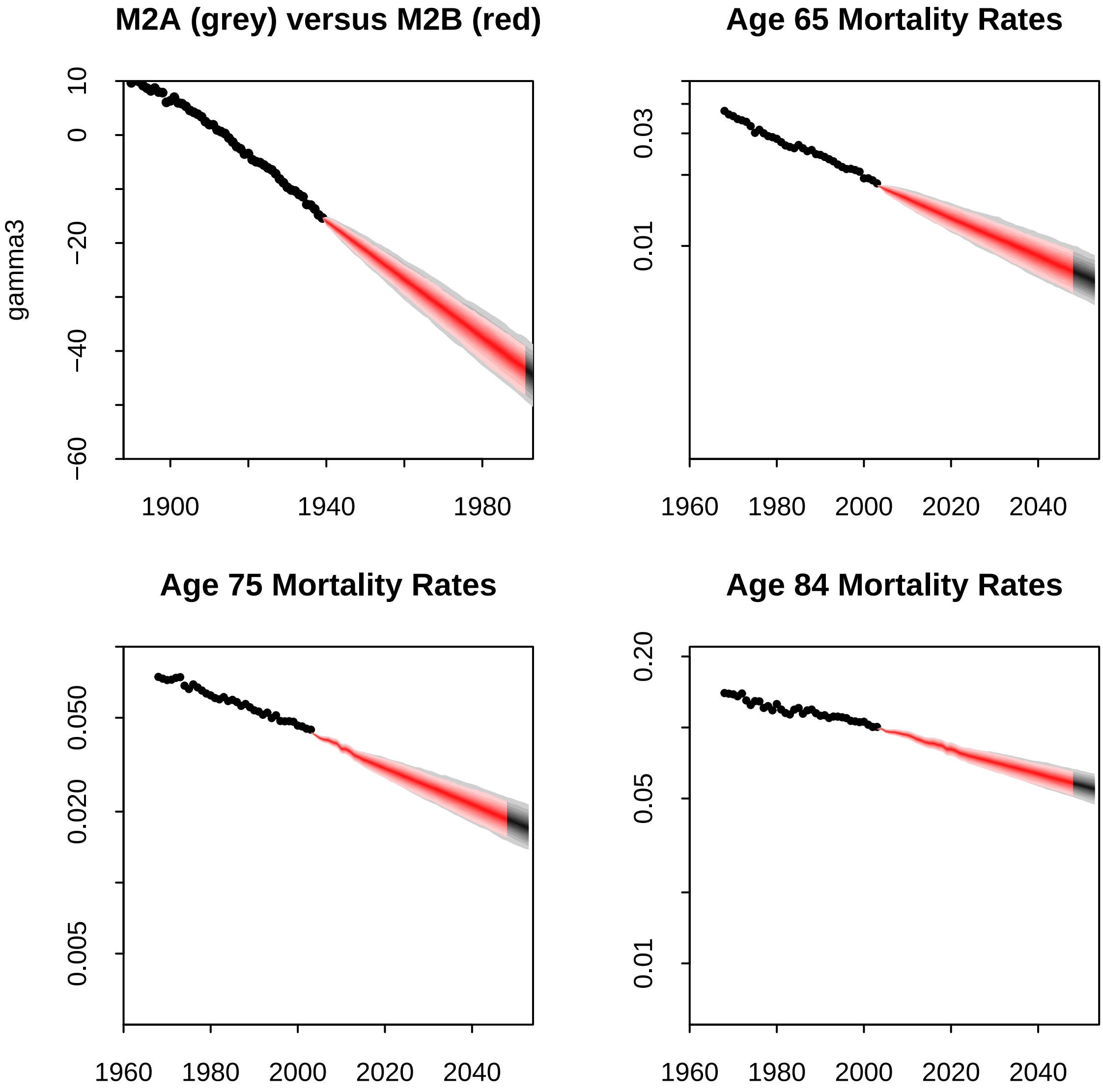


Figure 26: US, males: Fan charts comparing models M2A (grey fans) and M2B (red fans). Top left: historical (dots) and forecast (fans) values for the cohort effect, $\gamma_c^{(3)}$. Top right, bottom left and right: historical (dots) and forecast (fans) mortality rates, $q(t,x)$, for ages 65, 75 and 84.

