A quantitative comparison of stochastic mortality models using data from England & Wales and the United States

Andrew J.G. Cairns[a][b], David Blake[c], Kevin Dowd[d],
Guy D. Coughlan[e], David Epstein[e], Alen Ong[e], and Igor Balevich[f]

March 2007

Abstract

We compare quantitatively eight stochastic models explaining improvements in mortality rates in England & Wales and in the US. On the basis of the Bayes Information Criterion (BIC), we find that an extension of the Cairns, Blake & Dowd (2006b) model that incorporates the cohort effect fits the England & Wales data best, while for US data, the Renshaw & Haberman (2006) extension to the Lee & Carter (1992) model that also allows for a cohort effect provides the best fit. However, we identify problems with the robustness of parameter estimates of these models over different time periods. A different extension to the Cairns, Blake & Dowd (2006b) model that allows not only for a cohort effect, but also for a quadratic age effect, while ranking below the other models in terms of the BIC, exhibits parameter stability across different time periods for both data sets. This model also shows, for both data sets, that there have been approximately linear improvements over time in mortality rates at all ages, but that the improvements have been greater at lower ages than at higher ages, and that there are significant cohort effects.

Keywords

Stochastic mortality; CBD-Perks models; Lee-Carter models; age effect; period effect; cohort effect; maximum likelihood; Bayes Information Criterion; robustness; parameter stability.

[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.

[f] Pension Advisory Group, JPMorgan Securities Inc., 270 Park Avenue, New York, NY 10017-2070, USA

1 Introduction

It has become increasingly clear that mortality improvements in countries where reliable data exist are driven by an underlying process that is stochastic. Since the early 1990s, a number of stochastic models have been developed to analyse these mortality improvements. These include the Lee-Carter model and its extensions (Lee and Carter (1992), Brouhns et al. (2002), Renshaw and Haberman (2003, 2006), CMI (2005, 2006)), the P-splines model (Currie et al. (2004), Currie (2006), CMI (2005, 2006)), and the Cairns et al. (2006b) model (a stochastic version of the Perks (1932) model). Wong-Fupuy and Haberman (2004) have compared a variety of deterministic and stochastic mortality models in a qualitative sense, but so far as we are aware, there have been no quantitative comparisons of these models. This study undertakes such a comparison.

We consider a range of both new and existing models. In the early part of the paper we compare these on the basis of a set of desirable, *qualitative* criteria: ease of implementation, parsimony, transparency, the ability to generate sample paths, the ability to generate forecast percentiles, allowance for parameter uncertainty, the incorporation of cohort effects (see Willets (1999, 2004) and Richards et al. (2006)), and the ability to produce a non-trivial correlation structure. The study then pays considerable attention to two important, *quantitative* criteria that can only be evaluated when the model is fitted to the data: consistency with historical data; and robustness relative to the range of data employed.

We find that no single model dominates on the basis of all these criteria. If we rank models using an objective model selection criterion based on the statistical quality of fit, then an extension of the Cairns et al. (2006b) model fits the England & Wales data best, while the Renshaw and Haberman (2006) model fits the US data best. However, if we take the robustness of parameter estimates into account then the preferred model is a different extension of the Cairns et al. (2006b) model that allows for both a cohort effect and a period effect that is quadratic in age.

1.1 Notation

We consider eight models in this paper and it is important that we use consistent and clear notation throughout.

Calendar year t is defined as running from time t to time $t+1$.

We define $m(t,x)$ to be the crude (that is, unsmoothed) death rate for age x in calendar year t. More specifically,

$$m(t,x) = \frac{\#\text{deaths during calendar year } t \text{ aged } x \text{ last birthday}}{\text{average population during calendar year } t \text{ aged } x \text{ last birthday}}.$$

The average population is usually approximated by an estimate of the population aged x last birthday in the middle of the calendar year.

A second measure of mortality is the mortality rate $q(t,x)$. This is the *probability* that an individual aged exactly x at exact time t will die between t and $t+1$.

A third measure is the *force* of mortality, $\mu(t,x)$. This is interpreted as the instantaneous death rate at exact time t for individuals aged exactly x at time t. For these individuals, for small dt, the probability of death between t and $t+dt$ is approximately $\mu(t,x) \times dt$.

1.2 Relationship between $m(t,x)$ and $q(t,x)$

The death rate, $m(t,x)$, and the mortality rate, $q(t,x)$, are typically very close to one another in value. With some assumptions, we can formalise this relationship more precisely:

- Assumption 1: For integers t and x, and for all $0 \leq s, u < 1$, $\mu(t+s, x+u) = \mu(t,x)$ – that is, the force of mortality remains constant over each year of integer age and over each calendar year.
- Assumption 2: We assume that we have a stationary population – that is, the size of the population at all ages remains constant over time.

These assumptions imply that

(a) $m(t,x) = \mu(t,x)$;

(b) $q(t,x) = 1 - \exp[-\mu(t,x)] = 1 - \exp[-m(t,x)]$.

Relationship (a) is often used in the analysis of death rate data (see, for example, Brouhns et al, 2002).

Relationship (b) is useful in the analysis of parametric models for mortality that are formulated in terms of $q(t,x)$.

Assumptions 1 and 2 do not normally hold exactly, but the resulting relationship between $m(t,x)$ and $q(t,x)$ is generally felt to provide an accurate approximation.

2 Data

We will now discuss the general characteristics of both the England & Wales and the US males data.

The primary motivation for this study is to compare various mortality models and determine which are best suited to forecasting mortality at higher ages. This reflects a concern with longevity risk — the risk that experienced survival rates might be higher than anticipated — to which pension plans and annuity providers are exposed.

As a consequence we will use data at higher ages only (ages 60 to 89 inclusive) when we make our comparisons of the different models.

2.1 England & Wales: crude death rates

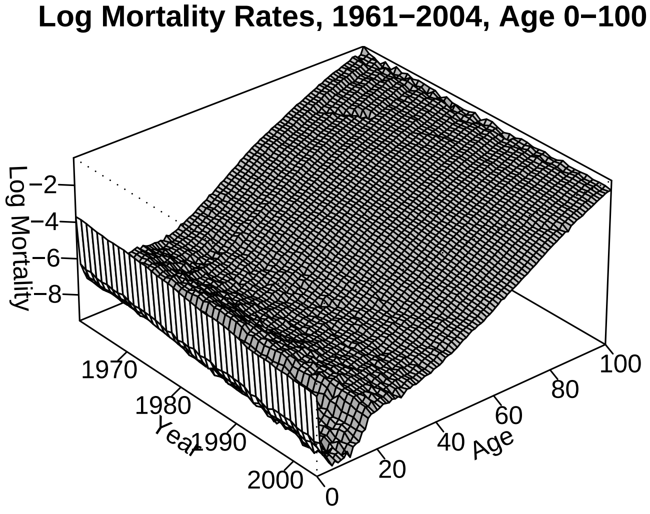
Crude mortality rates for England & Wales males between 1961 and 2004 are plotted in Figure 1. The upper plot shows rates for all ages on the same vertical (logarithmic) scale. This gives us a general impression of how rates at different ages are related. The lower plot shows how death rates at selected ages have changed over time relative to their average values over 1961 to 1965. These plots clearly show marked improvements in mortality over time, but that the rates of improvement have been different at different ages and have also been fairly erratic.

2.1.1 Data issues

A typical dataset consists of numbers of deaths, $D(t,x)$, and the corresponding exposures, $E(t,x)$, over a range of years t and ages x. Numbers of deaths are normally regarded as being reasonably accurate, although they become less reliable at very high ages. The exposure $E(t,x)$ represents the average, during calendar year t, of the number of people alive who were aged x last birthday. This quantity is normally not known with a high degree of accuracy, even in census years, and has to be estimated by the Office for National Statistics (ONS) (or its equivalent in other countries) taking account of recorded births and deaths and net immigration.

In the analysis that follows we shall exclude a number of seemingly unreliable data points (t,x):

- The 1886 cohort (that is, $t-x=1886$). Death rates became markedly out of line with neighbouring cohorts during the 1960's. This might be the result of poorly calculated exposures (that is, estimates of average population size at each age).
- Death rates at and above age 85 in the years up to including 1970. Accurate exposures were not estimated by the ONS (or its predecessors) at these ages until 1971.



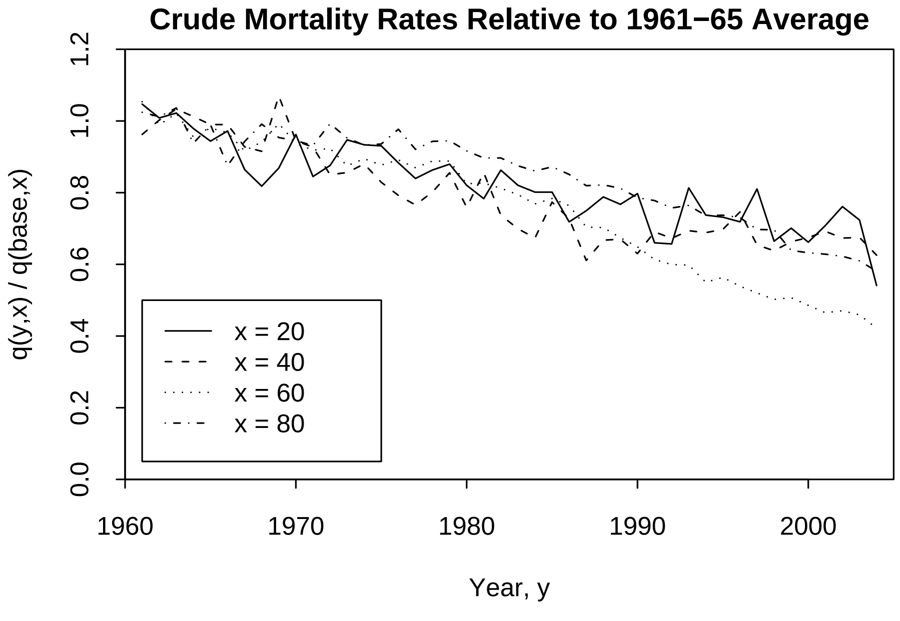


Figure 1: Crude mortality rates for England & Wales males between 1961 and 2004. Top: The surface of mortality rates $q(y, x)$ plotted on a logarithmic scale for the years 1961 to 2004. Bottom: Mortality rates for the years 1961 to 2004 at selected ages relative to average (baseline) mortality rates between 1961 and 1965.

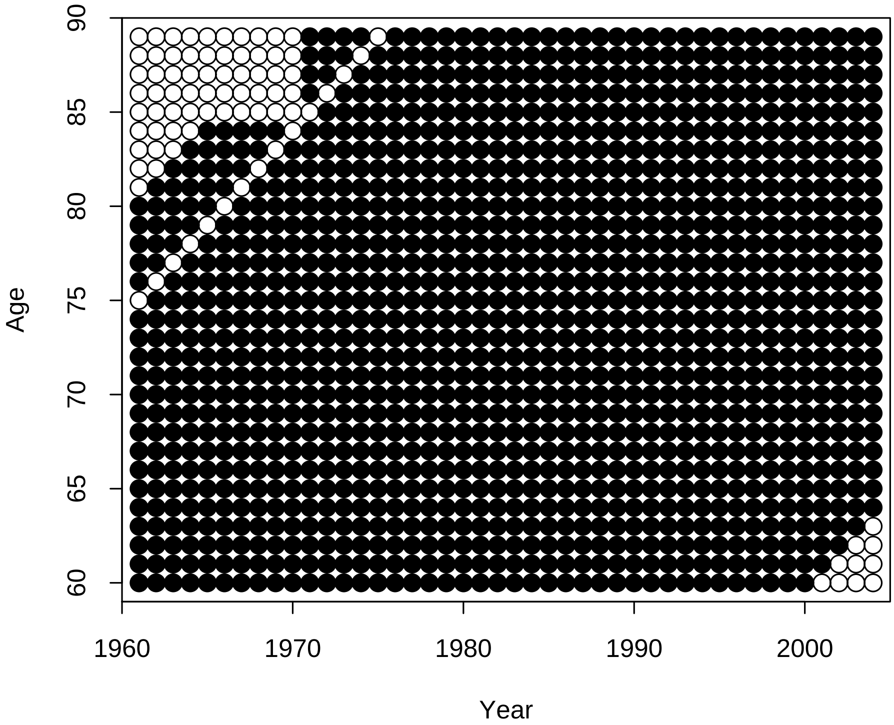


Figure 2: Data points that are included (solid dots) out of the England & Wales males mortality data ages 60-89, years 1961-2004. Excluded data are: 1886 cohort; 1961-1970 ages 85-89; and cohorts with fewer than five observations.

Additionally, since some of the models we are fitting are cohort models, we will exclude all cohorts that have fewer than 5 observations (after taking account of the exclusions above).[1]

A graphical representation of the excluded data is given in Figure 2.

[1] The reliability of the estimates of the $\gamma^{(i)}_{t-x}$ cohort parameters (defined later) depends on the number of observations for each cohort. At one extreme, if we have just one observation then the $\gamma^{(i)}_{t-x}$ parameter can be chosen so that the fitted death rate is exactly equal to the observed rate, a so-called quality of fit that can be achieved without affecting any of the other estimated death rates. In effect, the single observation allows us to overfit the model, whereas the estimated $\gamma^{(i)}_{t-x}$ parameter is, in reality, subject to substantial parameter uncertainty. With more observations in a given cohort, the estimated $\gamma^{(i)}_{t-x}$ parameter becomes more reliable. Consequently, we wish to exclude cohorts that have too few observations. However, if we exclude too many cohorts then we are left with relatively little data. We therefore adopt a compromise and exclude cohorts with fewer than 5 observations as can be seen in Figure 2.

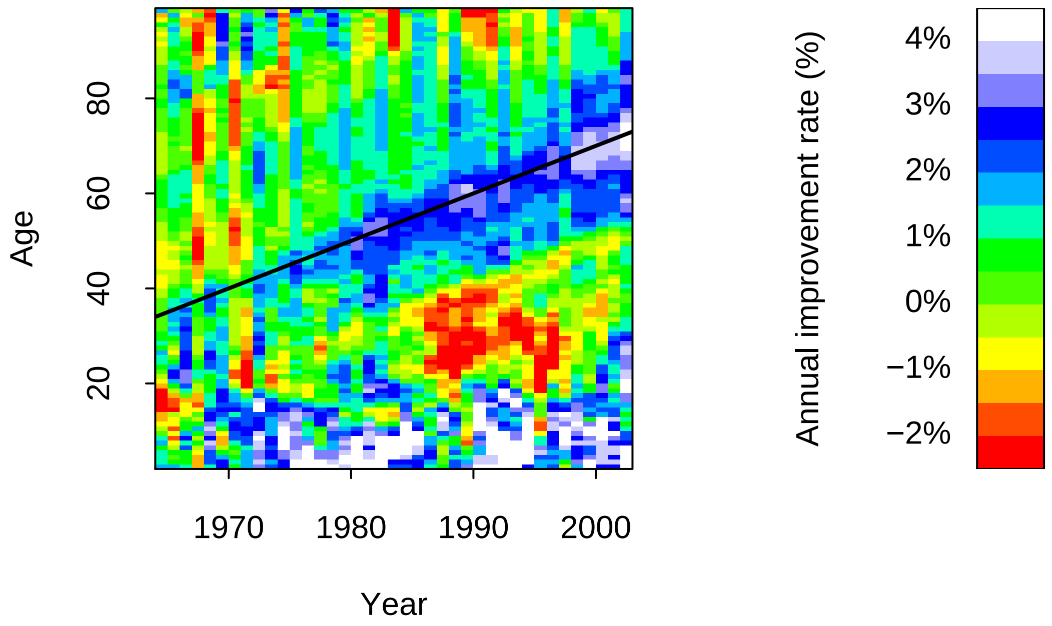


Figure 3: Improvement rates in mortality for England & Wales by calendar year and age relative to mortality rates at the same age in the previous year. Red cells imply that mortality is deteriorating; green small rates of improvement, and blue and white strong rates of improvement. The black diagonal line follows the progress of the 1930 cohort.

2.1.2 The cohort effect

Some of the models we employ incorporate what is commonly called the "cohort effect". The rationale for its incorporation lies in an analysis of the rates at which mortality has been improving at different ages and in different years. Rates of improvement are plotted in Figure 3 (see, also Willets, 2004, and Richards et al., 2006). A black and white version of this graph can be found in the Appendix, Figure 38.

In line with previous authors (see, for example, Willets, 2004, Richards et al., 2006) we can note the following points. In certain sections of the plot, we can detect strong diagonals of similar colours. Most obviously, cohorts born around 1930 have strong rates of improvement between ages 40 and 70 relative to, say, cohorts born 10 years earlier or 10 years later. The cohort born around 1950 seems to have worse mortality than the immediately preceeding cohorts.

There are other ways to illustrate the cohort effect and these can be found in Appendix A.

2.2 United States: crude death rates

This paper also analyses data for US males aged 60 to 89 over the period 1968 to 2003.[2] In our introductory remarks here we will focus on those aspects of the US data that are different from the England & Wales (EW) data.

With the EW data in a given year, t, we have identified (Cairns, Blake and Dowd, 2006b) that logit $q(t, x)$ (that is, $\log q(t, x)/(1 - q(t, x))$) is reasonably linear in x. While this is approximately true for the US data, we can see that in some years (for example, in 2003 as illustrated in Figure 4) there is a small degree of curvature in the plot of x versus logit $q(t, x)$. The curvature is not all that prominent, but it does turn out to be significant when we compare models with and without a quadratic term in logit $q(t, x)$, and it is an effect that changes over time.

We have also seen a prominent cohort effect in the EW data. Cohort effects are certainly evident in the US data as can be seen in Figure 5. However, above age 60 the magnitude of the effect is much smaller than the the England & Wales cohort effect. A black and white version of this graph can be found in the Appendix, Figure 39.

2.2.1 Data issues

Accurate exposures data are not available for the period 1968 to 1979 for ages above 84. Consequently we have used data for ages 85 to 89 only after 1979.

In contrast to the EW data, the US data do not appear to have any individual cohorts that have identifiable problems. However, we found that the exposures data[3] were, in general, less reliable as estimates of the underlying population sizes at specific ages in specific years. This is most apparent if we follow the exposures data for a specific cohort over time.[4] Exposures data for the cohorts born in 1928, 1918, 1908 and 1898 are plotted in Figure 6. We would normally expect to see a relatively smooth progression in the exposures data from one year to the next. The decrease in the exposure from one year to the next should reflect the numbers of deaths and net immigration from the cohort. If net immigration is zero this should result in a fairly-smooth, downwards progression of values in each plot. Instead we see for each of the cohorts in Figure 6 that the pattern is to some extent erratic, especially for the top left plot. This might be explained by a volatile pattern of net immigration. However, it could also be explained by errors in the underlying data (particularly the exposures data). The corresponding plots for England & Wales are much smoother.