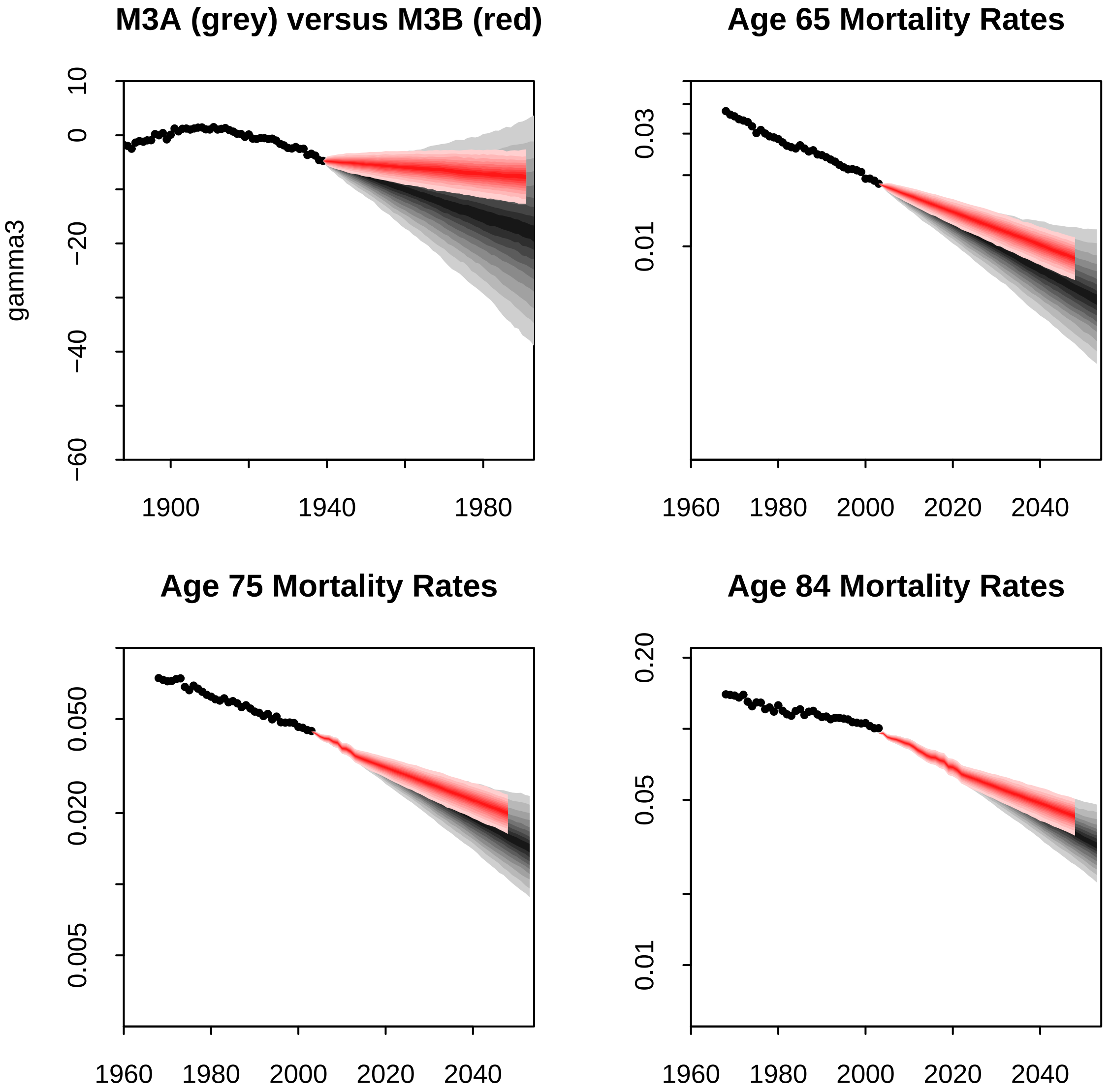


Figure 27: US, males: Fan charts comparing models M3A (grey fans) and M3B (red fans). Top left: historical (dots) and forecast (fans) values for the cohort effect, $\gamma_c^{(3)}$. Top right, bottom left and right: historical (dots) and forecast (fans) mortality rates, $q(t,x)$, for ages 65, 75 and 84.

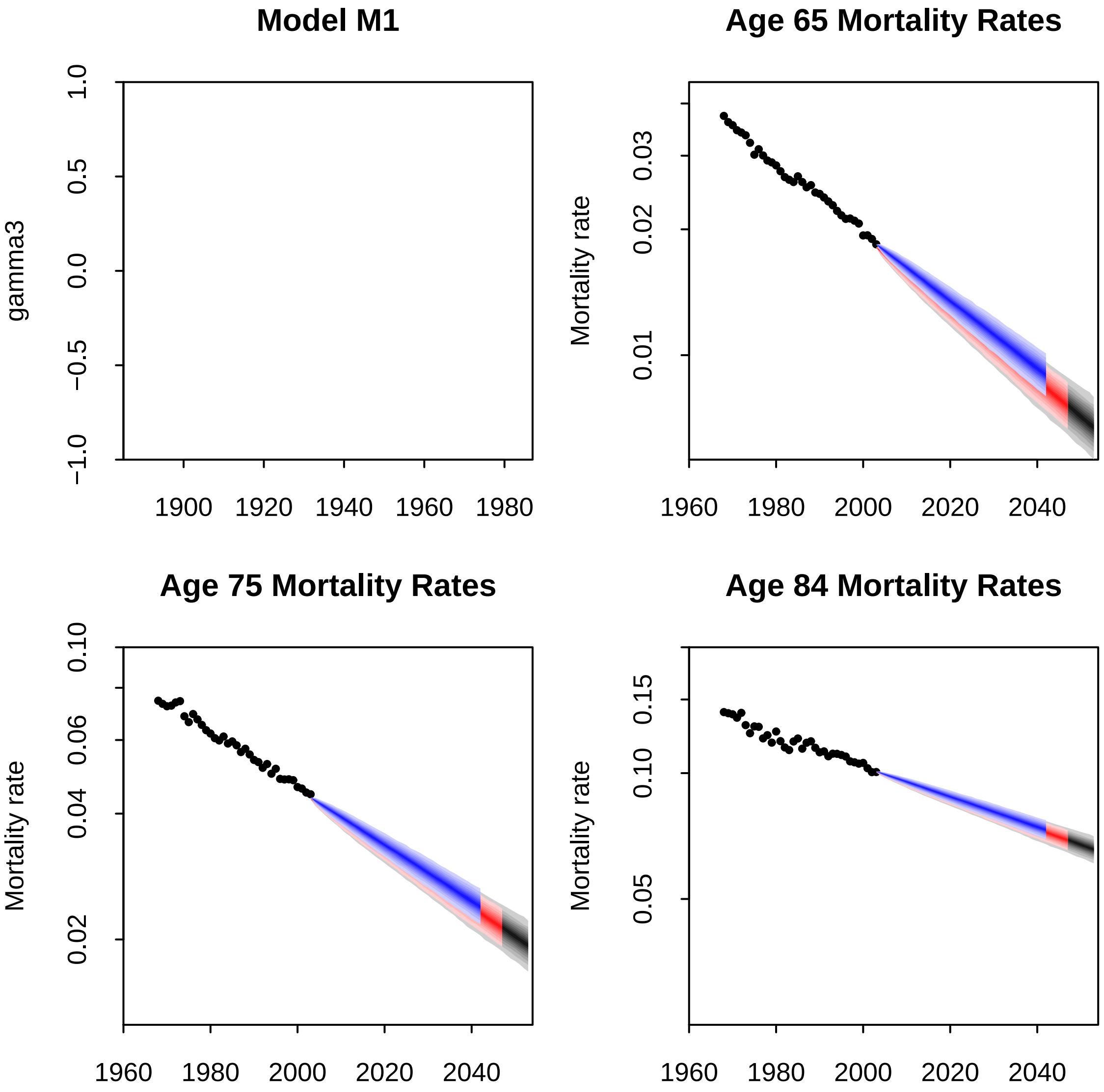


Figure 28: US, males: Model M1. Cohort effect (absent for this model) and mortality rates for ages 65, 75 and 84. Dots and grey fans: historical data from 1968 to 2003 used to estimate the historical $\kappa_t^{(2)}$; forecasting model uses the 36 $\kappa_t^{(2)}$ values. Dots and red fans: historical data from 1968 to 2003 used to estimate the historical $\kappa_t^{(2)}$; forecasting model uses the 24 most recent $\kappa_t^{(2)}$ values. Blue fans: historical data from 1980 to 2003 used to estimate the historical $\kappa_t^{(2)}$; forecasting model uses the full 24 $\kappa_t^{(2)}$ values.

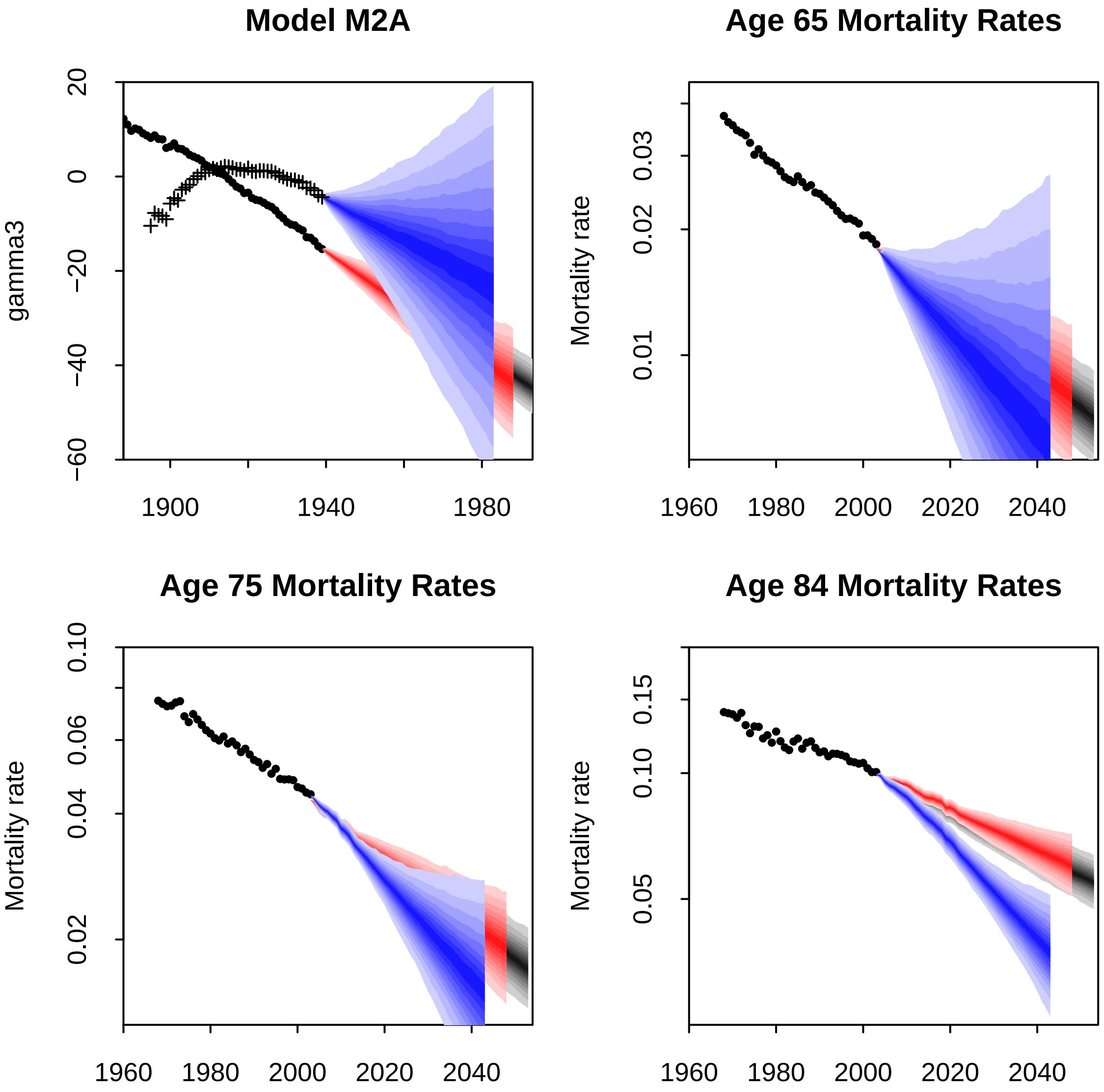


Figure 29: US, males: Model M2A. Cohort effect and mortality rates for ages 65, 75 and 84. Dots and grey fans: historical data from 1968 to 2003 used to estimate the historical $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 36 $\kappa_t^{(2)}$ values and the 52 $\gamma_c^{(3)}$ values. Dots and red fans: historical data from 1968 to 2003 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 blue fans: historical data from 1980 to 2003 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.