Further evidence of the lower reliability of the exposures data is provided by the correlation between the change in the cohort exposure ($\Delta E(1967 + t, x + t)$) and

[2] Data are available for higher ages, but (in common with most countries) age at death is often misreported at these high ages, meaning that estimated death rates become unreliable.

[3] For further explanation, see Section 3.

[4] For example, if we follow exposures data for what is labelled as the "1920" cohort, then we look at the sequence $E(1920, 0)$, $E(1921, 1), \ldots,$ $E(1980, 60)$, $E(1981, 61), \ldots$.

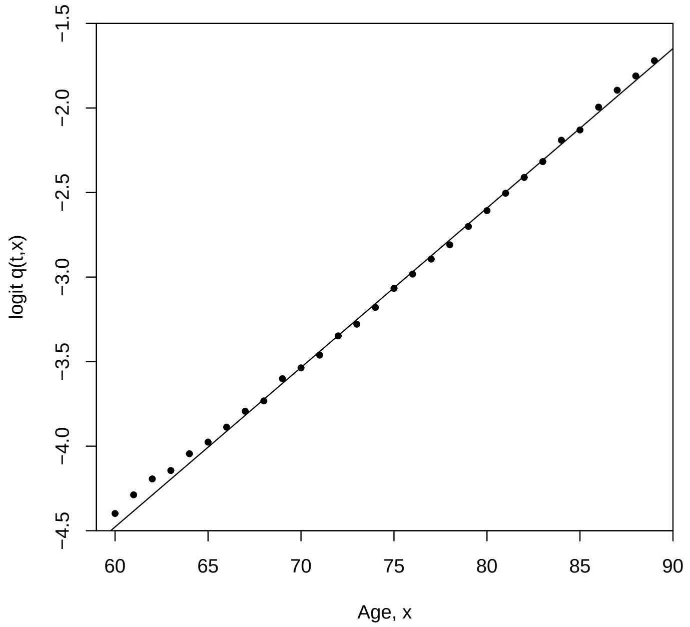


Figure 4: US mortality rates for the year 2003. Plot shows logit $q(t,x)$ for the year $t = 2003$ and ages $x = 60, \dots, 89$.

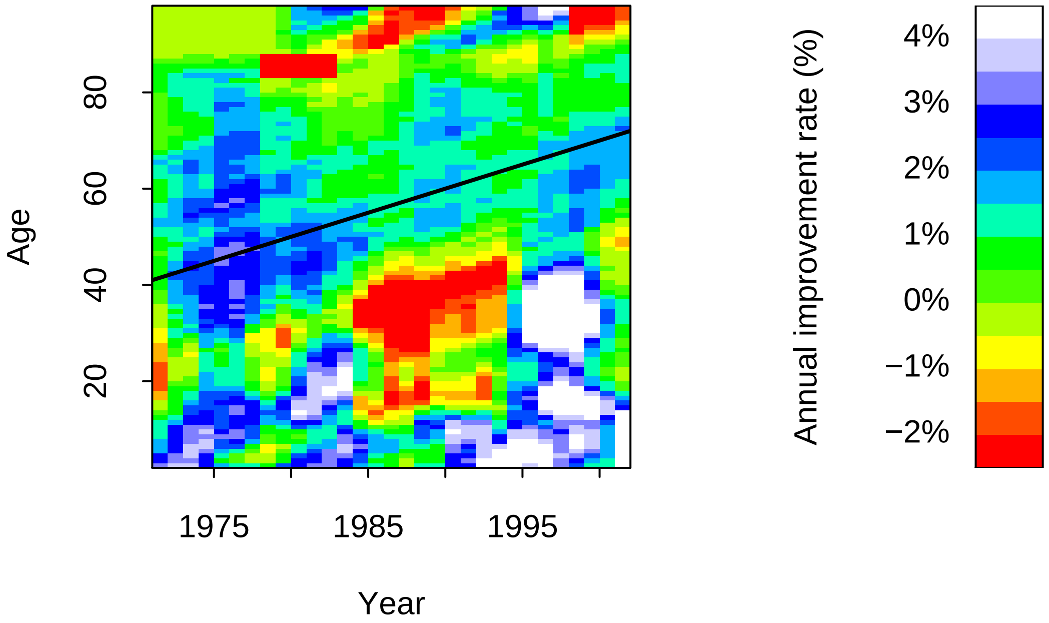


Figure 5: Improvement rates in mortality for the US by calendar year and age relative to mortality rates at the same age in the previous year. Red cells imply that mortality is deteriorating; green small rates of improvement, and blue and white strong rates of improvement.

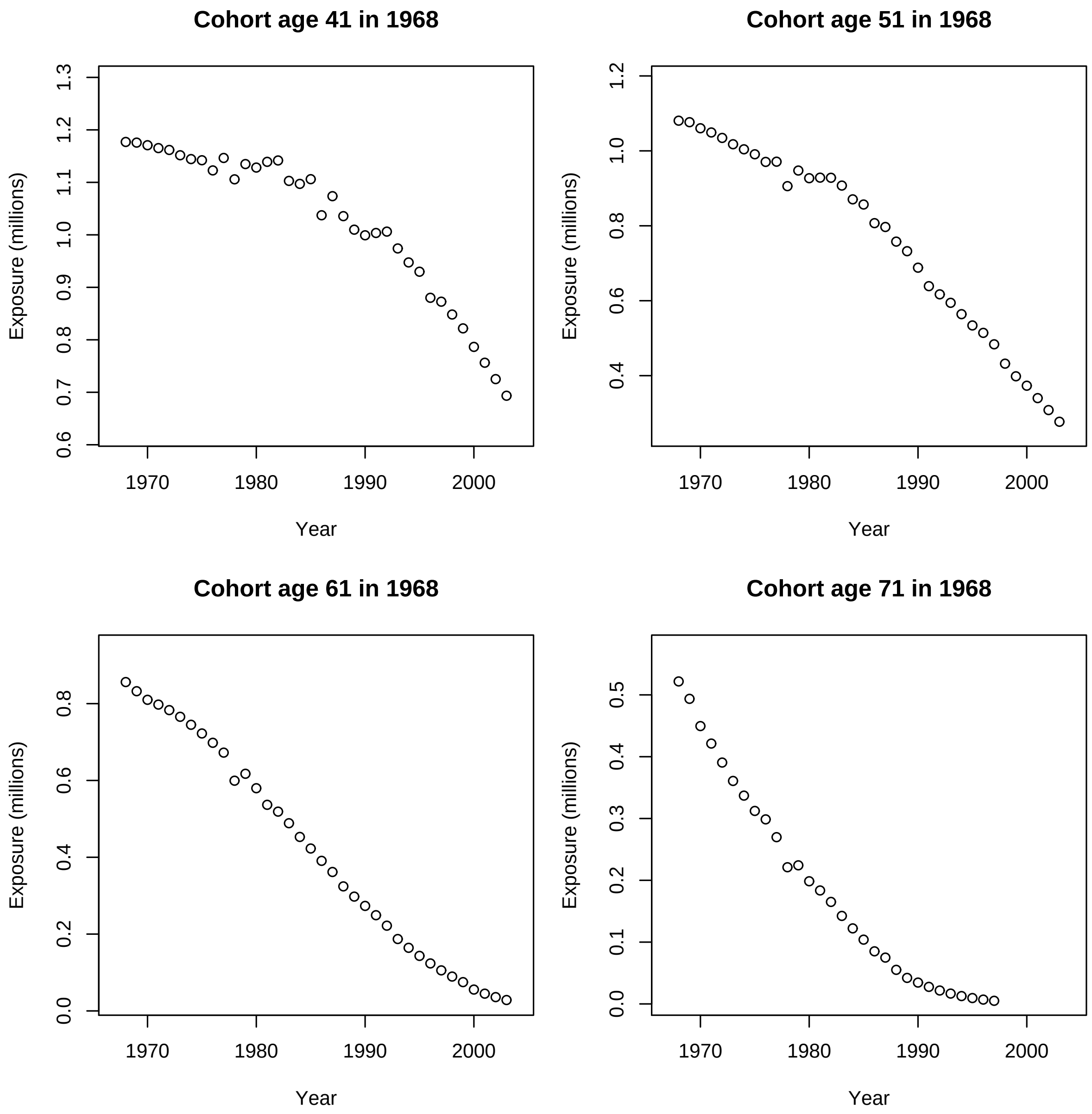


Figure 6: US Exposures for different cohorts over the period 1968 to 2003: $E(1967+t, x+t)$ for $x = 40, 50, 60, 70$.

the corresponding change in the death rate ($\Delta m(1967+t, x+t)$).[5] If the data were totally reliable then this correlation should be very close to zero. However, we find that there is a strong negative correlation (around -0.6) for the age 41 and 51 cohorts in Figure 6. The interpretation of this is that if we see an abnormally high death rate in a specific year then this can be largely explained by an abnormally low exposure rather than an abnormally high number of deaths.

In Figure 5 we see a distinct vertical band between 1984 and 1994 at the lower ages characterised by low mortality improvement rates (red), with high improvement rates after 1994 (white) which might suggest inaccuracies in the data. The red region corresponds to a period of apparently-worsening mortality rates, while the white region after 1995 corresponds to a dramatic improvement in mortality rates in 1996. This period sees the baby-boom generation moving through their 20's and 30's in the exposures data. The rise and fall in mortality rates might be linked to the HIV/AIDS epidemic, but the fall would need to be linked to the successful introduction of a treatment for the virus. This paper focuses on mortality above age 60, so we do not pursue this feature of the US data.

3 Estimation

We have data covering ages $x_1, \ldots, x_{n_a}$ and calendar years $t_1, \ldots, t_{n_y}$. For each age x and year t we have an exposure of $E(t,x)$ (that is, $E(t,x)$ is the average size of the population aged x last birthday during year t) and $D(t,x)$ deaths during year t recorded as age x last birthday at the date of death.

We will model the number of deaths using the Poisson model commonly employed in the literature on mortality modelling: namely $D(t,x)$ has a Poisson distribution with mean $E(t,x) \times m(t,x)$ (or $D(t,x) \sim Po\left(E(t,x)m(t,x)\right)$): see, for example, Brouhns et al. (2002).

Our analysis is complicated slightly by the fact that some of the models we consider directly model the death rate $m(t,x)$ while others model the mortality rate $q(t,x)$. In order to ensure that our comparison of the different models is carried out in a consistent way, our analyses of the models for $q(t,x)$ involve an additional step. First, for a given set of parameters we calculate the $q(t,x)$. We then transform these into death rates using the identity $m(t,x) = -\log[1-q(t,x)]$. We can then calculate the likelihood for all models consistently based on the $m(t,x)$ values.

For a given model, we use ϕ to represent the full set of parameters, and the notation for $m(t,x)$ is augmented to read $m(t,x;\phi)$ to indicate its dependence on the parameters. Where we have a model for $q(t,x) = q(t,x;\phi)$ we define

$$m(t,x;\phi) = -\log[1-q(t,x;\phi)].$$

For all models the log-likelihood is

$$l(\phi; D, E) = \sum_{t,x} D(t,x)\log[E(t,x)m(t,x;\phi)] - E(t,x)m(t,x;\phi) - \log[D(t,x)!],$$

[5] In these expressions Δ is the difference operator that takes the difference between the observations at t and $t-1$.

Model	formula
M1	$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}$
M2	$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta^{(3)}\gamma_{t-x}^{(3)}$
M3	$\log m(t,x) = \beta_x^{(1)} + \kappa_t^{(2)} + \gamma_{t-x}^{(3)}$
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 mortality models: 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 $(x-\bar{x})^2$.

and estimation is by maximum likelihood.[6]

4 Mortality models

The data will cover the range $x_1, \ldots, x_{n_a}$ and $t_1, \ldots, t_{n_y}$ with increments of one in each case. Models will be labelled M1, M2 etc. and are listed in Table 1.

Additionally we will use the following conventions:

- The $\beta_x^{(i)}$ functions will reflect age-related effects.
- The $\kappa_t^{(i)}$ functions will reflect period-related effects.
- The $\gamma_c^{(i)}$ functions will reflect cohort-related effects, with $c = t - x$.

All of the models we shall look at with the exception of the P-splines model will be of the form $\log m(t,x) = \sum_i \beta_x^{(i)}\kappa_t^{(i)}\gamma_{t-x}^{(i)}$ or logit $q(t,x) = \sum_i \beta_x^{(i)}\kappa_t^{(i)}\gamma_{t-x}^{(i)}$.

The method of recording the calendar year of death and the age last birthday at death means that the death count covers individuals born on 1 January in calendar

[6] Note that $D(t,x)!$ means "$D(t,x)$ factorial".

year $t-x-1$ through to 31 December $t-x$ (that is, two years) with a peak representation on 1 January $t-x$. The cohort index $c=t-x$ takes its values from the second of these years. As an example, our previously-referred-to 1886 cohort (Section 2.1.1) covers individuals born between 1 January 1885 and 31 December 1886.

4.1 Model M1

Lee and Carter (1992) propose the following model for death rates:

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

For this and some of the other models there is an *identifiability* problem in parameter estimation. To see this, note that the revised parametrisation

$$\begin{aligned} \log m(t,x) &= \tilde{\beta}_x^{(1)} + \tilde{\beta}_x^{(2)}\tilde{\kappa}_t^{(2)} \\ \text{where } \tilde{\beta}_x^{(1)} &= \beta_x^{(1)} + b\beta_x^{(2)} \\ \tilde{\beta}_x^{(2)} &= \beta_x^{(2)}/a \\ \text{and } \tilde{\kappa}_t^{(2)} &= a(\kappa_t^{(2)} - b) \end{aligned}$$

results in identical values for $\log m(t,x)$, and this means that we cannot distinguish between the two parametrisations. To circumvent this problem we need to impose two constraints on the parameters to prevent the arbitrary revaluation of the parameters a and b above. To some extent the choice of constraints is a subjective one, although some are more natural choices than others. With the current model we use the following constraints:

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

The first is a natural constraint and implies that for each x the estimate for $\beta_x^{(1)}$ will be equal (at least approximately) to the mean over t of the $\log m(t,x)$. There has to be a second constraint to pin down both of the parameters a and b above. However, there is no natural choice for this constraint and, indeed, different choices can be seen in the different applications of the Lee-Carter model in the academic literature. The important point, here, is that the choice of the second constraint has no impact on the quality of the fit, or on forecasts of mortality.

4.2 Model M2

Renshaw and Haberman (2006) generalised the Lee-Carter model to include a cohort effect as follows:

$$\log m(t,x) = \beta_x^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x}^{(3)}.$$

Model M1 is then a special case where the $\beta_x^{(3)}$ and $\gamma_{t-x}^{(3)}$ are set to zero.

This model has similar identifiability problems to the previous model. We therefore impose the following constraints to ensure identifiability:

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

The first and third constraints mean that the estimate for $\beta_x^{(1)}$ will be (at least approximately) equal to the mean over t of the $\log m(t,x)$. The second and fourth constraints are similar to the second constraint in model M1, in that they have no natural choices, but the actual choice makes no difference to the quality of fit.

The original paper of Renshaw and Haberman (2006) chose to fix their estimates for $\beta_x^{(1)}$ at $n_y^{-1}\sum_t \log m(t,x)$; the remaining parameters are estimated using an iterative scheme (an approach which we adopt here). In contrast, we use the above expression for $\beta_x^{(1)}$ only as an initial estimate and include $\beta_x^{(1)}$ in the iterative scheme as well.

We found that parameter values in the iterative scheme converge very slowly to their maximum likelihood estimates. This suggests that some sort of identifiability problem remains. It is not clear if this problem is an exact one in the sense described above or approximate. An exact identifiability problem means that the likelihood function will be absolutely level in certain dimensions about the maximum. An approximate identifiability problem means that the likelihood function will be very flat, although not absolutely level, in certain dimensions.

4.3 Model M3

Currie (2006) introduces the simpler Age-Period-Cohort (APC) model

$$\log m(t,x) = \beta_x^{(1)} + \kappa_t^{(2)} + \gamma_{t-x}^{(3)}.$$

This is a special case of model M2 with $\beta_x^{(2)} = 1$ and $\beta_x^{(3)} = 1$. Currie (2006) uses P-splines to fit $\beta_x^{(1)}$, $\kappa_t^{(2)}$ and $\gamma_{t-x}^{(3)}$ to ensure smoothness. In our analysis of M3 we do not impose any smoothness conditions.

Without loss of generality we impose the following constraints:

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

We need one further constraint, since we can otherwise add $\delta\Big((t-\bar{t}) - (x-\bar{x})\Big)$ to $\gamma_{t-x}^{(3)}$, subtract $\delta(t-\bar{t})$ from $\kappa_t^{(2)}$ and add $\delta(x-\bar{x})$ from $\beta_x^{(1)}$ with no impact on the two constraints above. We propose here that the tilting parameter, δ, be chosen within an iterative scheme to minimise

$$S(\delta) = \sum_x \left(\beta_x^{(1)} + \hat{\sigma}_x^2(x-\bar{x}) - \bar{\beta}_x^{(1)}\right)^2,$$

where $\bar{\beta}_x^{(1)} = n_y^{-1}\sum_t \log m(t,x)$. This implies that

$$\delta = -\frac{\sum_x (x-\bar{x})(\beta_x^{(1)} - \bar{\beta}_x^{(1)})}{\sum_x (x-\bar{x})^2}.$$

Given that the $\kappa_t^{(2)}$ and $\gamma_{t-x}^{(3)}$ already satisfy the first two constraints, we revise our parameter estimates according to the formulae:

$$\begin{aligned}
\tilde{\kappa}_t^{(2)} &= \kappa_t^{(2)} - \delta(t-\bar{t}) \\
\tilde{\gamma}_{t-x}^{(3)} &= \gamma_{t-x}^{(3)} + \delta\Big((t-\bar{t}) - (x-\bar{x})\Big) \\
\tilde{\beta}_x^{(1)} &= \beta_x^{(1)} + \delta(x-\bar{x}).
\end{aligned}$$

Note that models M1 to M3 can be described as belonging to the family of generalised Lee-Carter models.

4.4 Model M4

Currie et al. (2004) propose the use of B-splines and P-splines to fit the mortality surface:

$$\log m(t,x) = \sum_{i,j} \theta_{ij} B_{ij}^{ay}(x,t)$$

with smoothing of the θ_{ij} in the age and cohort directions.

Appendix B discusses the construction of B-splines and how they are fitted.

4.5 Model M5

Cairns, Blake and Dowd (2006b) (CBD) fitted the following model to mortality rates $q(t,x)$:

$$\text{logit } q(t,x) = \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(2)}\kappa_t^{(2)}.$$

For this model simple parametric forms were assumed for $\beta_x^{(1)}$ and $\beta_x^{(2)}$:

$$\begin{aligned}
\beta_x^{(1)} &= 1, \\
\text{and } \beta_x^{(2)} &= (x-\bar{x})
\end{aligned}$$

where $\bar{x} = n_a^{-1}\sum_i x_i$ is the mean age in the sample range (in our analysis, therefore, $\bar{x} = 74.5$). Thus

$$\text{logit } q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}).$$

This model has no identifiability problems.