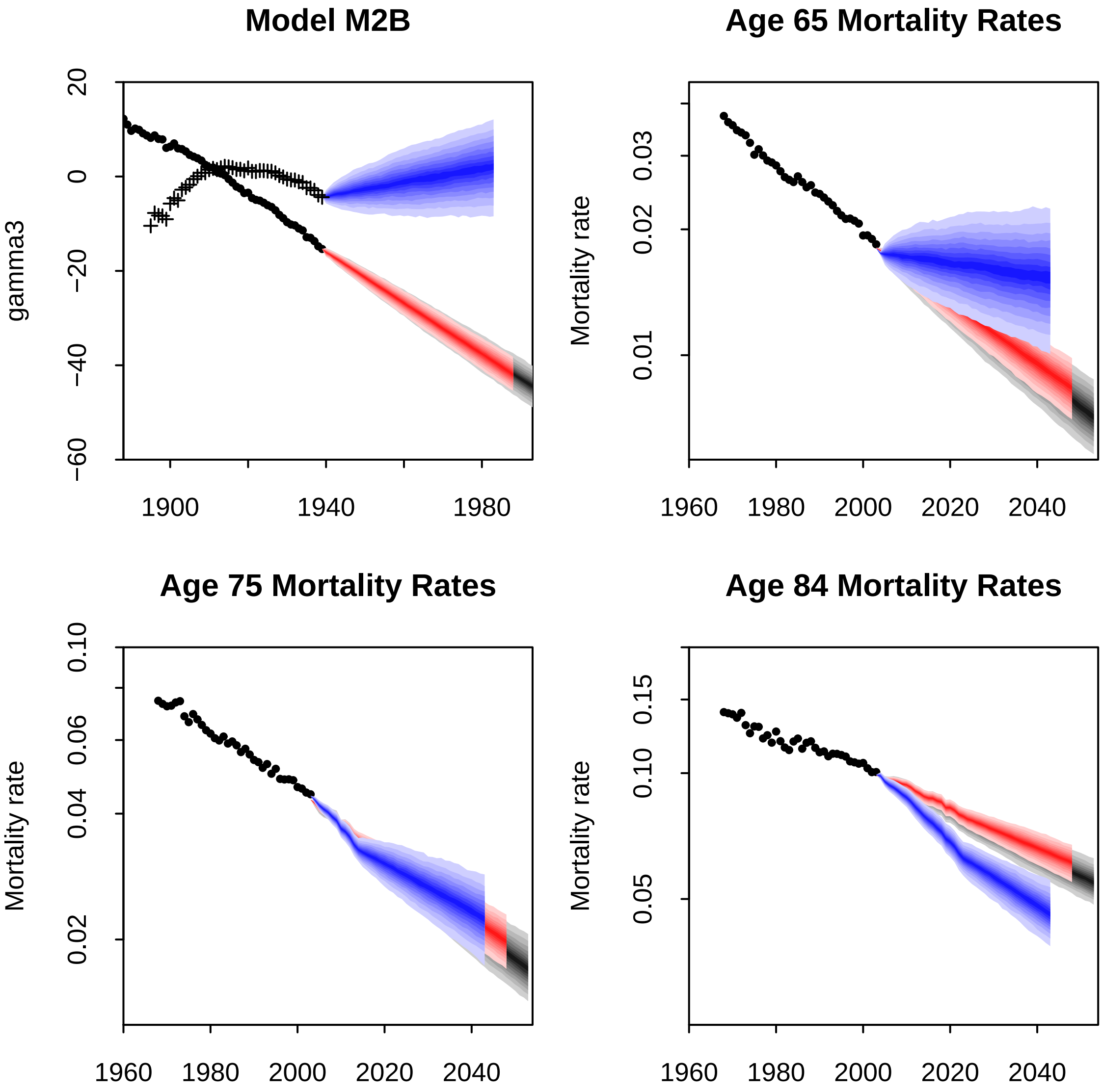


Figure 30: US, males: Model M2B. Cohort effect and mortality rates for ages 65, 75 and 84. Dots and grey fans: historical data from 1968 to 2003 used to estimate the historical $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 36 $\kappa_t^{(2)}$ values and the 52 $\gamma_c^{(3)}$ values. Dots and red fans: historical data from 1968 to 2003 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 blue fans: historical data from 1980 to 2003 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.

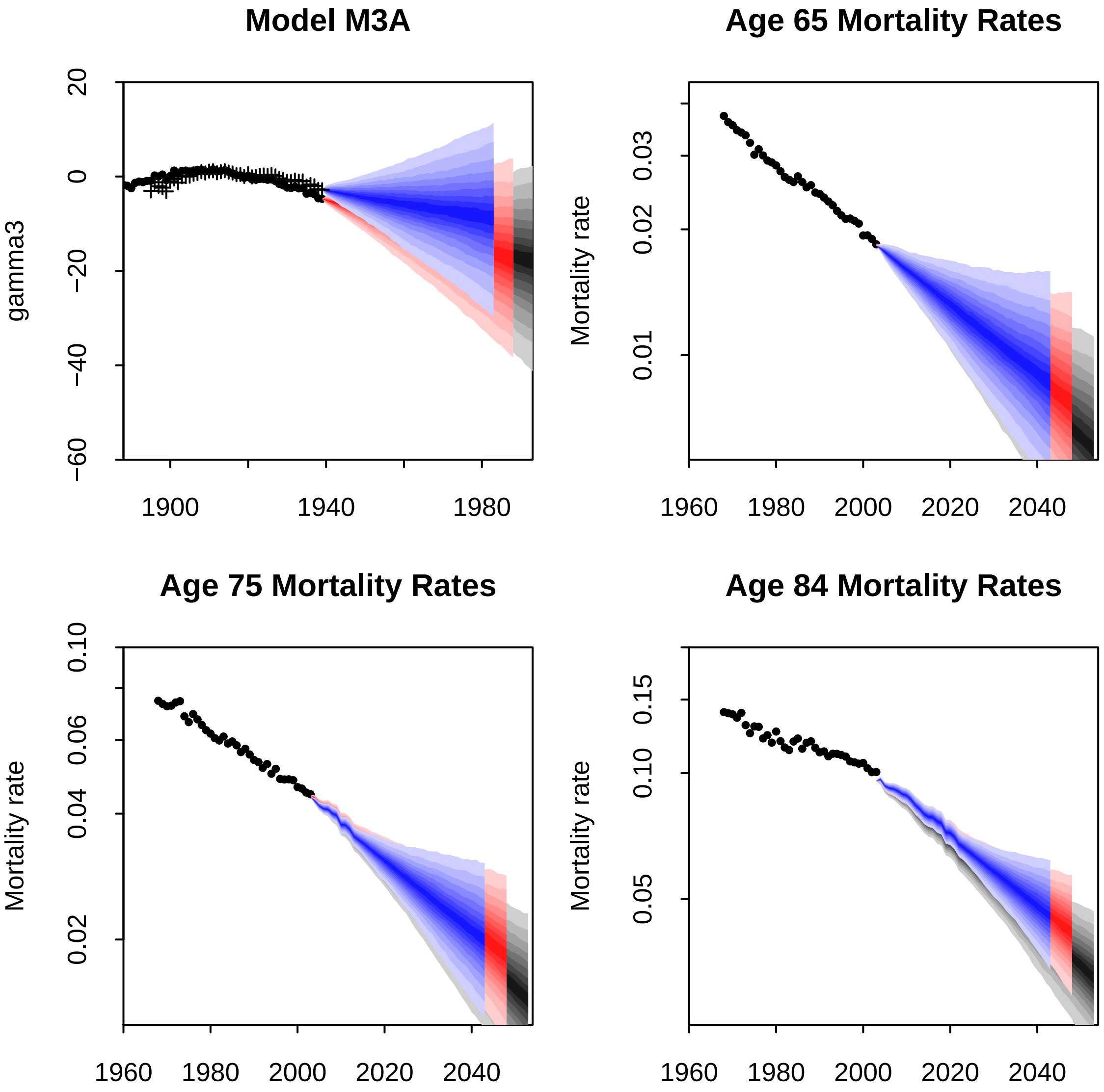


Figure 31: US, males: Model M3A. Cohort effect and mortality rates for ages 65, 75 and 84. Dots and grey fans: historical data from 1968 to 2003 used to estimate the historical $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 36 $\kappa_t^{(2)}$ values and the 52 $\gamma_c^{(3)}$ values. Dots and red fans: historical data from 1968 to 2003 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 blue fans: historical data from 1980 to 2003 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.

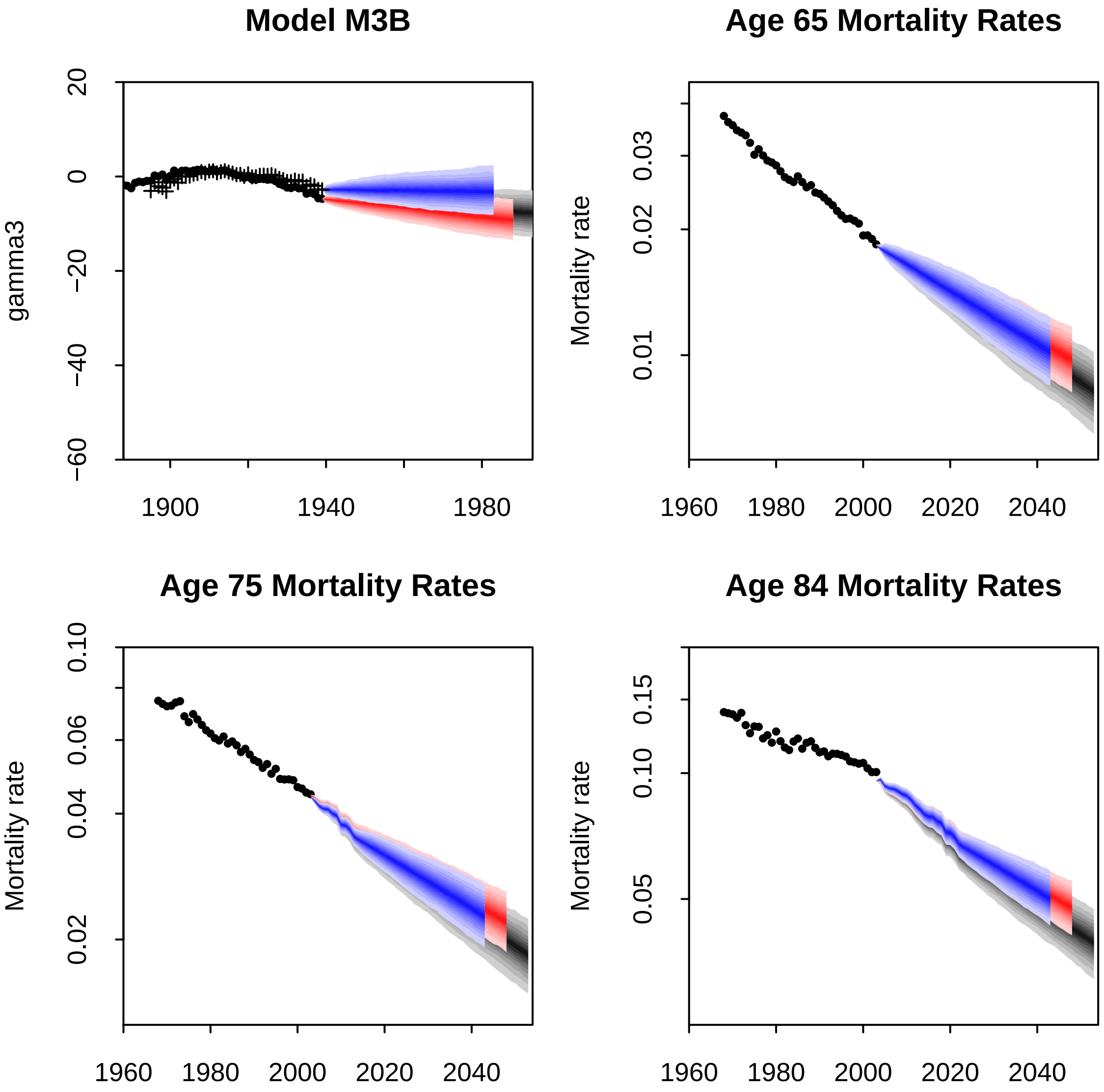


Figure 32: US, males: Model M3B. Cohort effect and mortality rates for ages 65, 75 and 84. Dots and grey fans: historical data from 1968 to 2003 used to estimate the historical $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 36 $\kappa_t^{(2)}$ values and the 52 $\gamma_c^{(3)}$ values. Dots and red fans: historical data from 1968 to 2003 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 blue fans: historical data from 1980 to 2003 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.