4.6 Model M6

This model is the first generalisation of the CBD model to include a cohort effect.

$$\text{logit } q(t,x) = \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(2)}\kappa_t^{(2)} + \beta_x^{(3)}\gamma_{t-x}^{(3)}$$

For this model, simple parametric forms were assumed for $\beta_x^{(1)}$, $\beta_x^{(2)}$ and $\beta_x^{(3)}$:

$$\begin{aligned} \beta_x^{(1)} &= 1, \\ \beta_x^{(2)} &= (x-\bar{x}), \\ \text{and } \beta_x^{(3)} &= 1. \end{aligned}$$

Thus

$$\text{logit } q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \gamma_{t-x}^{(3)}.$$

As with other models we have an identifiability problem. Here we can switch from $\gamma_{t-x}^{(3)}$ to $\tilde{\gamma}_{t-x}^{(3)} = \gamma_{t-x}^{(3)} + \phi_1 + \phi_2(t-x-\bar{x})$ and, with corresponding adjustments to $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$, there is no impact on the fitted values of the $q(t,x)$. This requires two constraints to prevent arbitrary use of ϕ_1 and ϕ_2. The constraint we have used here (which acts to constrain both ϕ_1 and ϕ_2) is that if we use least squares to fit a linear function of $t-x$ to $\gamma_{t-x}^{(3)}$ then the fitted linear function is identically equal to zero. In doing so our estimates of the $\gamma_{t-x}^{(3)}$ will be centred around zero and there will be no constant trend up or down.

4.7 Model M7

This model is a generalisation of model M6 that adds a quadratic term into the age effect. The inclusion of the quadratic term is inspired by the possibility of some curvature identified in the logit $q(t,x)$ plots in the US data. Thus

$$\text{logit } q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x-\bar{x}) + \kappa_t^{(3)}\left((x-\bar{x})^2 - \hat{\sigma}_x^2\right) + \gamma_{t-x}^{(4)}.$$

Here the constant $\hat{\sigma}_x^2 = n_a^{-1}\sum_i (x-\bar{x})^2$ is the mean of $(x-\bar{x})^2$.

As with model M6, we have an identifiability problem and we can switch from $\gamma_{t-x}^{(4)}$ to $\tilde{\gamma}_{t-x}^{(4)} = \gamma_{t-x}^{(4)} + \phi_1 + \phi_2(t-x-\bar{x}) + \phi_3(t-x-\bar{x})^2$ and with corresponding adjustments to $\kappa(1)_t$, $\kappa(2)_t$ and $\kappa(3)_t$, without there being an impact on the fitted values of the $q(t,x)$. This requires three constraints to prevent arbitrary choices over ϕ_1, ϕ_2 and ϕ_3. We impose the constraint that the fitted quadratic function from the least squares fit of a quadratic function of $(t-x)$ to $\gamma_{t-x}^{(4)}$ is identically equal to zero. This constraint fixes ϕ_1, ϕ_2 and ϕ_3. In doing so our estimates of the $\gamma_{t-x}^{(4)}$ will be fluctuating around zero, there will be no obvious trend up or down and there will be no systematic curvature.

4.8 Model M8

Our third generalisation of the CBD model builds on our experience from fitting model M2 (see the results in Section 6). This suggested that the impact of the cohort effect $\gamma^{(3)}_{t-x}$ for any specific cohort diminishes over time ($\beta^{(3)}_x$ = decreasing with x) instead of remaining constant ($\beta^{(3)}_x$ = constant). This gives us

$$\text{logit } q(t,x) = \beta^{(1)}_x \kappa^{(1)}_t + \beta^{(2)}_x \kappa^{(2)}_t + \beta^{(3)}_x \gamma^{(3)}_{t-x}$$

where

$$\begin{aligned} \beta^{(1)}_x &= 1, \\ \beta^{(2)}_x &= (x - \bar{x}), \\ \text{and } \beta^{(3)}_x &= (x_c - x) \end{aligned}$$

for some constant parameter x_c to be estimated. This results in

$$\text{logit } q(t,x) = \kappa^{(1)}_t + \kappa^{(2)}_t (x - \bar{x}) + \gamma^{(3)}_{t-x}(x_c - x).$$

To avoid identifiability problems we need to introduce one constraint:

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

Each of models M6 to M8 is an extension of model M5 with some allowance for the cohort effect. Consequently models M5 to M8 can be described as members of the family of generalised CBD-Perks models.

4.9 Philosophical differences between models

We comment here briefly on the underlying structural assumptions of different models. It is these assumptions and their general philosophy that guide us in the range of models that we consider.

All of models M1-M3 and M5-M8 share the same underlying assumption that the age, period and cohort effects are qualitatively different in nature. Specifically, there is randomness from one year to the next, perhaps caused by local environmental factors, which we do not observe between ages. In contrast, the P-splines model, M4, assumes that there is smoothness in the underlying mortality surface in the period effects as well as in the age and cohort effects. Models M5 to M8 differ from M1 to M3 in that they assume a functional relationship (and hence smoothness) between ages.

Depending on one's personal beliefs about the underlying randomness in the age, period and cohort effects, one might attach greater weight to models that are aligned with these beliefs. For example, if one believes that there should be an underlying smoothness between ages, that there is randomness between cohorts, and randomness from one year to the next, then greater weight might be placed on models M6 to M8.

5 Alternative ways of evaluating and comparing mortality models

We can evaluate and compare these eight models in a number of ways:

- We can compare their main properties against a set of desirable criteria.
- We can carry out a preliminary analysis to see if fitted models look to be 'reasonable'. We can evaluate models on a stand-alone basis by means of their goodness of fit. We might also derive their testable predictions and test these predictions.
- We can rank models using ranking criteria.
- Where models are special cases of others, we can test whether the special case restrictions are empirically acceptable. If they are accepted, then the principle of parsimony would lead us to prefer the special case model; and if the restrictions are not accepted, we would prefer the more general model.
- Parameter estimates should be robust relative to the range of data employed. For example, if we use England & Wales data for 1981 to 2004 we would hope to see similar parameter estimates to those found using data from 1961 to 2004.

We now proceed to apply each of these criteria in the sections that follow.

5.1 Comparing the main features of the different models

Table 2 lists some criteria that we might consider desirable in a mortality model. These criteria are:

- Ease of implementation: other things being equal, we would prefer a model that is easier to implement than one that is not.
- Parsimony: other things being equal, a model with fewer parameters is preferable to a model with a large number of parameters.
- Transparency: how much of the model and its output is treated as a "black box"?
- Whether the model has the ability to generate sample paths: this can be useful for other tasks such as pricing longevity-linked financial instruments.[7]

[7] We refer here to sample paths for the *underlying* (and unobserveable) death rates $\hat{m}(t,x)$. A different type of sample path can be constructed when we look at *crude* (observable) death rates under the Poisson model: $m(t,x) = D(t,x)/E(t,x)$ where $D(t,x) \sim Po(\hat{m}(t,x)E(t,x))$.

Model	M1	M2	M3	M4	M5	M6	M7	M8
Ease of implementation	Y	?	Y	?	Y	Y	Y	?
Parsimony	Y	?	?	Y	Y	?	?	?
Transparency	Y	Y	Y	?	Y	Y	Y	Y
Ability to generate sample paths	Y	Y	Y	N	Y	Y	Y	Y
Ability to generate percentiles	Y	Y	Y	Y	Y	Y	Y	Y
Allowance for parameter uncertainty	Y	Y	Y	Y	Y	Y	Y	Y
Incorporation of cohort effects	N	Y	Y	Y	N	Y	Y	Y
Non-trivial correlation structure	N	N?	N?	N	Y	Y	Y	Y

Table 2: The table shows whether each model satisfies each of the stated criteria. Where a criterion cannot be answered with a simple Y(es) or N(o) the ? indicates that the model lies somewhere in the middle, with further comments in the main text.

- Whether the model has the ability to generate forecast percentiles.[8]
- Allowance for parameter uncertainty.
- Incorporation of cohort effects.
- Ability to produce a non-trivial correlation structure: that is, correlation between the year-on-year changes in mortality rates at different ages.[9]

The desirability of models that satisfy the first six criteria considered in Table 2 is obvious, and we might desire the seventh criterion if we believed that cohort effects are important and needed to be allowed for (as in fact our later results suggest). The eighth criterion is desirable to the extent that such a correlation structure appears in historical data. However, it is also relevant in developing hedging strategies based on traded mortality-linked securities.

Two important, additional criteria that we can only evaluate when we fit the data are:

- The model should be consistent with historical data.
- Parameter estimates should be robust relative to the range of data employed.

[8] Percentiles of future mortality rates etc. combine information about parameter uncertainty resulting from the fitting process with uncertainty arising from the underlying stochastic dynamics.

[9] Statistical analysis of mortality rates points to changes in the $m(t,x)$ at different ages being imperfectly correlated. This is inconsistent with the dynamics of a one-factor model such as M1 under which changes in mortality rates at different ages are perfectly correlated. The existence of a non-trivial correlation structure implies, for example, that hedging of longevity-linked liabilities requires more than one hedging instrument.

5.2 Further remarks

- Ease of implementation: all of the models require some programming. However, some are straightforward to program, while others require a higher level of skill. Models M2 and M8 have ?'s in Table 2 to indicate that they take longer to converge to the maximum likelihood estimate than the other models. Model M4 is tougher to program, but the parameter estimation algorithm converges rapidly.

- Parsimony: Each of the models has a large number of parameters (see Table 3), so none could be described as parsimonious in any absolute sense. However, some models are more parsimonious than others and have fewer parameters. M1, M4 and M5 are recorded as parsimonious, therefore, in the sense that they have fewer (effective) parameters to estimate.

- All of the models except M4 were regarded as transparent, in that their outputs are straightforward to analyse, allowing us to explain, from a statistical perspective, changes in mortality over time. M4 seems less transparent: its output is a smooth surface fitted to historical data and then projected; there is little in the output to allow us to get a feel for the underlying dynamics. This does not mean that M4 is not a useful model *per se*, just that it needs to be used with more caution.

- Ability to generate sample paths: Only the P-splines model M4 fails on this point. M4 assumes that there is an underlying smoothness to the mortality surface and that the only uncertainty in forecasts (which can be substantial) is due to parameter and model uncertainty. In other words, if the true surface were known then we would have perfect knowledge of future underlying rates of mortality.[10]

- Allowance for parameter uncertainty: It has been demonstrated by Cairns et al. (2006b) and in CMI working paper 15 (2005) that parameter uncertainty forms a significant element of the uncertainty in forecasts of future mortality. Any model that does not allow for parameter uncertainty is, therefore, in danger of significantly underestimating uncertainty in its forecasts. Not all published works on models M1 to M8 discuss parameter uncertainty, but it can be included if desired using methods described in Cairns et al. (2006b) and CMI working paper 15 (2005).

- The correlation structure is described as trivial when there is perfect correlation between changes in mortality rates at different ages from one year to the next. This is the case for model M1 for example where there is a single time series process $\kappa_t^{(2)}$. For models M2 and M3 we also have perfect correlation (for the same reason) at all ages except at the youngest age where there is potentially additional randomness arising from the arrival of a new cohort with an unknown cohort effect. Models M5, M6 and M8 have two time series

[10] If we refer back to footnote 7 we can construct a different type of sample path under M4 by using the Poission model to simulate actual numbers of deaths $D(t, x)$ in the future.

processes (and M7 three) $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ as drivers which affect different ages in different ways. Consequently changes in underlying mortality rates at different ages are not perfectly correlated.

Model	Maximum log-likelihood	Effective number of parameters	BIC (rank)
M1	-8912.7	102	-9275.8 (6)
M2	-7735.6	203	-8458.1 (3)
M3	-8608.1	144	-9120.6 (5)
M4	-9245.9	74.2	-9372.9 (7)
M5	-10035.5	88	-10348.8 (8)
M6	-7922.3	159	-8488.3 (4)
M7	-7702.1	202	-8421.1 (2)
M8	-7823.7	161	**-8396.8 (1)**

Table 3: England & Wales males ages 60 to 89 and years 1961 to 2004. Maximum likelihood, effective number of parameters estimated and Bayes Information Criterion (BIC) for each model. Effective number of parameters takes account of the constraints on parameters or the effect of the penalty functions in the case of model M4. For M4 with $dx = 4$ and $dt = 4$ inter-knot distances, the BIC is optimised over the penalty weights. The optimal value quoted here is for $\lambda_a = 786$, and $\lambda_c = 2.8$.

6 England & Wales data analysis

6.1 Estimation and preliminary data analysis

We now proceed to the more objective model comparisons.

For each model, we estimated (as appropriate) the $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_{t-x}^{(i)}$ for each factor, i, age, x, year, t, and cohort $t-x$ by maximising the log-likelihood function. Estimates of the $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_{t-x}^{(i)}$ are plotted in Figures 7 to 13.

Values for the maximum likelihood, effective number of parameters (or degrees of freedom in estimation), and the Bayes Information Criterion (BIC) for each model are given in Table 3.

6.2 Model selection criteria

If one simply compares the maximum likelihoods attained by each model then it is natural for models with more parameters to fit the data "better". Indeed such improvements are guaranteed if models are nested: if one model is a special case of another then the model with more parameters will have a higher maximum likelihood with certainty even if the true model is the model with fewer parameters.

To avoid this problem we penalise models that are over-parametrised. Specifically for each parameter that we add into the model we need to see a "significant" improvement in the maximum likelihood rather than just an increase of any size. A number of such penalties have been proposed. Here we will focus on use the Bayes Information Criterion or BIC (see, for example, Cairns, 2000, or Hayashi, 2000).

A key point about the use of the BIC is that it provides us with a mechanism for striking a balance between quality of fit (which can be improved by adding in more parameters) and parsimony. A second, and equally important point about the BIC, is that it allows us to compare models that are not necessarily nested. For example, M1 and M3 are nested within M2, but M1 is not nested within M3 and vice versa.

A final point is that the BIC makes no assumptions about "prior" model rankings: that is, all models have equal status in terms of how we rank them. In contrast, hypothesis tests start from a null hypothesis which favours one specific model over the others.

The BIC for model r is defined as

$$BIC_r = l(\hat{\phi}_r) - \frac{1}{2}\nu_r \log N$$

where ϕ_r is the parameter vector for model r, $\hat{\phi}_r$ is its maximum likelihood estimate, $l(\hat{\phi}_r)$ is the maximum log likelihood, N is the number of observations (not counting those cells that have been excluded from the analysis), and ν_r is the effective number of parameters being estimated.[11]

The models can then be ranked, with the top model having the highest BIC.

Values for the BIC are given in Table 3 and we see that model M8 comes out top in the BIC rankings with M7 second.

6.3 Parameter estimates

In Figures 7 to 13 we have plotted the maximum-likelihood estimates for the various parameters in all models except for M4.

For those models that incorporate a cohort factor we can see a distinctive cohort effect. In model M2, for example, we can see from the second kink in $\gamma_{t-x}^{(3)}$ that cohort mortality was falling at a faster rate for males born after 1920. The same feature can be seen in M3 and M6.

For models M7 and M8 the cohort effect follows a different pattern. In M7 part of the cohort effect has been substituted by the additional quadratic age effect. The M8 cohort effect seems to follow a similar pattern except for the fact that it has been tilted a little.

[11] For example, model M1 requires estimates for 30 values of $\beta_x^{(1)}$, 30 values of $\beta_x^{(2)}$ and 44 values of $\kappa_t^{(2)}$, totalling 104, but we then deduct 2 from this total to reflect the two constraints $\sum_t \kappa_t^{(2)} = 0$ and $\sum_x \beta_x^{(2)} = 1$. For the P-splines model the concept of the effective number of parameters is more abstract, and the reader is referred to Currie et al. (2004) and references therein for further details.

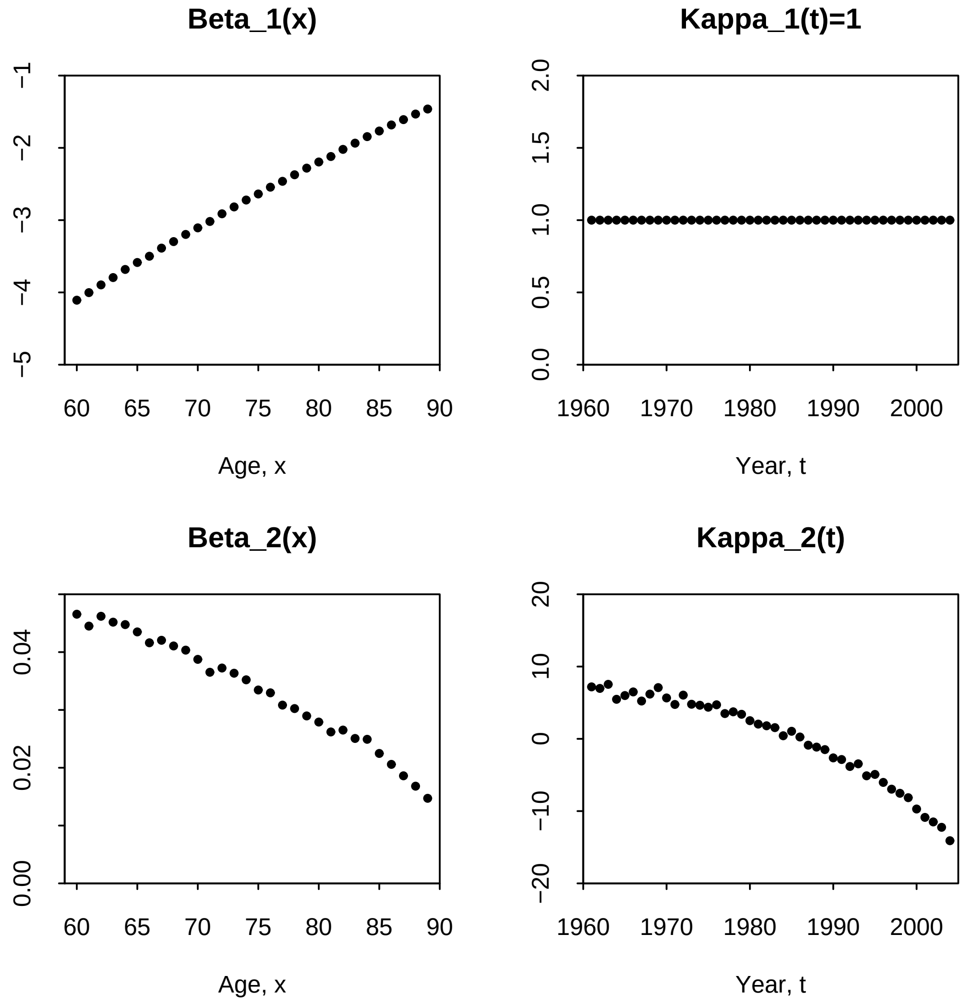


Figure 7: England & Wales data: Parameter estimates for model M1. $\kappa_t^{(1)} = 1$ is included for completeness.