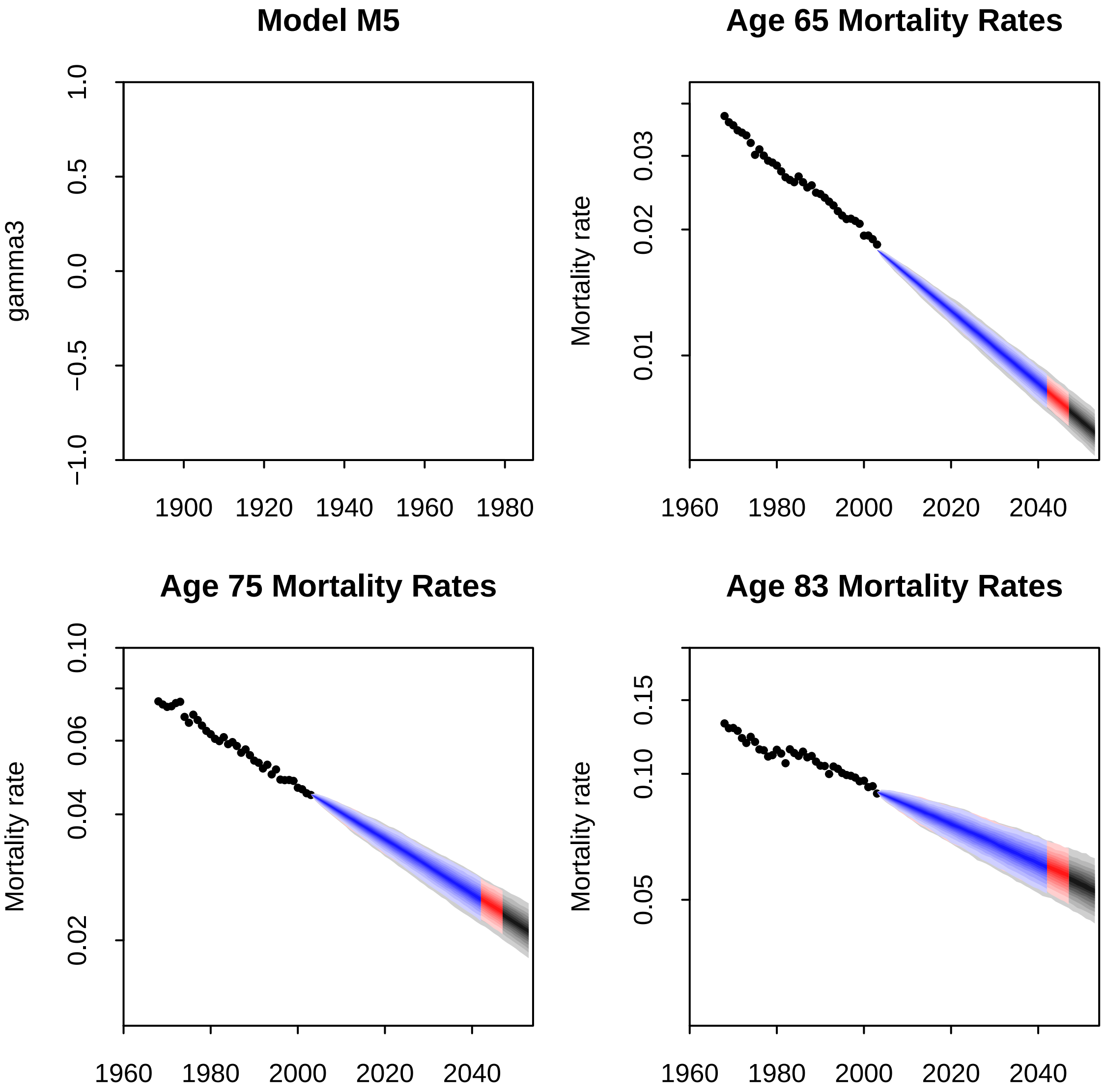


Figure 33: US, males: Model M5. Cohort effect (absent for M5) and mortality rates for ages 65, 75 and 84. Dots and grey fans: historical data from 1968 to 2003 used to estimate the historical $\kappa_t^{(i)}$; forecasting model uses the 36 $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ values. Dots and red fans: historical data from 1968 to 2003 used to estimate the historical $\kappa_t^{(i)}$; forecasting model uses the 24 most-recent $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ values. Blue fans: historical data from 1980 to 2003 used to estimate the historical $\kappa_t^{(i)}$; forecasting model uses the full 24 $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ values.

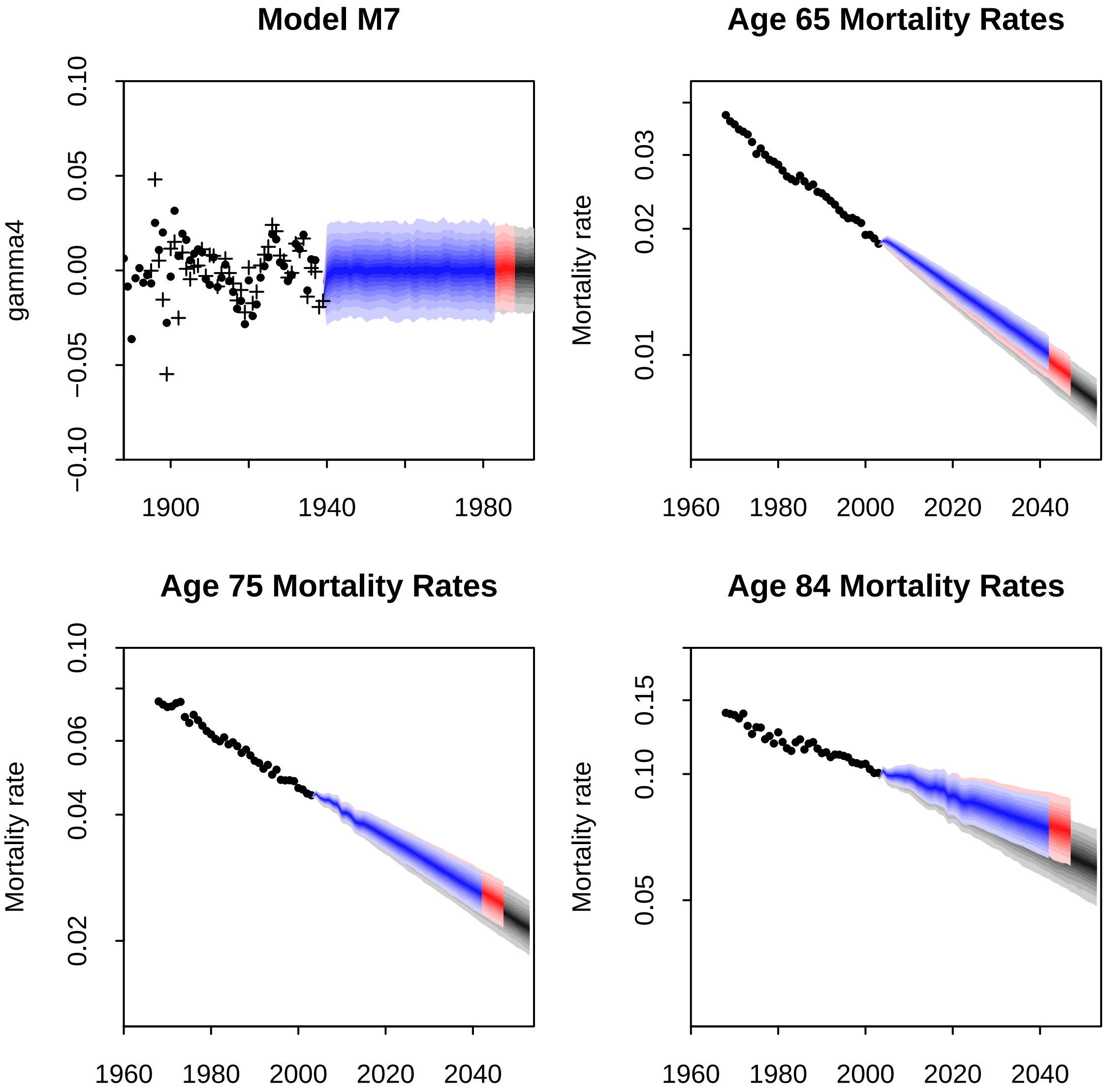


Figure 34: US, males: Model M7. Cohort effect and mortality rates for ages 65, 75 and 84. Dots and grey fans: historical data from 1968 to 2003 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 36 $\kappa_t^{(i)}$ values and 52 $\gamma_c^{(4)}$ values. Dots and red fans: historical data from 1968 to 2003 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 24 most-recent $\kappa_t^{(i)}$ values and the 45 most-recent $\gamma_c^{(4)}$ values. Crosses and blue fans: historical data from 1980 to 2003 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 24 fitted $\kappa_t^{(i)}$ values and the full 45 fitted $\gamma_c^{(4)}$ values.