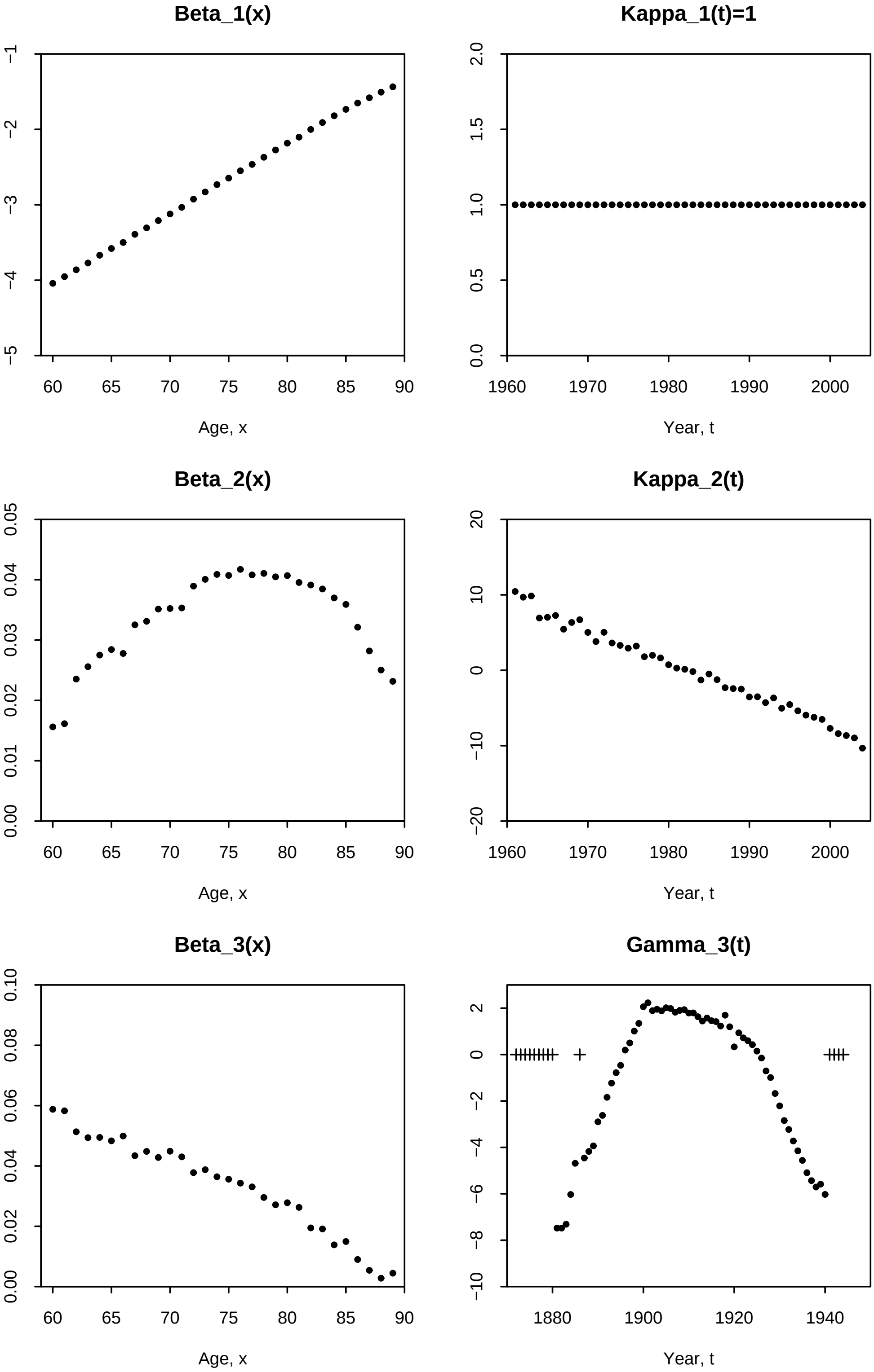


Figure 8: England & Wales data: Parameter estimates for model M2. Crosses in the bottom right plot correspond to excluded cohorts. $\kappa_t^{(1)} = 1$ is included for completeness.

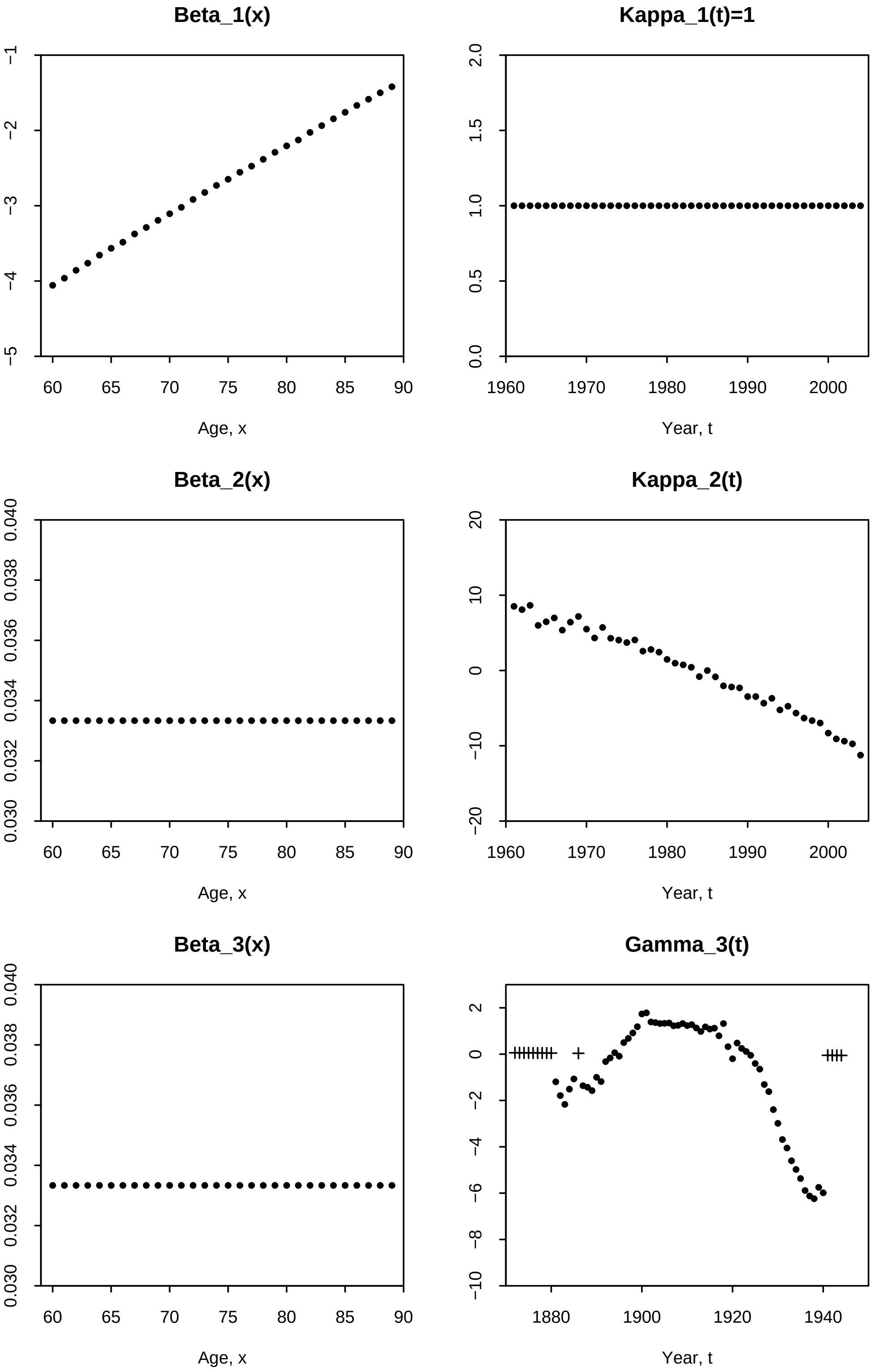


Figure 9: England & Wales data: Parameter estimates for model M3. Crosses in the bottom right plot correspond to excluded cohorts. $\kappa_t^{(1)} = 1$, $\beta_x^{(2)} = 1/30$ and $\beta_x^{(3)} = 1/30$ are included for completeness.

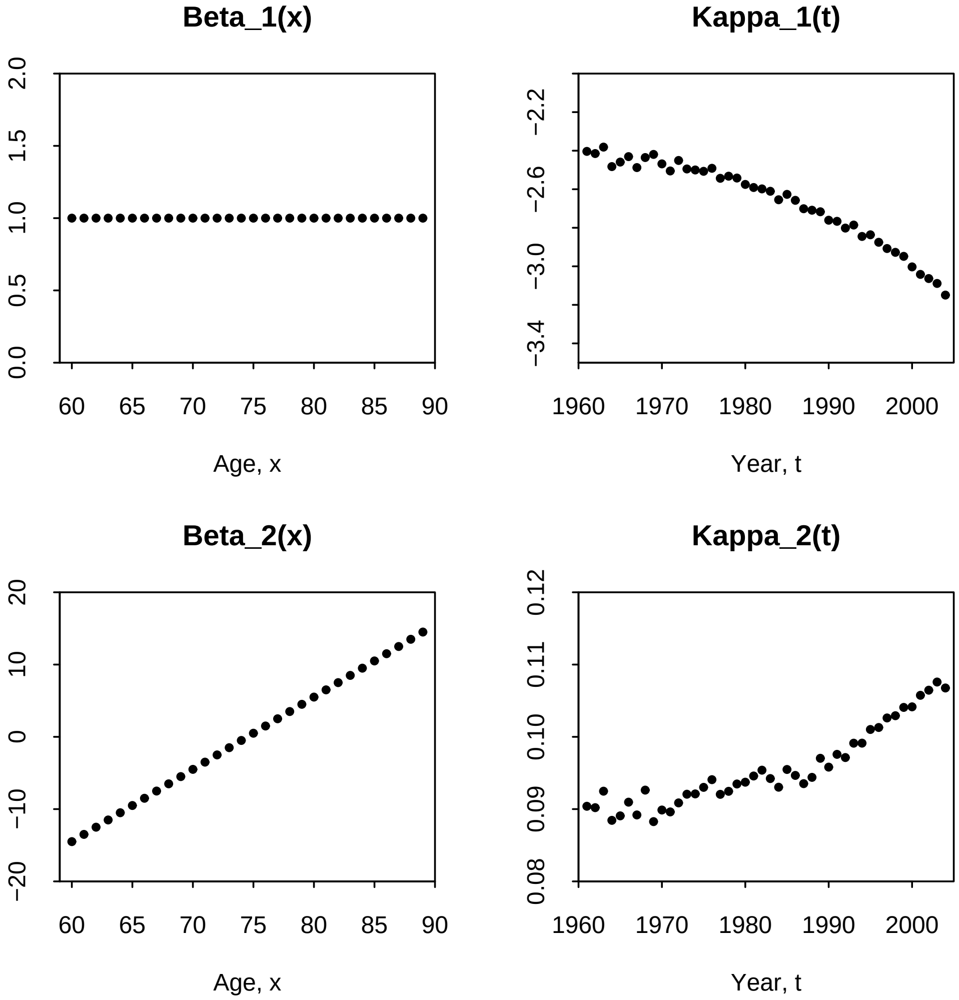


Figure 10: England & Wales data: Parameter estimates for model M5. $\beta_x^{(1)} = 1$ and $\beta_x^{(2)} = x - \bar{x}$ are included for completeness.

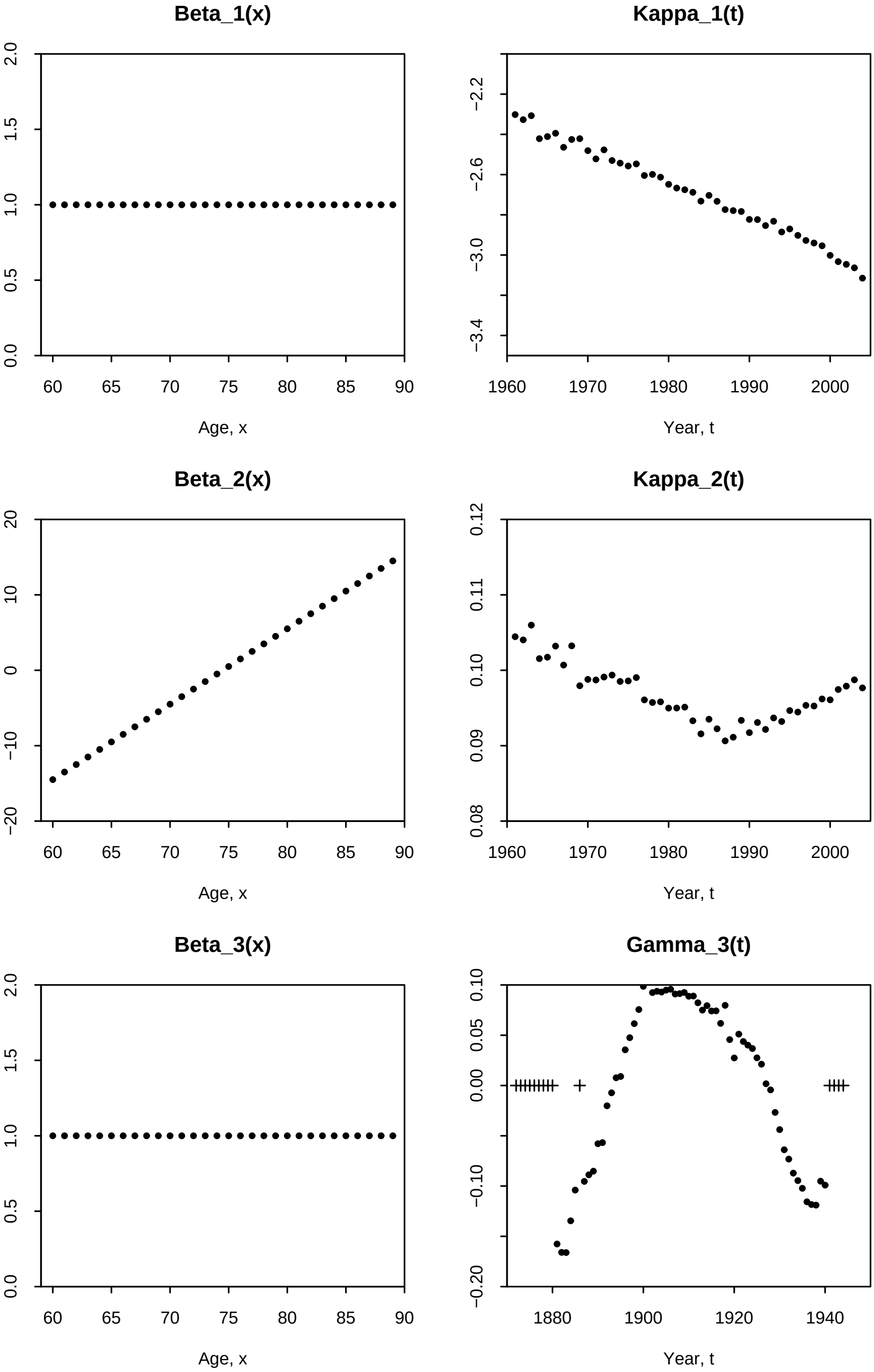


Figure 11: England & Wales data: Parameter estimates for model M6. Crosses in the bottom right plot correspond to excluded cohorts. $\beta_x^{(1)} = 1$, $\beta_x^{(2)} = x - \bar{x}$ and $\beta_x^{(3)} = 1$ are included for completeness.

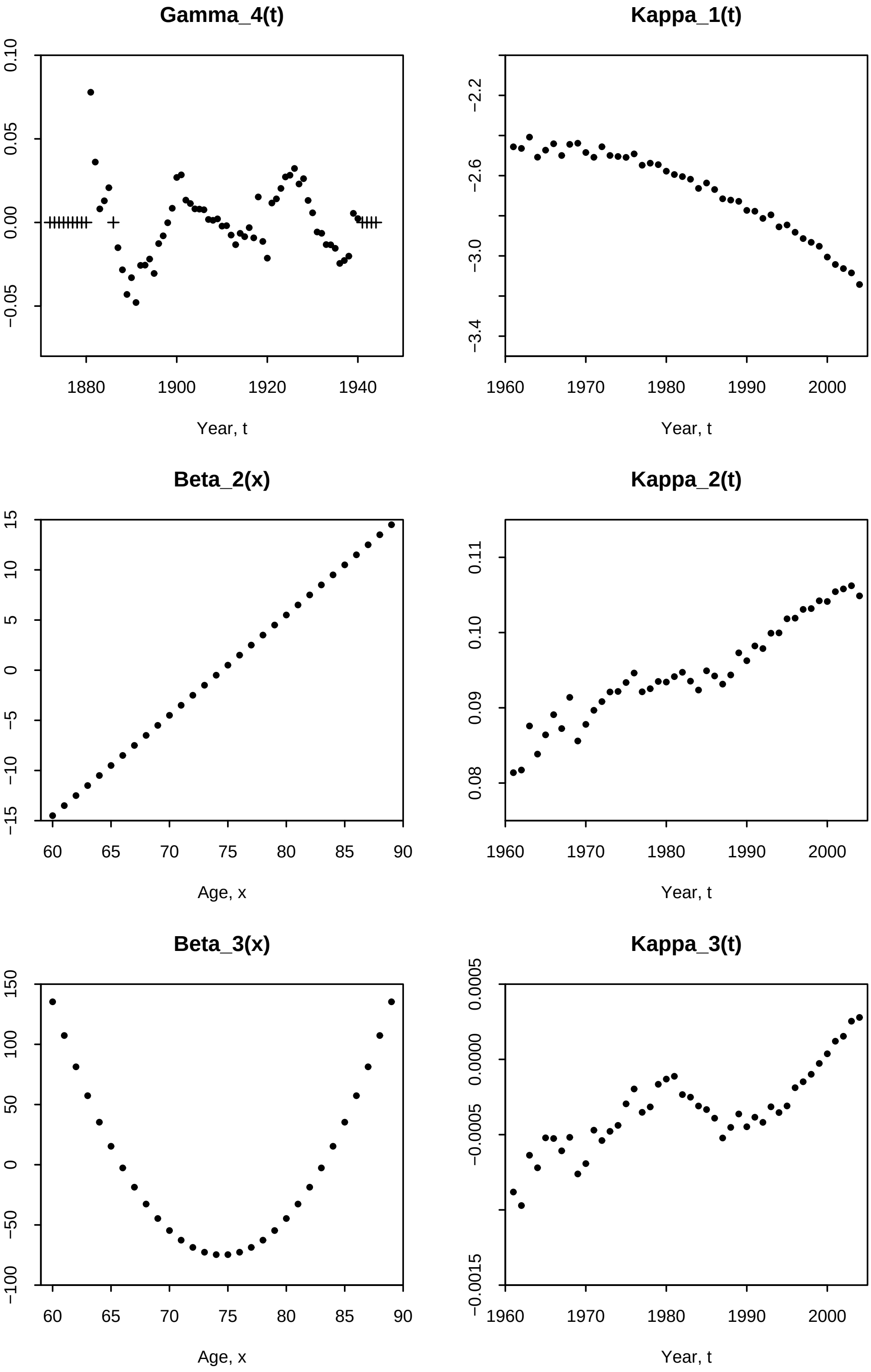


Figure 12: England & Wales data: Parameter estimates for model M7. Crosses in the bottom right plot correspond to excluded cohorts. $\beta_x^{(2)} = x - \bar{x}$ and $\beta_x^{(3)} = (x - \bar{x})^2 - \hat{\sigma}_x^2$ are included for completeness.

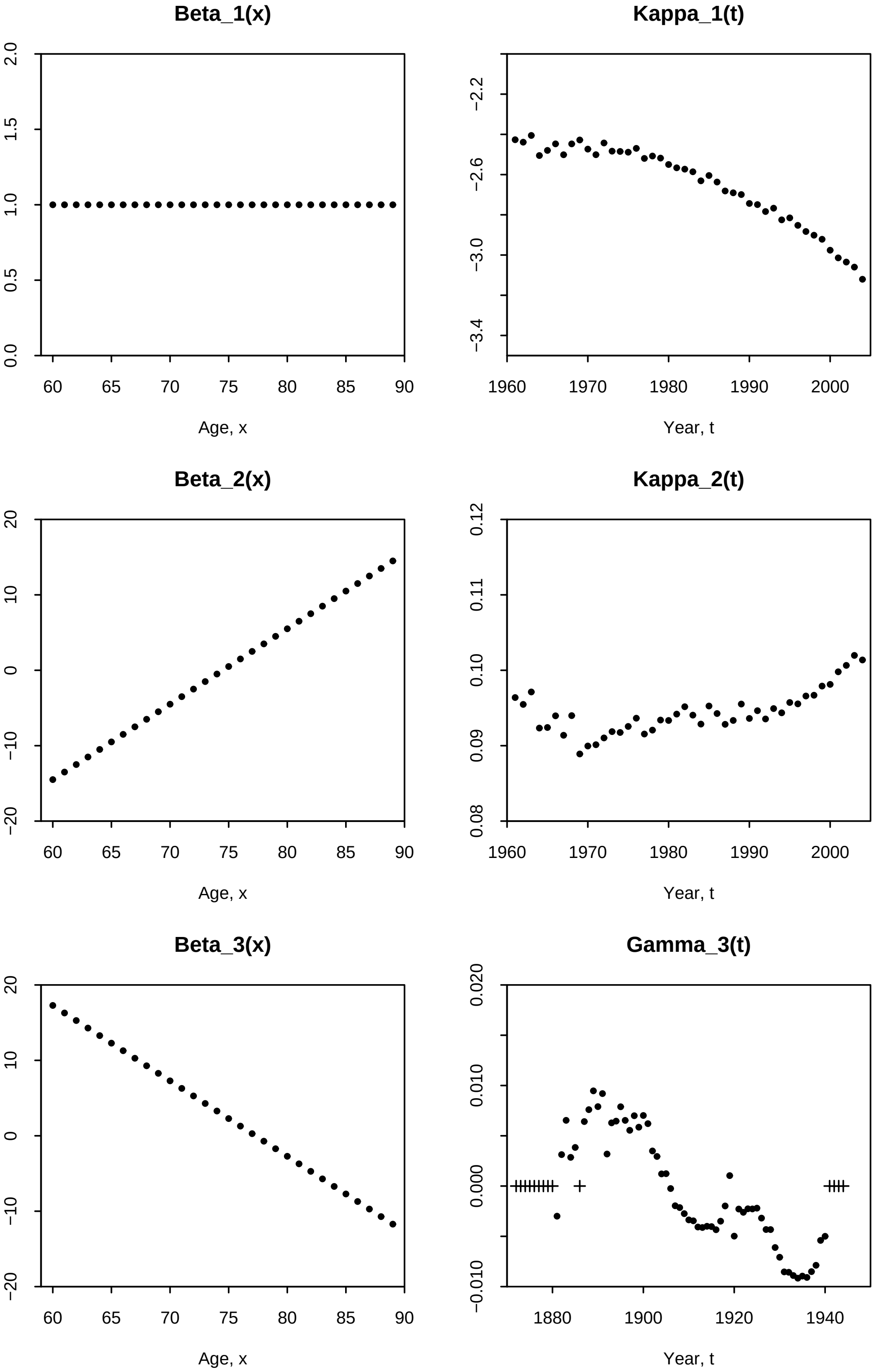


Figure 13: England & Wales data: Parameter estimates for model M8. Crosses in the bottom right plot correspond to excluded cohorts. $\beta_x^{(1)} = 1$ and $\beta_x^{(2)} = x - \bar{x}$ are included for completeness.

6.4 Robustness of parameter estimates

We can consider how robust parameter estimates are relative to changes in the period of data that we use to fit a given model.

We focus here on the highest BIC-ranked models M2, M7 and M8. For each model we have plotted (Figures 14 to 16) parameter estimates based on data from 1961 to 2004 (represented by dots) or from 1981 to 2004 (solid line). The plots reveal that out of the three models, M7 seems to be the most robust relative to changes in the period of data used: that is, the parameter estimates hardly change even if we use much less data. M8 looks reasonably robust, and the differences that we do see are, in fact, consequences of the constraint that $\sum_{t,x} \gamma_{t-x}^{(3)} = 0$ when we are summing over different years. M2 on the other hand seems clearly to produce results that lack robustness, since the parameter estimates change very significantly when we use less data. This means that we must question the reliability of projections made using M2.

6.5 Other methods of comparison

6.5.1 Standardised residuals

We can also analyse standardised residuals:

$$Z(t,x) = \frac{D(t,x) - E(t,x)\hat{m}(t,x;\hat{\phi})}{\sqrt{E(t,x)\hat{m}(t,x;\hat{\phi})}}. \tag{1}$$

Our construction of the likelihood assumes that these are i.i.d. and approximately $N(0,1)$.

Models that have higher likelihood have a lower variance of the standardised residuals. However, for each model the variance of the standardised residuals is significantly greater than 1 (for Models M2 and M8 it is around 1.6, taking account of the number of parameters estimated). Where this happens, the data are often referred to as having an overdispersed Poisson distribution. This overdispersion seems to be a general feature of mortality data in many countries. A possible source of this overdispersion lies in the fact that the exposures data are estimated.

What does this mean for forecasting? We conjecture that this overdispersion does not have a significant impact on our estimates of the future dynamics of the underlying mortality rates $q(t,x)$. However, a Poisson model might underestimate the future variability of the actual death rates relative to the underlying rates.

6.5.2 Pattern of standardised residuals

Embedded within our modelling hypothesis is an assumption that the death counts are independent for each age and year. If we add the Poisson assumption to this and if our hypothesis is also true, the standardised residuals (equation 1) will be approximately i.i.d. standard normal random variables.

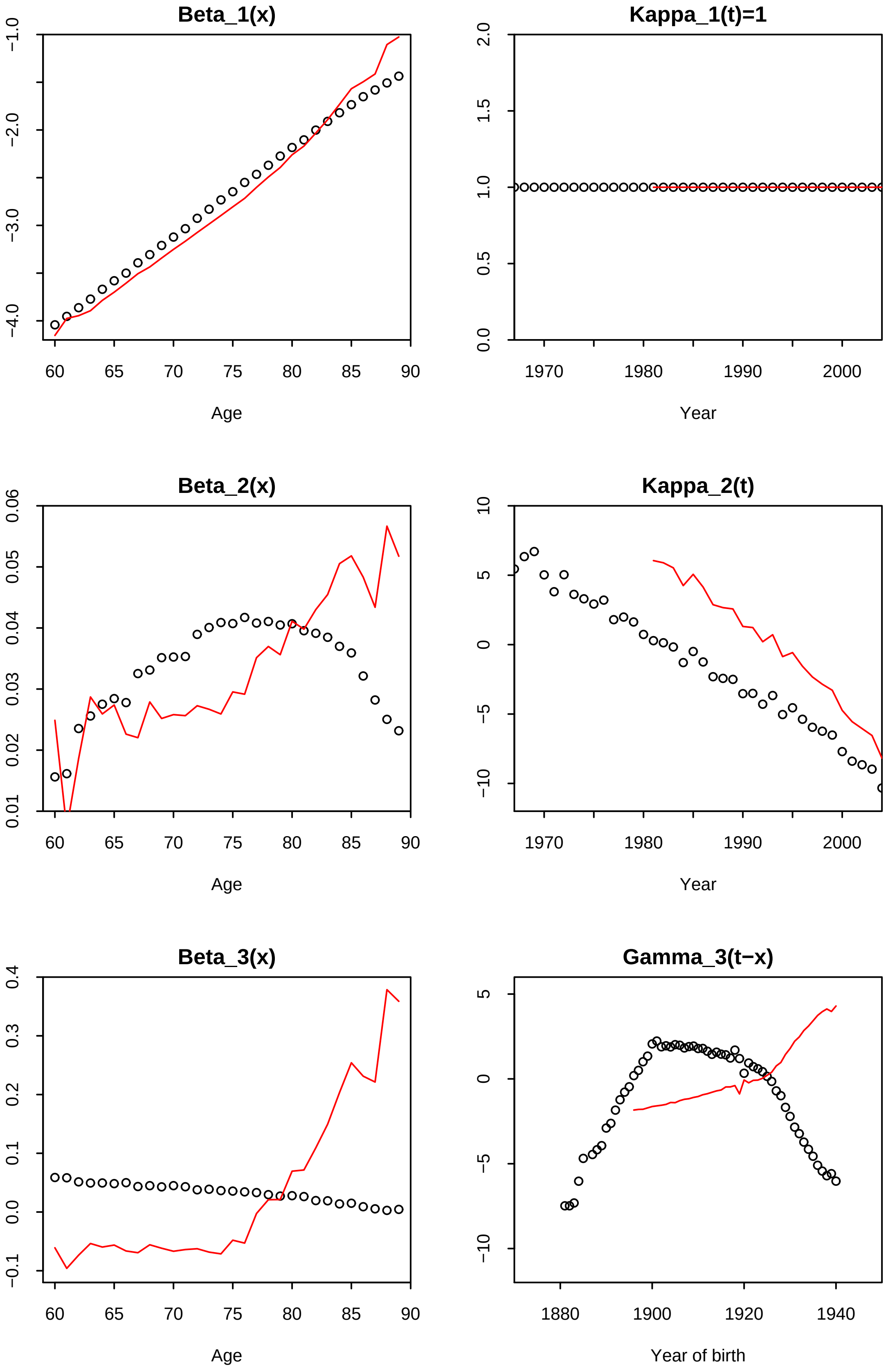


Figure 14: England & Wales data: Parameter estimates for model M2 derived from (a) data from 1961 to 2004 (dots) or (b) data from 1981 to 2004 (solid lines).

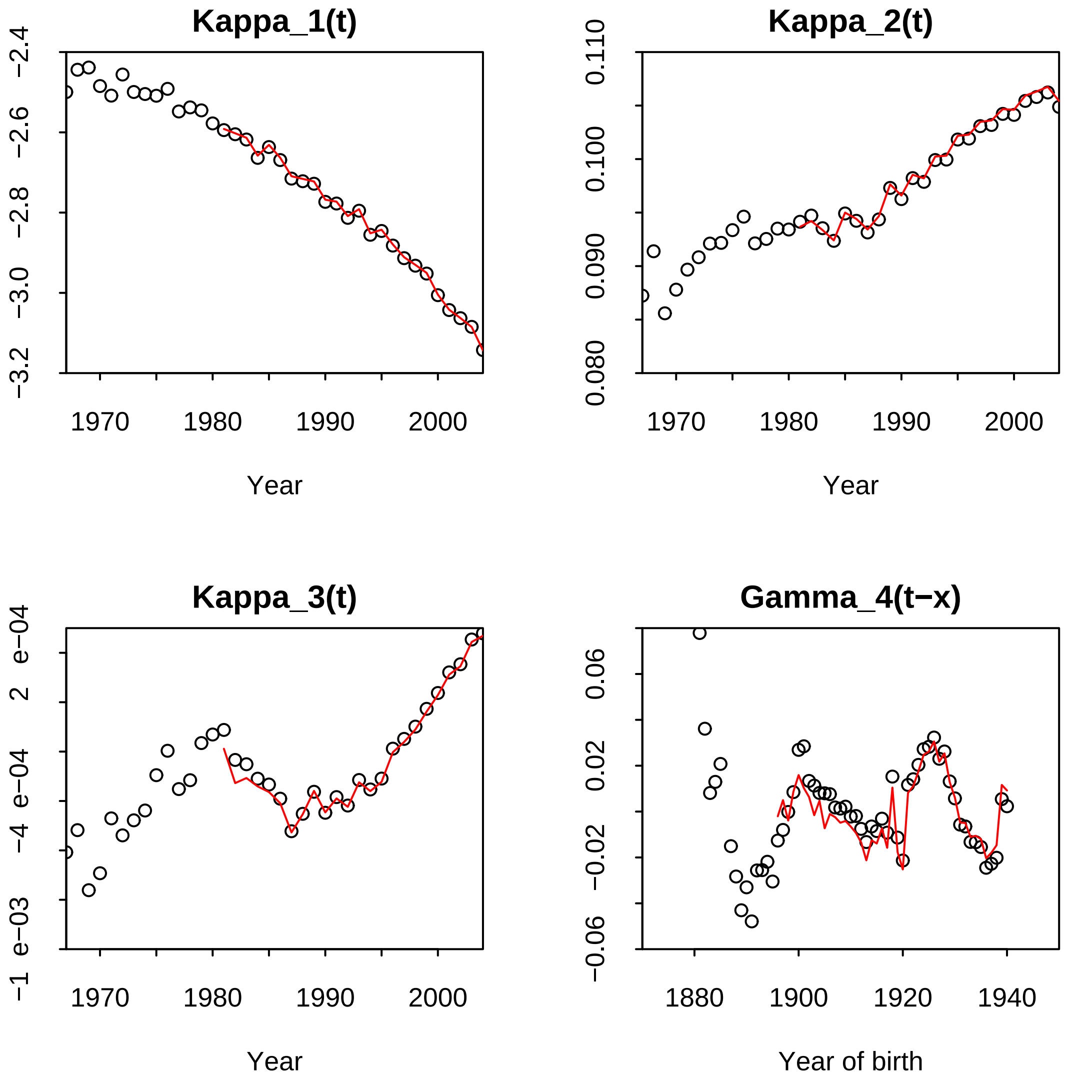


Figure 15: England & Wales data: Parameter estimates for model M7 derived from (a) data from 1961 to 2004 (dots) or (b) data from 1981 to 2004 (solid lines).

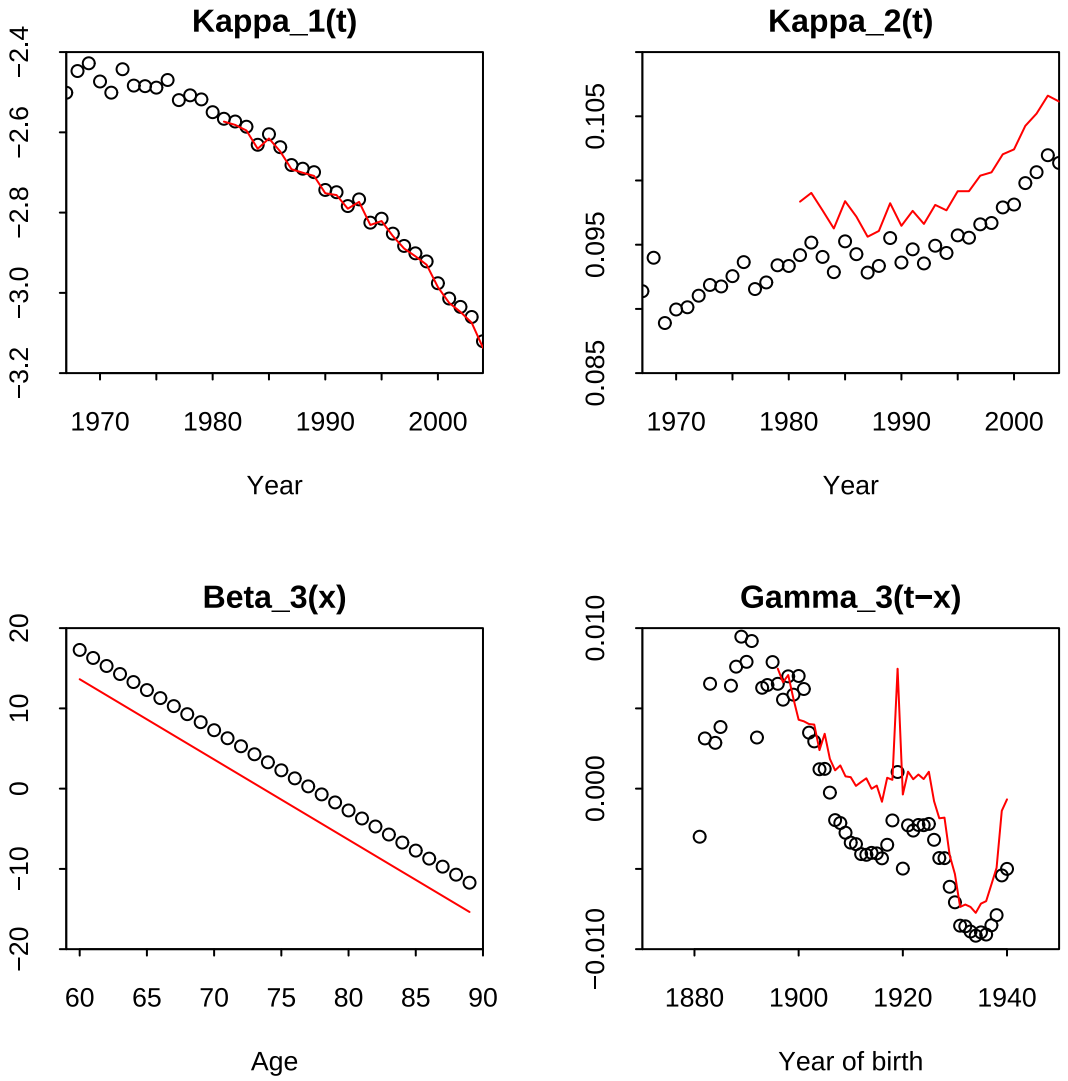


Figure 16: England & Wales data: Parameter estimates for model M8 derived from (a) data from 1961 to 2004 (dots) or (b) data from 1981 to 2004 (solid lines).

A simple means of considering the i.i.d. assumption is to look at the pattern of positive and negative standardised residuals.[12] These are plotted in Figures 17 to 20.[13] Red cells imply that $Z(x,t) > 0$ while blue implies a standardised residual is negative. The i.i.d. hypothesis implies that there should be a random pattern of reds and blues. If the plots, in contrast, reveal some form of pattern or clumping of reds and blues then the i.i.d. assumption becomes inappropriate.