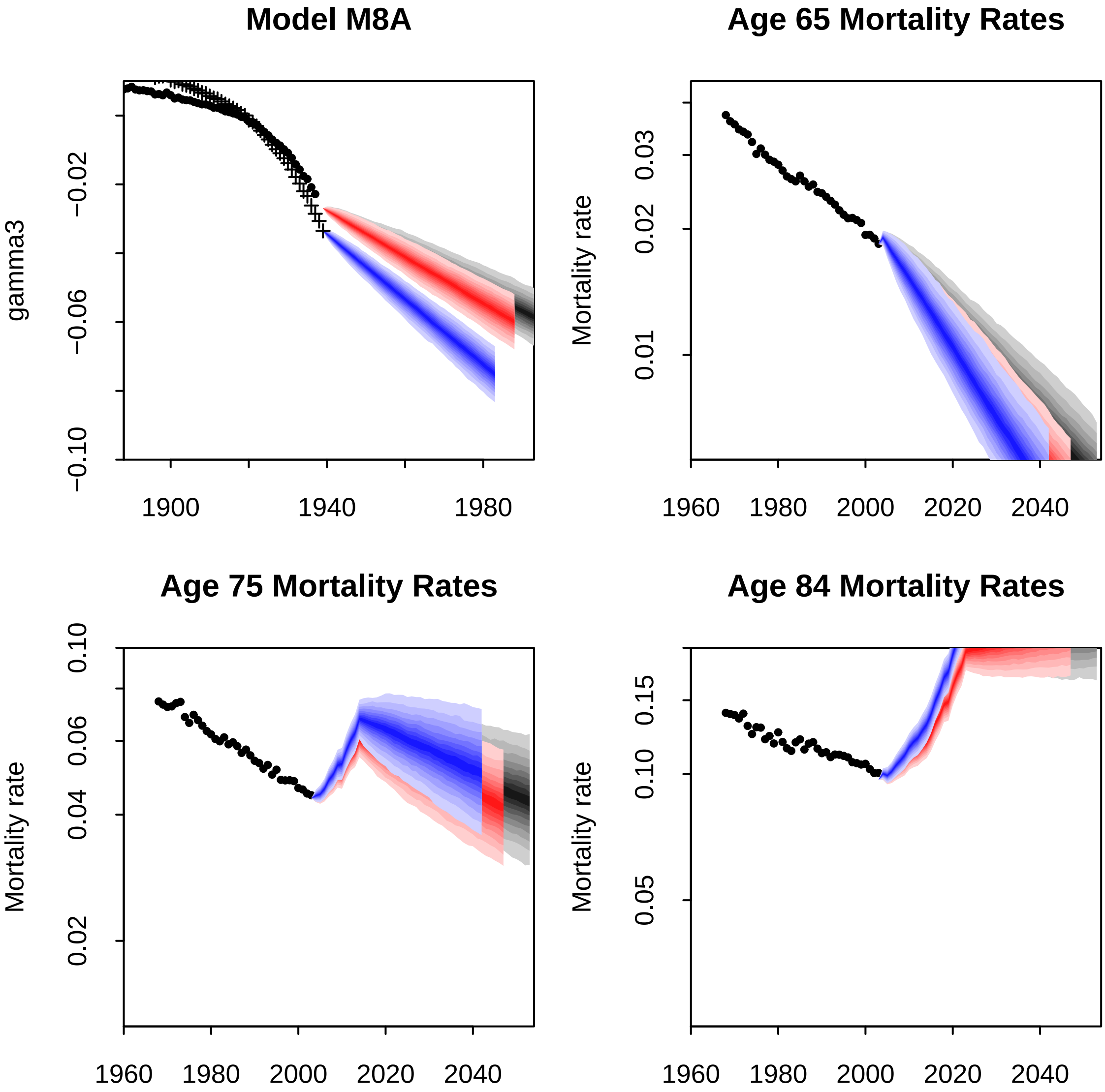


Figure 35: US, males: Model M8A. Cohort effect and mortality rates for ages 65, 75 and 84. Dots and grey fans: historical data from 1968 to 2003 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 36 $\kappa_t^{(i)}$ values and 52 $\gamma_c^{(3)}$ values. Dots and red fans: historical data from 1968 to 2003 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the 24 most-recent $\kappa_t^{(i)}$ values and the 45 most-recent $\gamma_c^{(3)}$ values. Crosses and blue fans: historical data from 1980 to 2003 used to estimate the historical $\kappa_t^{(i)}$ and $\gamma_c^{(i)}$; forecasting model uses the full 24 fitted $\kappa_t^{(i)}$ values and the full 45 fitted $\gamma_c^{(3)}$ values.