For models M1, M3 and M5 we can see strong clustering of reds and blues. Models M1 and M5 do not incorporate a cohort effect and the clustering of reds and blues show up strong diagonals: strong evidence for the existence of the cohort effect. M6 shows a little bit of clustering. M4, on first glance, looks reasonably random, but closer inspection reveals distinct vertical bands which suggest that there is a genuine random period effect that is being smoothed out too much under M4. M2, M7 and M8 all look reasonably random, and so all pass this "visual" test.

[12] Renshaw and Haberman (2006) use a helpful alternative graphical method to investigate the independence of residuals.

[13] Black and white versions of these figures can be found in the appendices in Figures 42 to 45.

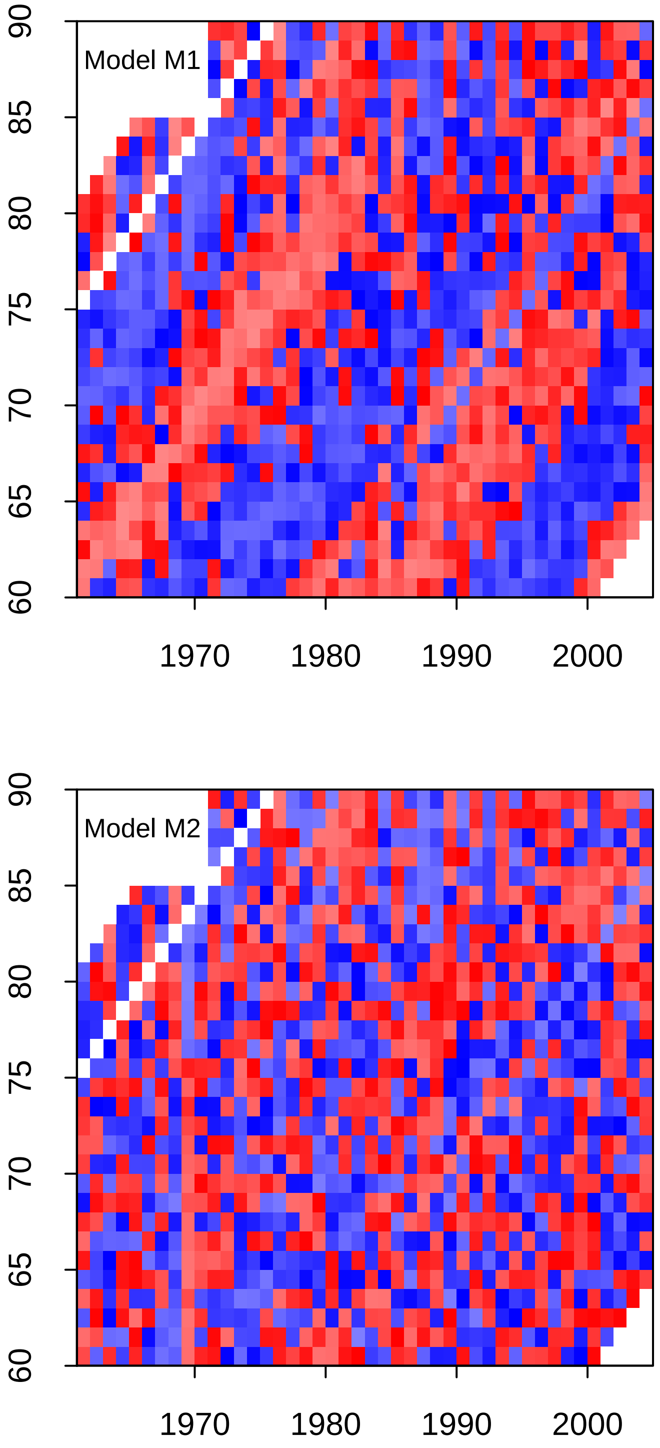
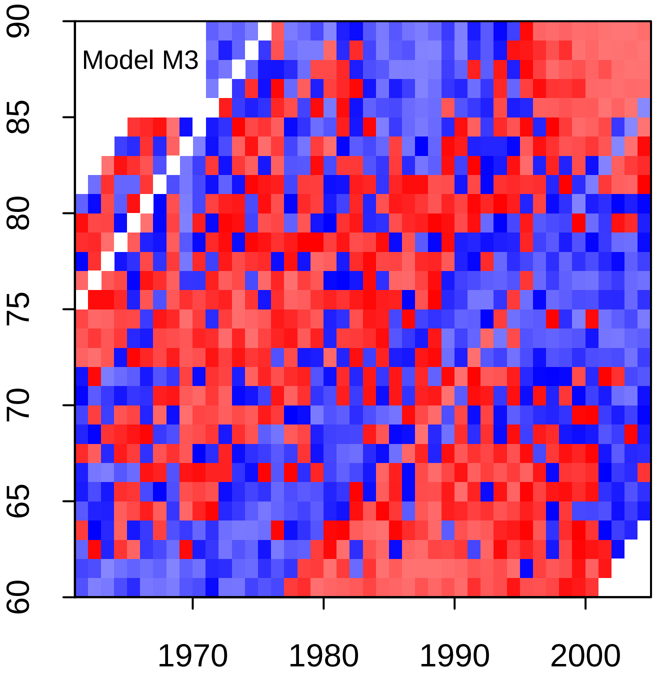


Figure 17: England & Wales data: Standardised residuals $Z(t, x)$ for Models M1 (top) and M2 (bottom). Red cells mean $Z(t, x) > 0$, blue means $Z(t, x) < 0$, white means the cell was excluded from the analysis. Light red and light blue cells indicate $Z(t, x)$ is far from zero. Dark red and blue indicate $Z(t, x)$ is closer to zero.



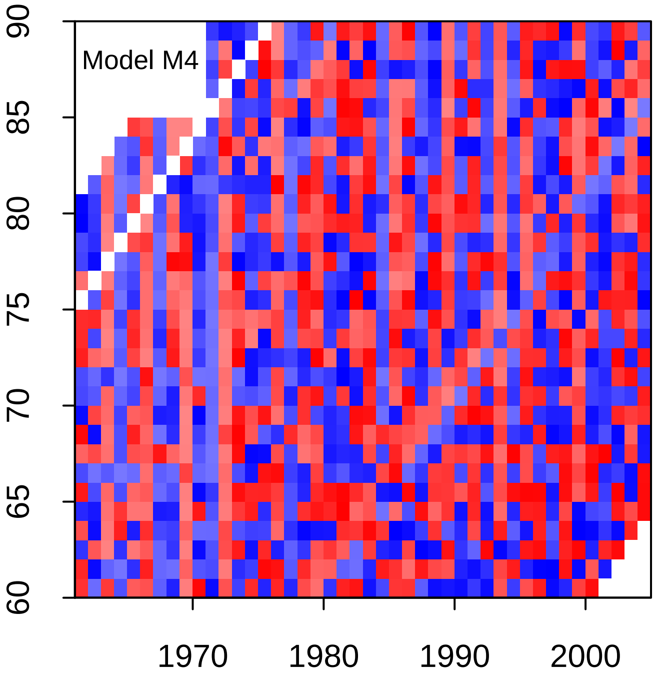
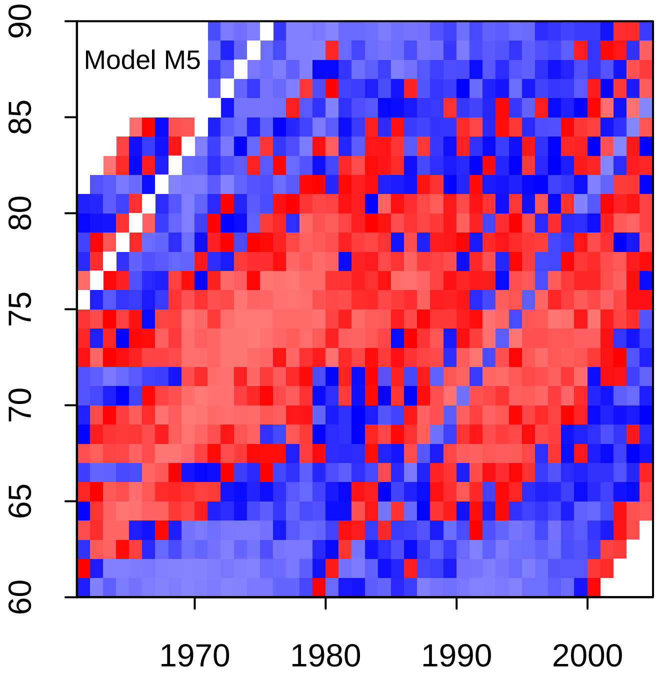


Figure 18: England & Wales data: Standardised residuals $Z(t, x)$ for Models M3 (top) and M4 (bottom). Red cells mean $Z(t, x) > 0$, blue means $Z(t, x) < 0$, white means the cell was excluded from the analysis. Light red and light blue cells indicate $Z(t, x)$ is far from zero. Dark red and blue indicate $Z(t, x)$ is closer to zero.



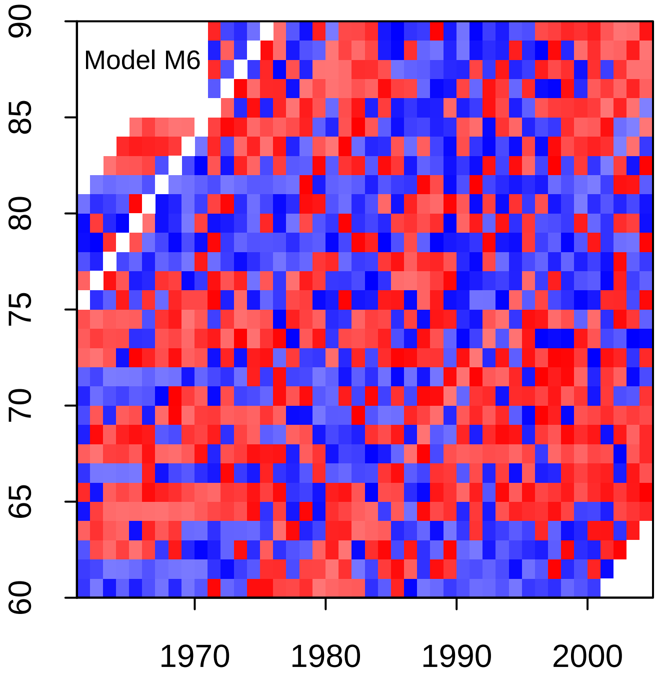


Figure 19: England & Wales data: Standardised residuals $Z(t, x)$ for Models M5 (top) and M6 (bottom). Red cells mean $Z(t, x) > 0$, blue means $Z(t, x) < 0$, white means the cell was excluded from the analysis. Light red and light blue cells indicate $Z(t, x)$ is far from zero. Dark red and blue indicate $Z(t, x)$ is closer to zero.

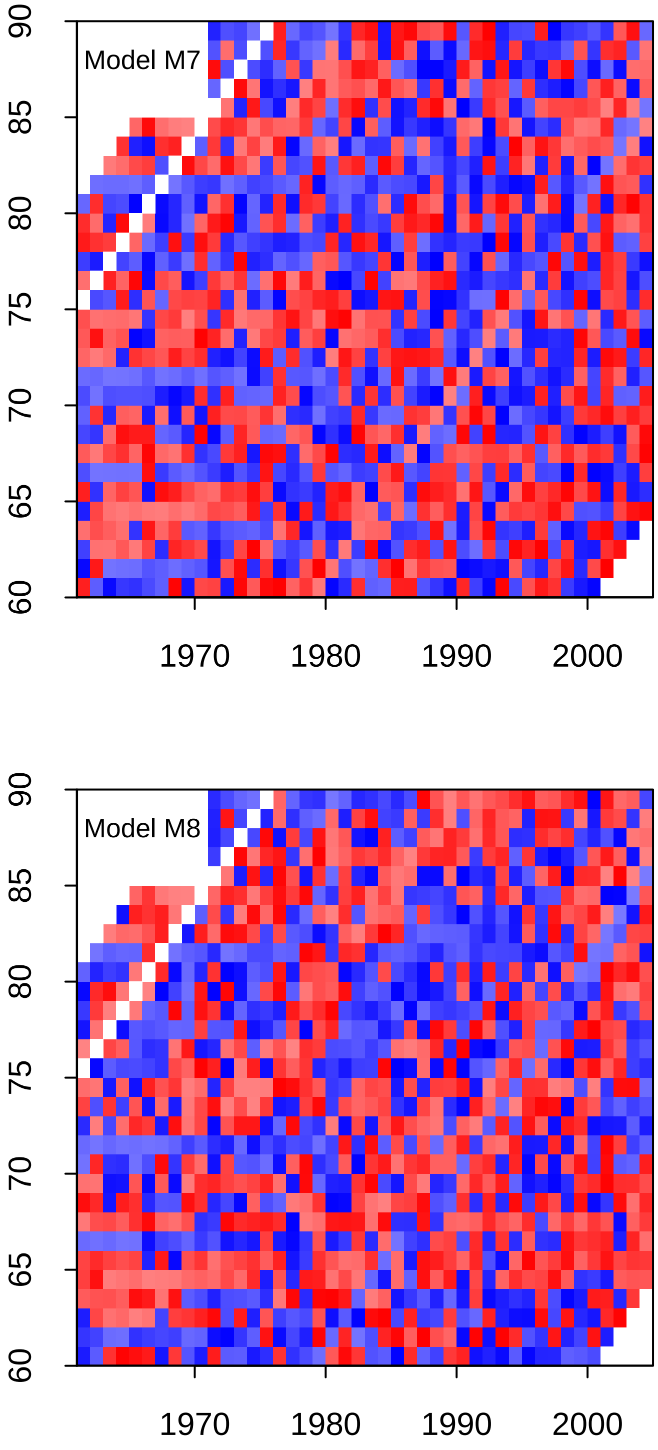


Figure 20: England & Wales data: Standardised residuals $Z(t, x)$ for Models M7 (top) and M8 (bottom). Red cells mean $Z(t, x) > 0$, blue means $Z(t, x) < 0$, white means the cell was excluded from the analysis. Light red and light blue cells indicate $Z(t, x)$ is far from zero. Dark red and blue indicate $Z(t, x)$ is closer to zero.

H_0: restricted model	H_1: general model	LR test statistic	d.f.	p-value
M1	M2	2354.3	101	< 0.000001
M3	M2	1745.0	59	< 0.000001
M5	M6	4226.5	71	< 0.000001
M5	M7	4666.8	114	< 0.000001
M6	M7	440.3	43	< 0.000001
M5	M8	4423.7	74	< 0.000001
M6	M8	197.2	2	< 0.000001

Table 4: England & Wales data: Likelihood Ratio test results for various pairs of general and nested models.

6.5.3 Comparing models that are nested

Some models are nested within one of the others: that is, they are special cases of more general models. For example, model M1 is nested within model M2 being a special case of M2 with $\beta_x^{(3)} = 0$ for all x and $\gamma_{t-x}^{(3)} = 0$ for all $t - x$. In such circumstances we can use the likelihood ratio test to test the null hypothesis that the nested or restricted model is the correct model versus the alternative hypothesis that the more general model is correct. For the nesting above, let $\hat{l}_1$ be the maximum log likelihood for model M1 and $\hat{l}_2$ be the maximum likelihood for model M2. Model M1 requires the estimation of $\nu_1 = 102$ parameters, while M2 requires $\nu_2 = 203$. The likelihood-ratio test statistic is

$$2(\hat{l}_2 - \hat{l}_1).$$

If the null hypothesis is true this should have approximately a chi-squared distribution with $\nu_2 - \nu_1$ degrees of freedom (d.f.). Thus we reject the null hypothesis in favour of the more general model if the test statistic is too large: specifically, if

$$2(\hat{l}_2 - \hat{l}_1) > \chi^2_{\nu_2-\nu_1,\alpha},$$

where α is the significance level. Alternatively we can calculate the p-value for this test as

$$p = 1 - {\chi^2_{\nu_2-\nu_1}}^{-1}\Big(2(\hat{l}_2 - \hat{l}_1)\Big).$$

The eight models consider here include seven nested pairs. Each pair is considered in Table 4. In each case the null hypothesis is rejected overwhelmingly in favour of the more general model.

These results support our earlier results based on the BIC. Additionally their decisive rejection of models M1 and M5, in particular, gives a clear indication that the cohort effect is a key feature of England & Wales males mortality data.

6.6 Model risk: robustness of projections

We will now carry out a number of tests to assess the impact of model choice on key outputs associated with projections of mortality rates into the future. We focus on models M2, M5, M7 and M8.

We first consider the impact of model choice on projected values of the survivor index $S(t, 65)$: the proportion out of the cohort aged 65 (and still alive) in 2004 who are still alive in year $2004 + t$.[14] In Figure 21 (top) we have plotted the mean and 90% confidence interval for the survivor index $S(t, 65)$ for a cohort aged 65 in 2004. It can be seen that these forecasts are little affected by the choice between models M5, M7 and M8. M2 is slightly more out of line but consistent with the others. In Figure 21 (bottom) we have plotted the variance of $\log S(t, 65)$ over time. Again the differences are relatively small between M5, M7 and M8, although they are perhaps more prominent at 25 years. M2 stands out as having a much lower variance, with the implication that model risk might be an issue. For example, a financial contract that contains some optionality based on $S(t, x)$ might have quite a different value under M2 from M5, M7 and M8.

6.7 Model risk: annuity values

As a second exercise we calculated the value in 2004 of an annuity payable until age 90 (the maximum age in our projection model) for males aged 60, 65, 70 and 75 in 2004 at a constant rate of interest of 4%:[15]

$$a_x(2004) = \sum_{t=1}^{90-x} 1.04^{-t} E[S(t, x)].$$

For ages 65, 70, 75, we have already estimated the cohort effect for models M2, M7 and M8. For age 60, $\gamma_{1944}^{(3/4)}$ is not known (that is the value of $\gamma_{t-x}^{(3)}$ or $\gamma_{t-x}^{(4)}$ for the cohort born in $t - x = 1944$). Since a single value is required we adopt a subjective approach and consider two values for $\gamma_{1944}^{(3/4)}$ for each model with the aim of assessing the sensitivity of results to this parameter. The values that we choose fall at the upper and lower ends of what we regard as the plausible range of outcomes for the 1944 cohort based on the historical estimates for earlier cohorts. Looking at Figure 12 we might speculate that extreme high and low values for $\gamma_{1944}^{(4)}$ in model M7 might be the final value plus and minus 0.04. Similarly, the information in Figure 13 suggests that $\gamma_{1944}^{(3)}$ in model M8 might lie between the final value plus and minus 0.01. Finally for M2 we choose values of -8 and -4 based on information in Figure 8.

Values for these ages are given in Table 5. The four models turn out to give relatively similar annuity values at all of the ages considered. At age 60 the differences in value

[14] Projections are based on fitted parameters and use a random walk model in simulations based on the historical estimates of the $\kappa_t^{(i)}$, as proposed in Cairns, Blake and Dowd (2006b) for model M5. Further details are given in appendix C.

[15] The aim here is to identify differences in values between models. We make no attempt therefore to introduce a market price of longevity risk as proposed by Cairns, Blake and Dowd (2006b).

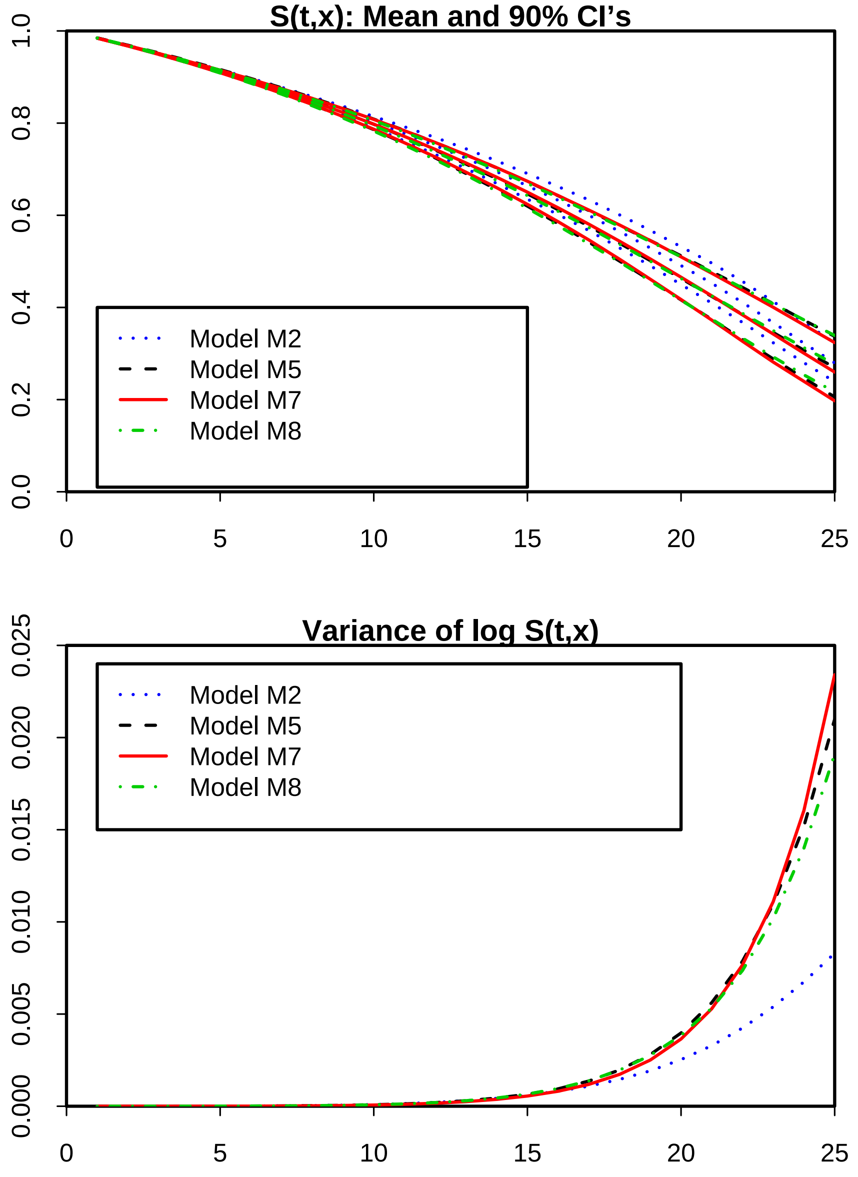


Figure 21: England & Wales data: Top: 5% and 95% quantiles for the survivor index $S(t, 65)$ for models M2, M5, M7 and M8, with the mean of $S(t, 65)$ running down the middle. Bottom: $Var[\log S(t, 65)]$ for models M2, M5, M7 and M8.

Model	M2 $\gamma_{1944}^{(3)}$ $= -8$	M2 $\gamma_{1944}^{(3)}$ $= -4$	M5	M7 $\gamma_{1944}^{(4)}$ $= -0.0398$	M7 $\gamma_{1944}^{(4)}$ $= 0.0402$	M8 $\gamma_{1944}^{(3)}$ $= -0.0209$	M8 $\gamma_{1944}^{(3)}$ $= -0.0009$
$x = 60$	13.220	13.659	13.472	13.557	13.350	13.669	13.363
$x = 65$	11.591		11.449	11.451		11.427	
$x = 70$	9.505		9.325	9.354		9.314	
$x = 75$	7.270		7.220	7.240		7.190	

Table 5: England & Wales data: Annuity values for various ages based on data from 1984 to 2004. Values for $\gamma_{1944}^{(3/4)}$, are estimated.

are slightly bigger (2.4% between the largest and the smallest). This is not surprising given we are now having to estimate a cohort effect at this age. However, despite the fact that we have used a relatively-extreme range of values for $\gamma_{1944}^{(3/4)}$ the range of annuity values is not that wide. We therefore conclude that the cohort effect, while statistically significant (in the sense of Table 3), has an economically small effect, for the annuity values considered here.

Model	Maximum log- likelihood	Effective number of parameters	BIC (rank)
M1	-12265.4	94	-12590.0 (6)
M2	-9737.4	187	**-10383.2 (1)**
M3	-11854.2	128	-12296.3 (3)
M5	-16121.3	72	-16370.00 (7)
M6	-11948.4	135	-12414.7 (5)
M7	-11631.7	170	-12218.9 (2)
M8	-11841.1	137	-12314.3 (4)

Table 6: US males ages 60 to 89 and years 1968 to 2003. Maximum likelihood, effective number of parameters estimated and Bayes Information Criterion (BIC) for each model. Effective number of parameters takes account of the constraints on parameters.

7 Analysis of US data

7.1 Estimation and preliminary data analysis

The US males' mortality data were analysed using models M1 to M3 and M5 to M8[16] using the data from 1968 to 2003 and for males aged 60 to 89. Over the period 1968 to 1979 ages 85-89 were excluded on the basis that the data at those ages in those years are not reliable.

For each model we estimated the $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_{t-x}^{(i)}$ parameters using maximum likelihood using data covering the period 1968 to 2003. Maximum likelihoods are given in Table 6. Estimates of the parameters themselves are plotted in Figures 22 to 28. In these plots the dots are parameter estimates based on data from 1968 to 2003, while lines are based on data from 1980 to 2003.

7.2 Model selection criteria

Section 6.2 describes how and why we choose to penalise the maximum likelihood according to the number of parameters estimated, resulting in the Bayes Information Criterion (BIC).

The BIC for each model is given in the final column in Table 6. In contrast to the results in Table 3 for the England & Wales (EW) data, model M2 now comes out significantly better than the other models. However, in the subsections that follow we will discuss graphical diagnostic tests that suggest M2 might be over-fitting the data (especially when it is suspected – Section 2.2.1 – that the exposures data contain significant errors), and lead us to doubt its robustness.

[16] Model M4 is a popular one used in the UK but not currently in the USA. Our earlier analysis of the EW data did not find that M4 fitted very well in comparison with M2, M7 and M8. We have not therefore used it in our analysis of the US data.

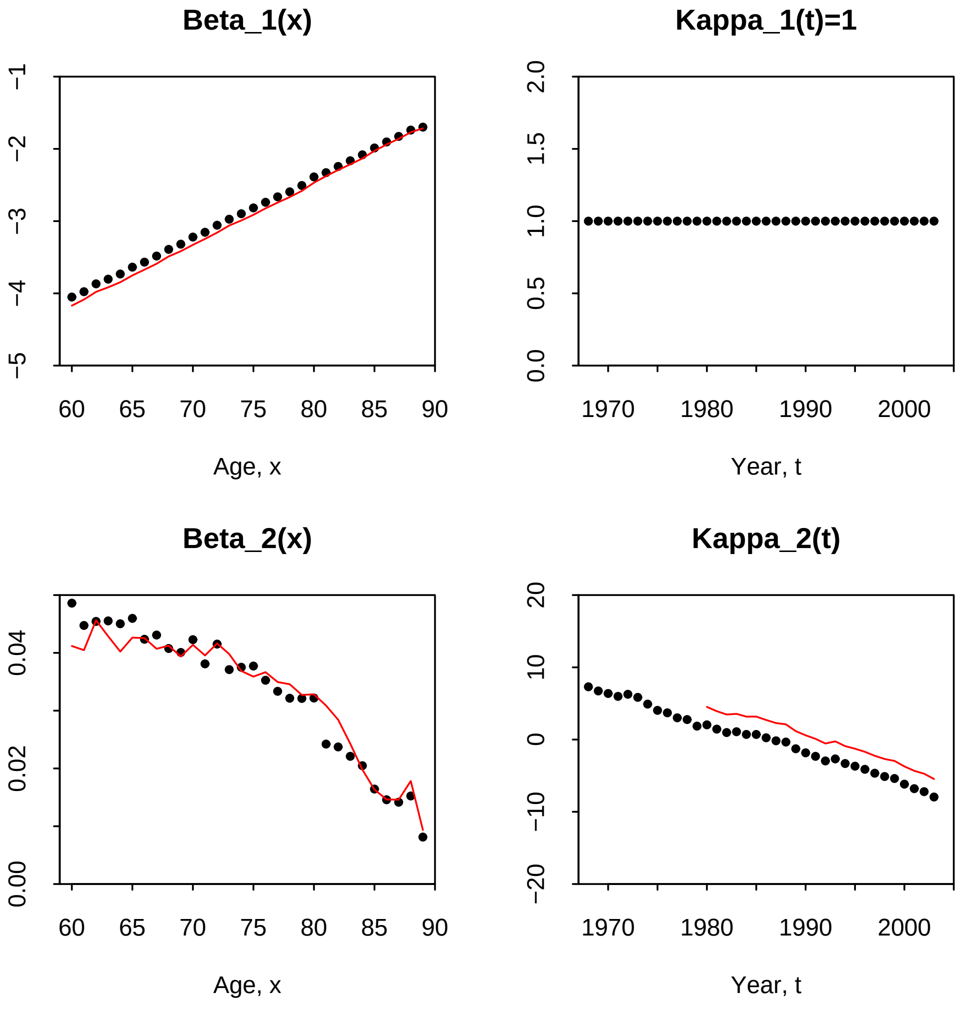


Figure 22: US data: (a) 1968 to 2003 (dots) (b) 1980 to 2003 (solid lines). Parameter estimates for model M1.

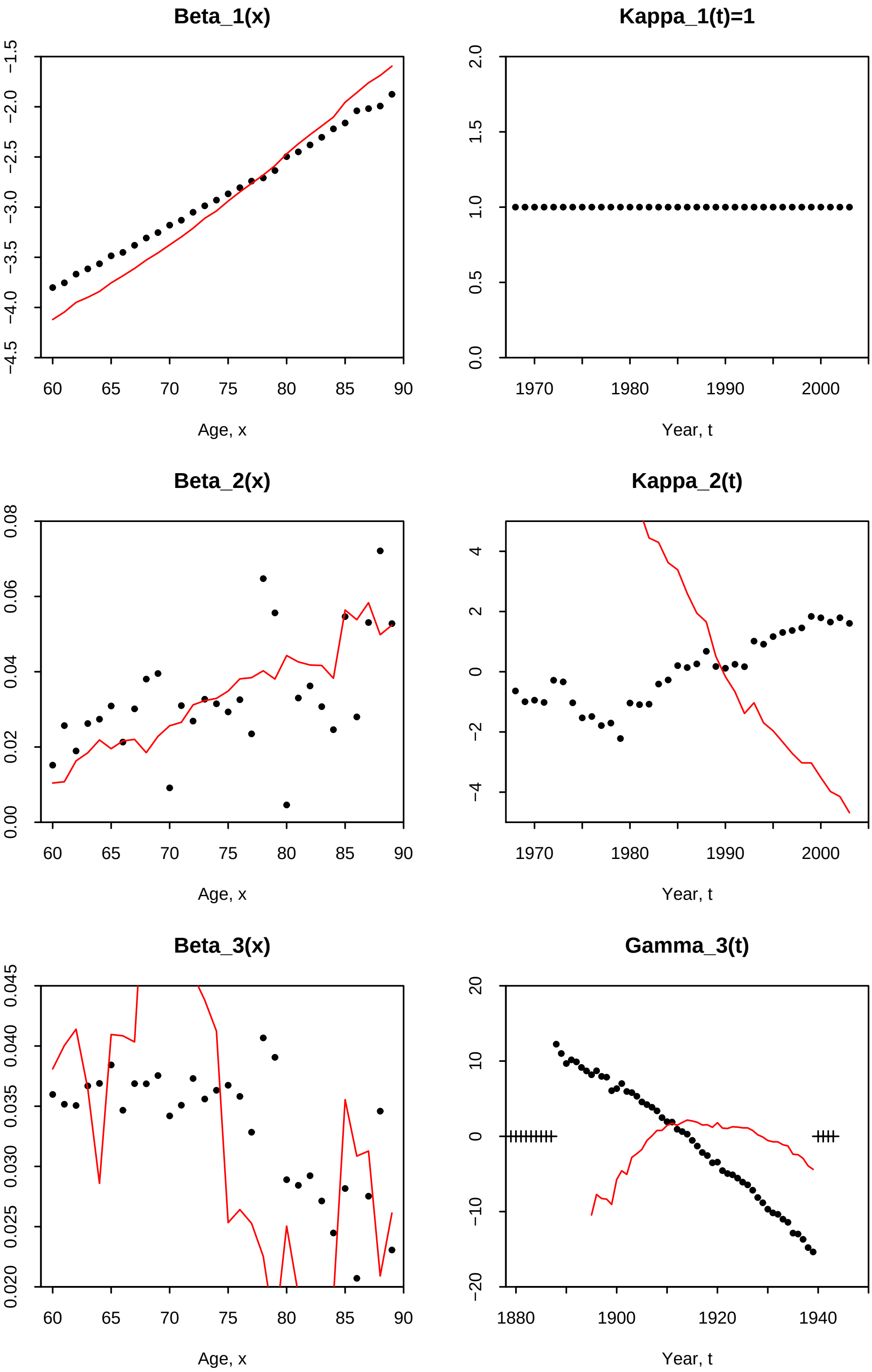


Figure 23: US data: (a) 1968 to 2003 (dots) (b) 1980 to 2003 (solid lines). Parameter estimates for model M2. Crosses in the bottom right plot correspond to excluded cohorts.

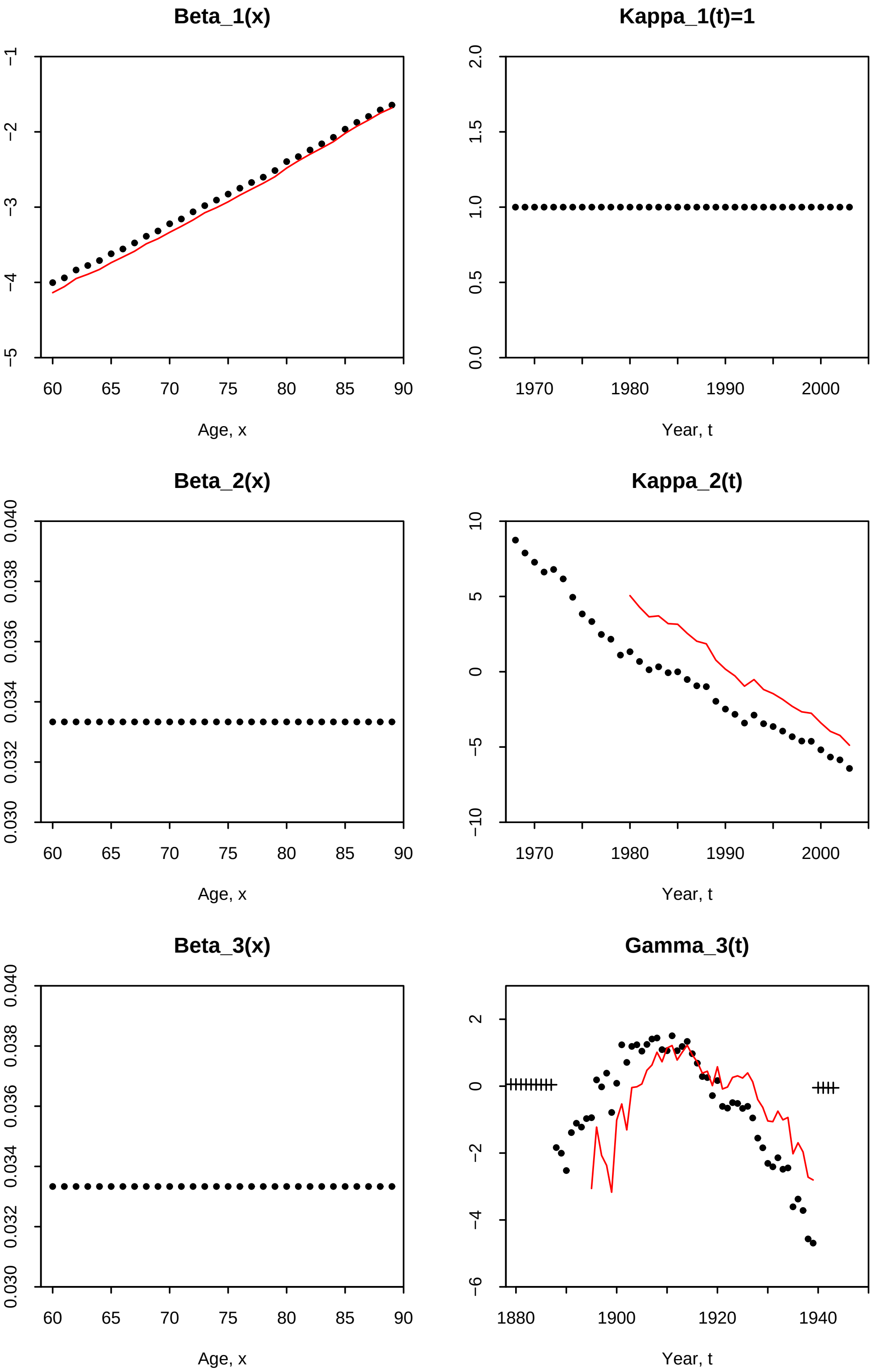


Figure 24: US data: (a) 1968 to 2003 (dots) (b) 1980 to 2003 (solid lines). Parameter estimates for model M3. Crosses in the bottom right plot correspond to excluded cohorts.