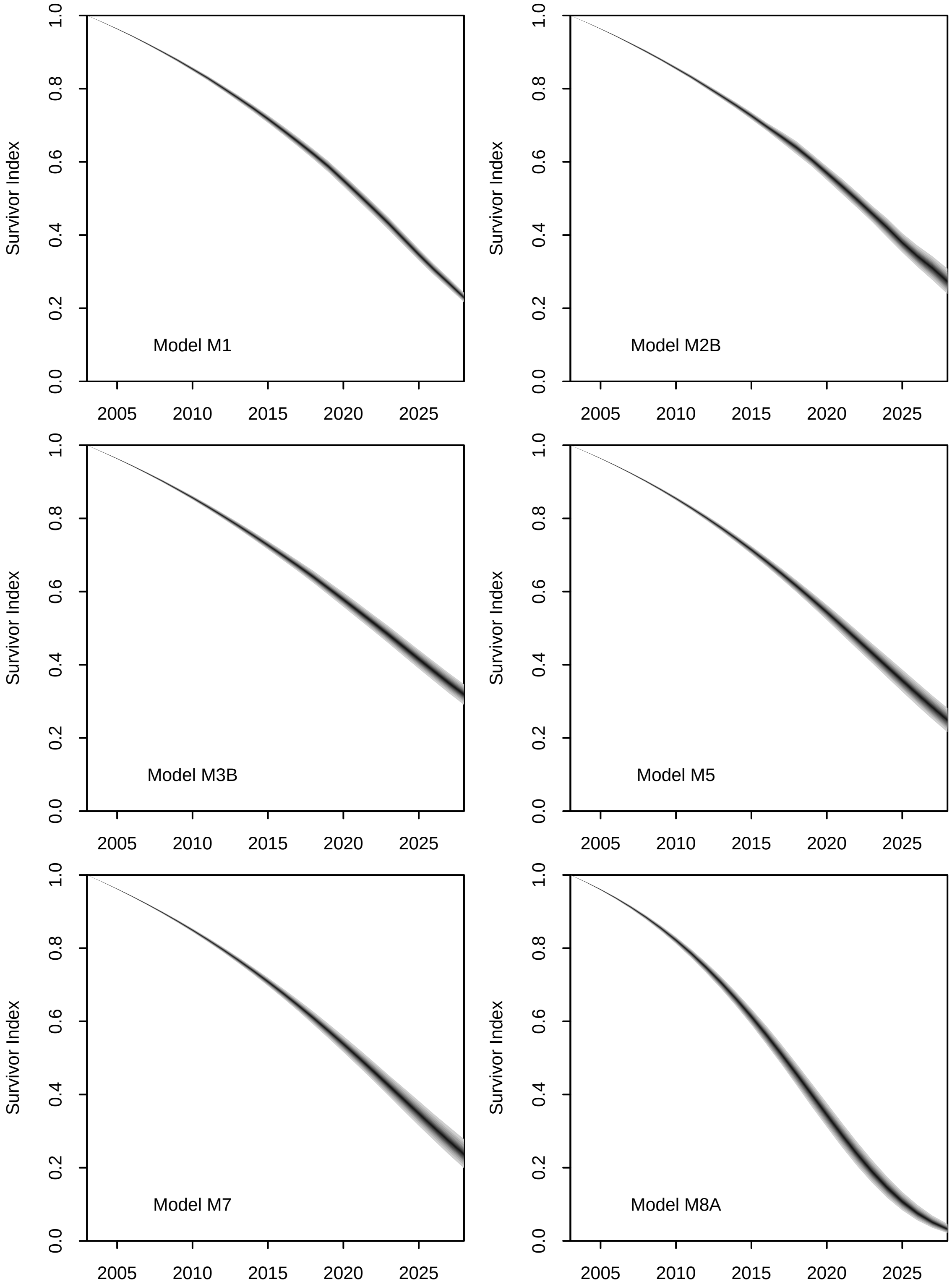


Figure 36: US, males: Fan charts for the survivor index $S(t, 65)$ for the cohort aged 65 at the start of 2005, for models M1, M2B, M3B, M5, M7 and M8A.

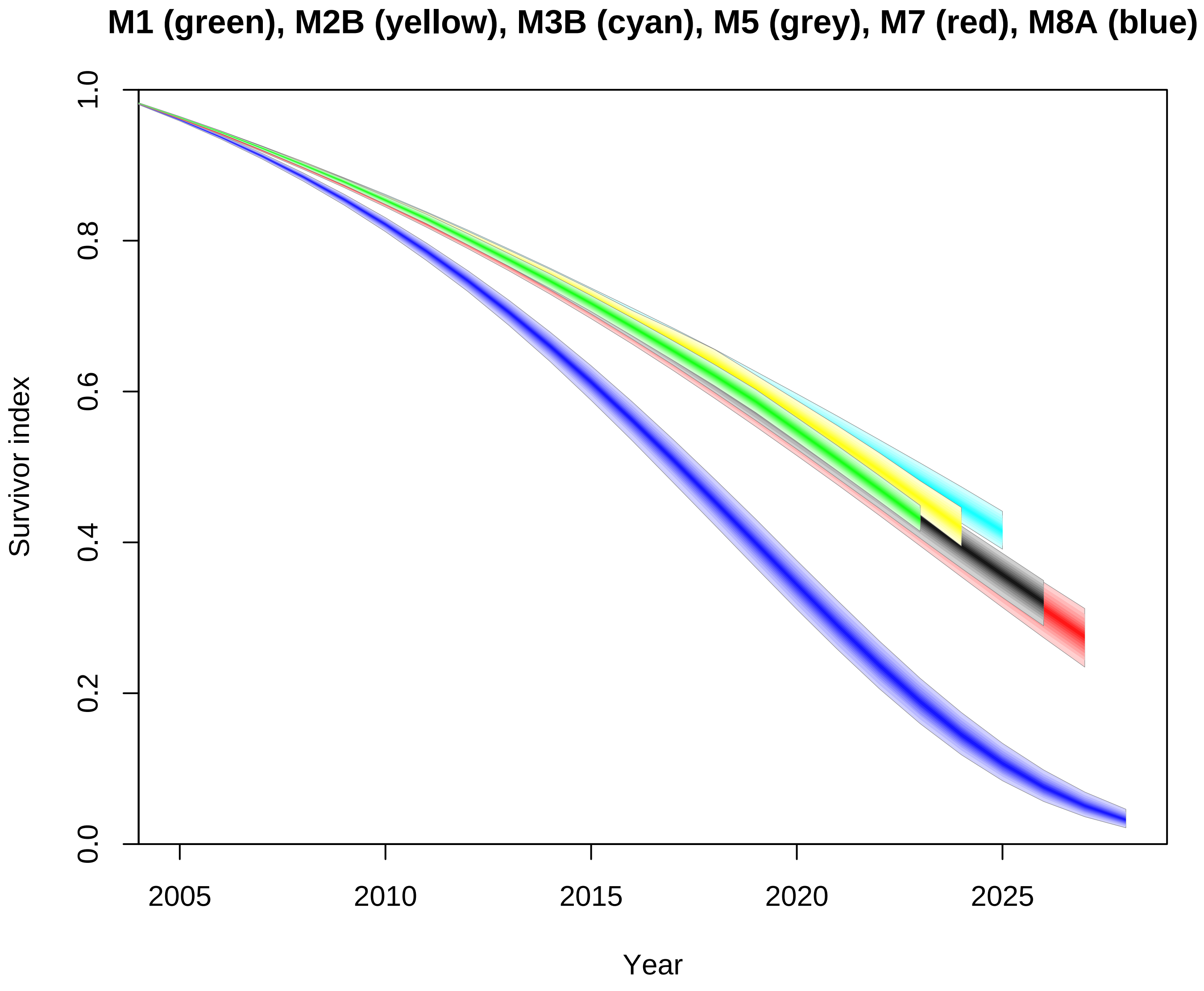


Figure 37: US, males: Fan charts for the survivor index $S(t, 65)$ for the cohort aged 65 at the start of 2005, for models M1 (green), M2B (yellow), M3B (cyan), M5 (grey), M7 (red) and M8A (blue).