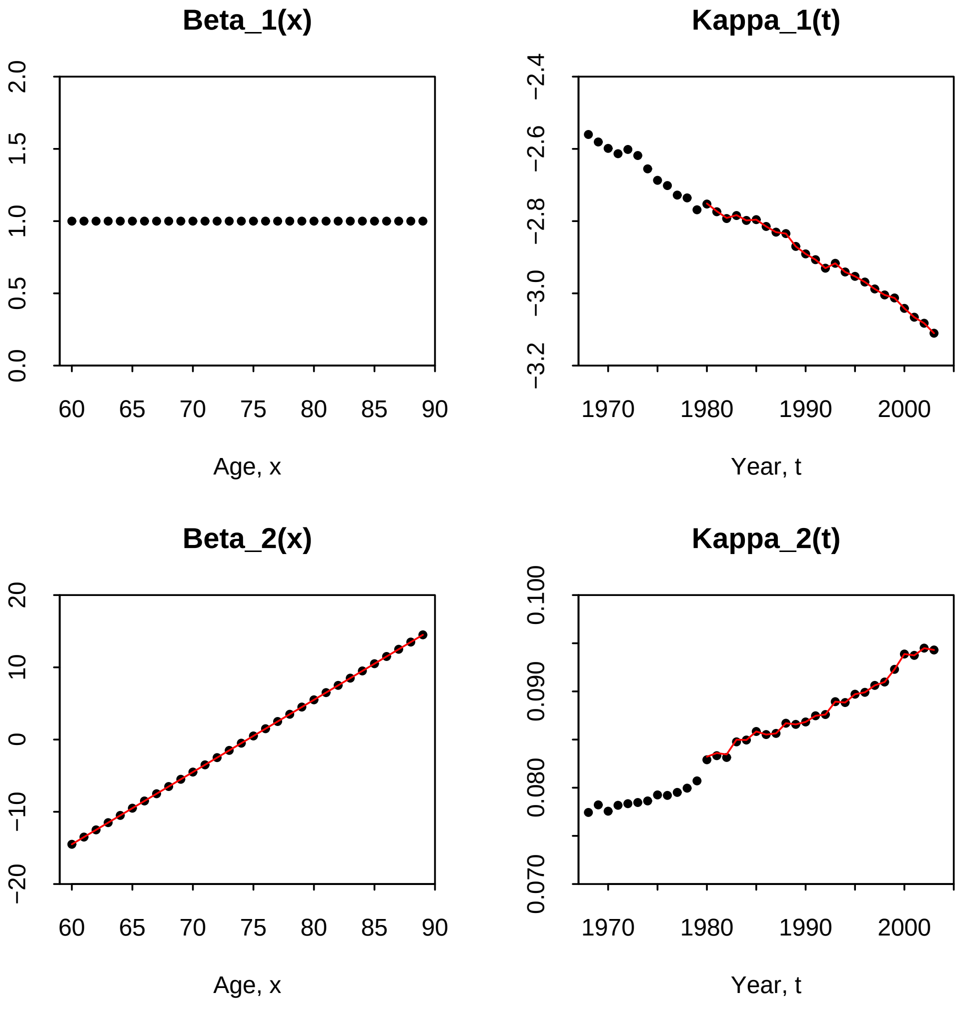


Figure 25: US data: (a) 1968 to 2003 (dots) (b) 1980 to 2003 (solid lines). Parameter estimates for model M5.

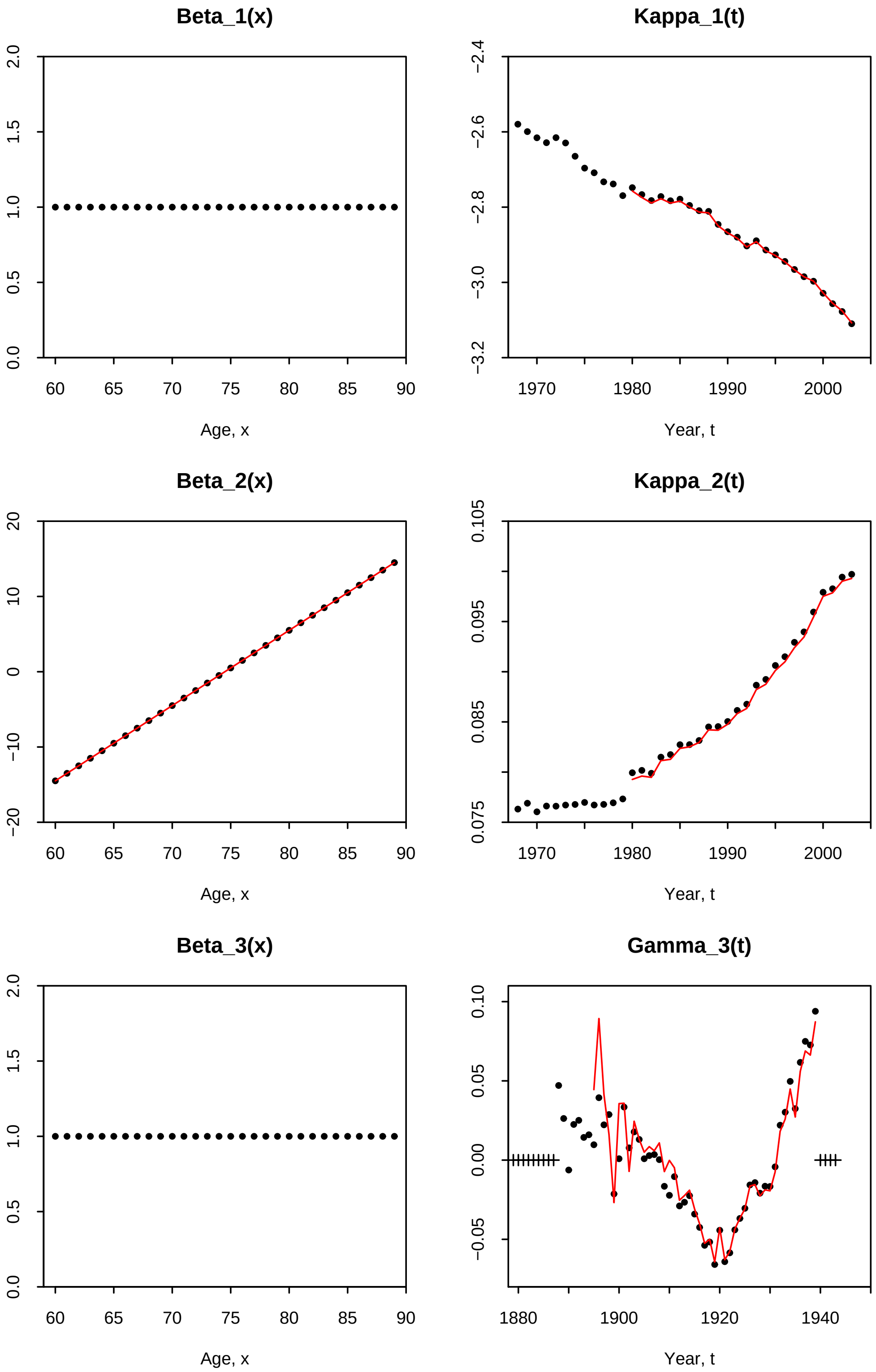


Figure 26: US data: (a) 1968 to 2003 (dots) (b) 1980 to 2003 (solid lines). Parameter estimates for model M6. Crosses in the bottom right plot correspond to excluded cohorts.

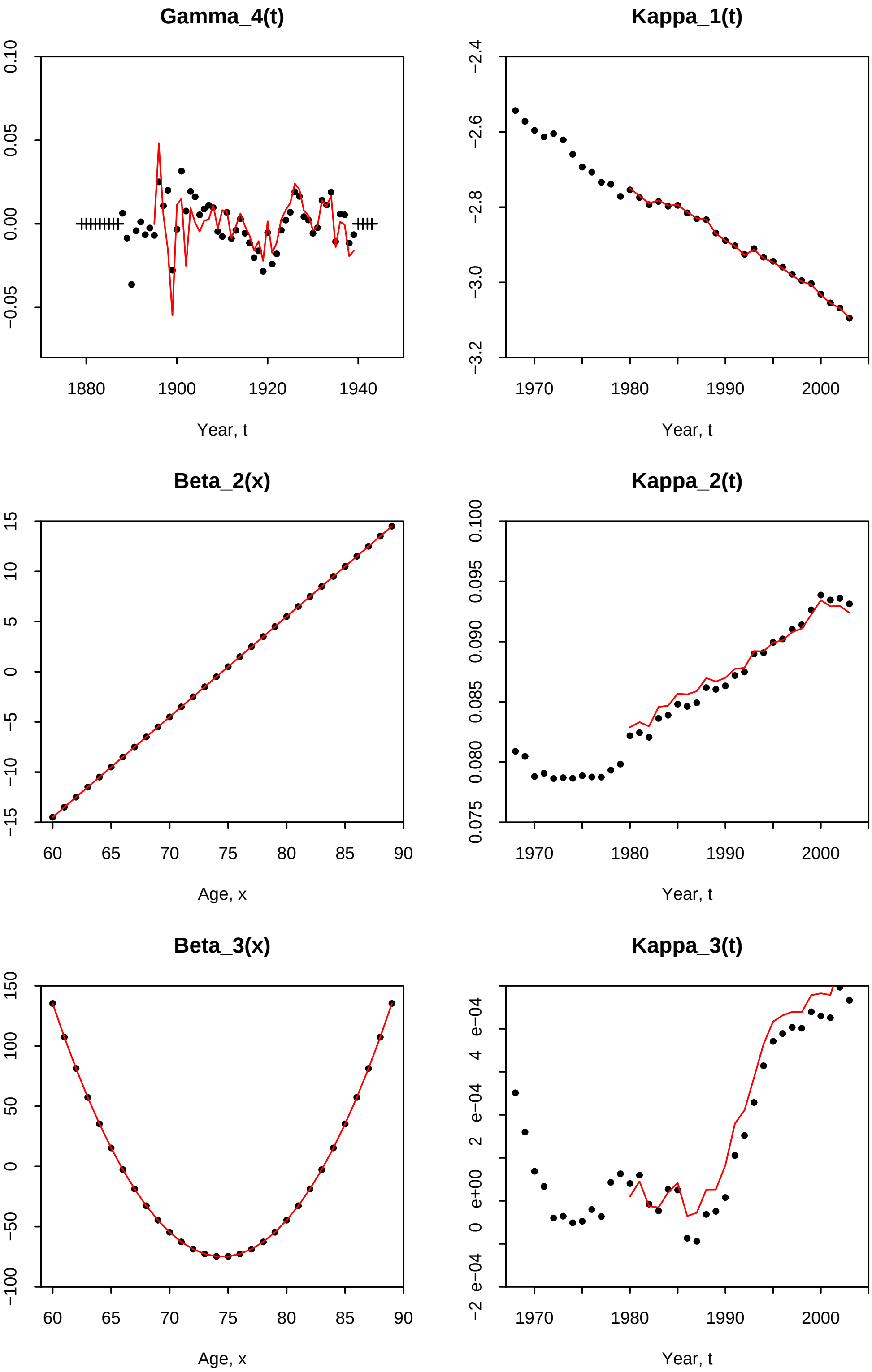


Figure 27: US data: (a) 1968 to 2003 (dots) (b) 1980 to 2003 (solid lines). Parameter estimates for model M7. Crosses in the bottom right plot correspond to excluded cohorts.

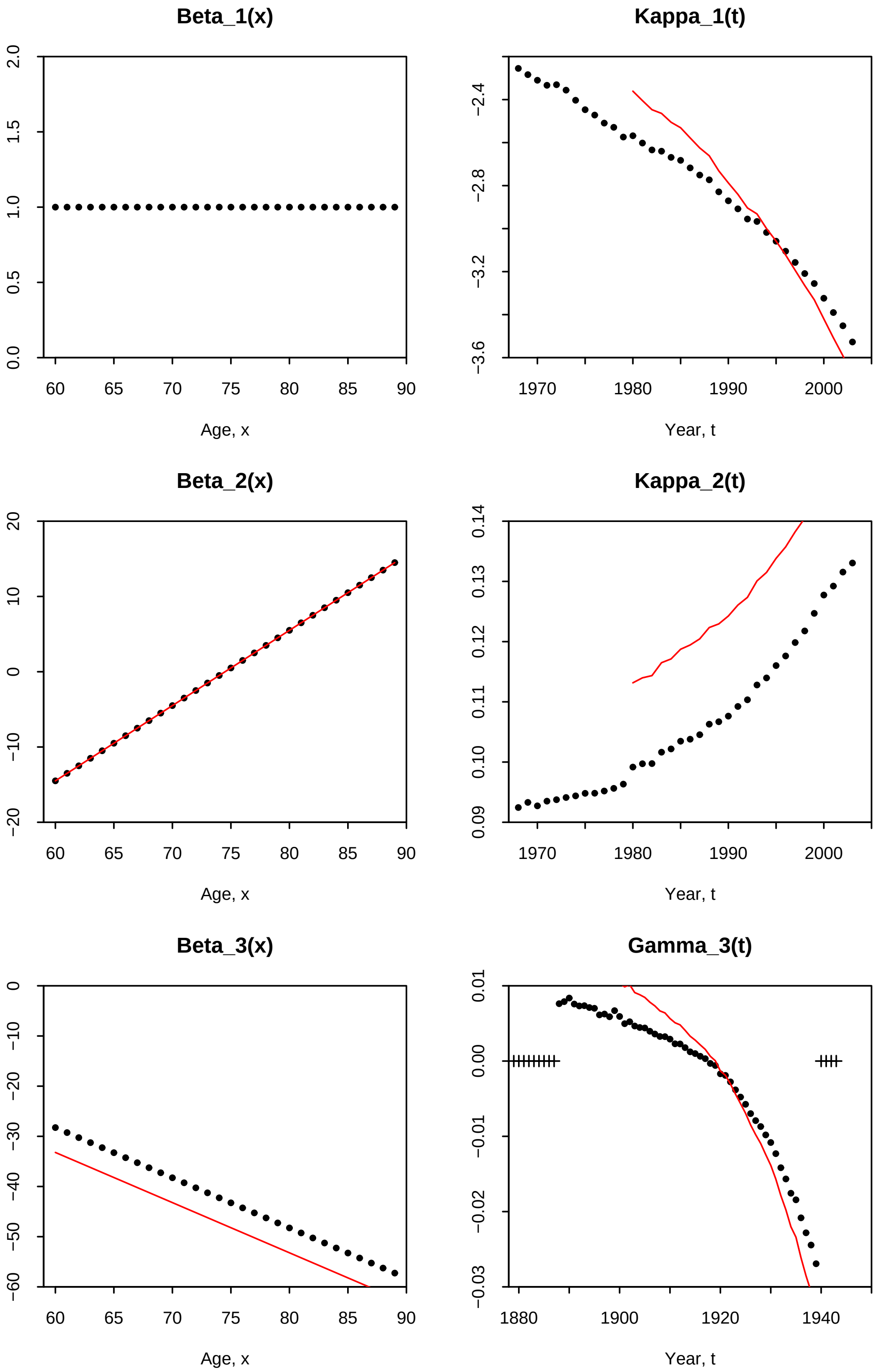


Figure 28: US data: (a) 1968 to 2003 (dots) (b) 1980 to 2003 (solid lines). Parameter estimates for model M8. Crosses in the bottom right plot correspond to excluded cohorts.

7.3 Parameter estimates

Parameter estimates for the seven models are plotted in Figures 22 to 28. For those that incorporate a cohort effect, this is quite prominent. However, the form of the effect does seem to vary from one model to another (for example, M3 versus M7). We will focus our remaining remarks on models M2, M7 and M8.

Figure 23 should be contrasted with its EW counterpart, Figure 8. The patterns here are quite different. The strong, almost linear trend in $\gamma_{t-x}^{(3)}$ is rather indicative of a steady period effect which is independent[17] of the $\kappa_t^{(2)}$ period effect. Consquently, although M2 scores the highest BIC, Figure 23 suggests that a model with an additional period factor might be required.

A further concern about the suitability of M2 form comes to mind when we look at estimates for $\beta_x^{(2)}$ and $\beta_x^{(3)}$. These display a much higher degree of randomness than we saw in the EW data. As we have discussed in Section 4.9 we would expect to see smoothness in each of the age effects: it is difficult to think of any biological or environmental factors that would result in this level of randomness in $\beta_x^{(2)}$ and $\beta_x^{(3)}$. This randomness might suggest M2 is overfitting the US data.

Now compare Figure 27 (model M7) with its EW counterpart Figure 12. The pattern of development of the various parameters in M7 is relatively consistent between the two countries. The main difference that we can identify under model M7 is that $\gamma_{t-x}^{(4)}$ has a less well-defined pattern in the US results, and a greater degree of randomness. This suggests that cohort-related trends in mortality are less important in the US than in England & Wales. What remains of a cohort effect in the US data might be the result of over-fitting or perhaps due to genuine environmental factors that affect each cohort in their year of birth and which vary randomly from year to year (e.g. influenza epidemics).

For model M8 (Figure 28 compared with Figure 13) we see similar issues to M2 in terms of the trend in $\gamma_{t-x}^{(3)}$. Again this trend points to a systematic period effect that is independent of the existing period effects modelled by $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$.

7.4 Robustness of parameter estimates

Figures 22 to 28 also include parameter estimates for each model based on data from 1980 to 2004 (solid lines in the plots). If we compare these with the original parameter estimates based on data from 1968 to 2003 (dots) we can make similar points to the England & Wales data for each of the models.

The simpler models, M1, M3 and M5, tend to show greater robustness.[18]

M7 again seems to be the most robust out of M2, M7 and M8 while M2 again has problems, leading us to question its reliability as a means of projecting mortality rates.

[17] That is, $\beta_x^{(2)}$ and $\beta_x^{(3)}$ are not identical.

[18] Much of the shifts we see in M1 and M3 could be eliminated by adjusting the constraints.

H_0: restricted model	H_1: general model	LR test statistic	d.f.	p-value
M1	M2	5056.0	93	< 0.000001
M3	M2	4233.8	59	< 0.000001
M5	M6	8345.7	63	< 0.000001
M5	M7	8979.2	98	< 0.000001
M6	M7	633.4	35	< 0.000001
M5	M8	8560.5	66	< 0.000001
M6	M8	214.7	2	< 0.000001

Table 7: US data: Likelihood Ratio test results for various pairs of general and nested models.

7.5 Other methods of comparison

7.5.1 Standardised residuals

Recall that the likelihood function assumes that the standardised residuals (see equation 1) are approximately i.i.d. $N(0,1)$: and specifically should have variance 1.

The variances of the standardised residuals for the US data are very much higher than for England & Wales. Using 1968 to 2003 data the variance is around 7.5 for model M2 and 11.5 for M7 and M8 (this size of difference being consistent, of course, with the differences between the log-likelihoods). Using data from 1980 to 2003 these fall to about 3.3 and 7.5 respectively. If the data were wholly reliable, the Poisson assumption the right one and the model the correct one, then this variance should be around one. The high values we see here, therefore, lend weight to our earlier remarks concerning inaccuracies in the exposures data.

7.5.2 Patterns of standardised residuals

In Figures 29 to 32 we have plotted standardised residuals for each of the models. For the underlying model to be valid the positive and negative residuals should be randomly distributed.

All seven plots exhibit some degree of clustering. Out of these M2 looks the most random but comparison of this with its England & Wales counterpart (Figure 17) suggests that M2 fits the US data less well in terms of the i.i.d. assumption.

7.5.3 Comparing models that are nested

We carried out likelihood ratio tests on models that are nested (that is, where one model is a special case of another), as an alternative to model selection using the BIC.

Test results for all seven nested pairs are presented in Table 7. These results support our earlier conclusions based on the BIC: namely that the more complex models succeed in fitting the data better than the simpler models. Unfortunately the nesting

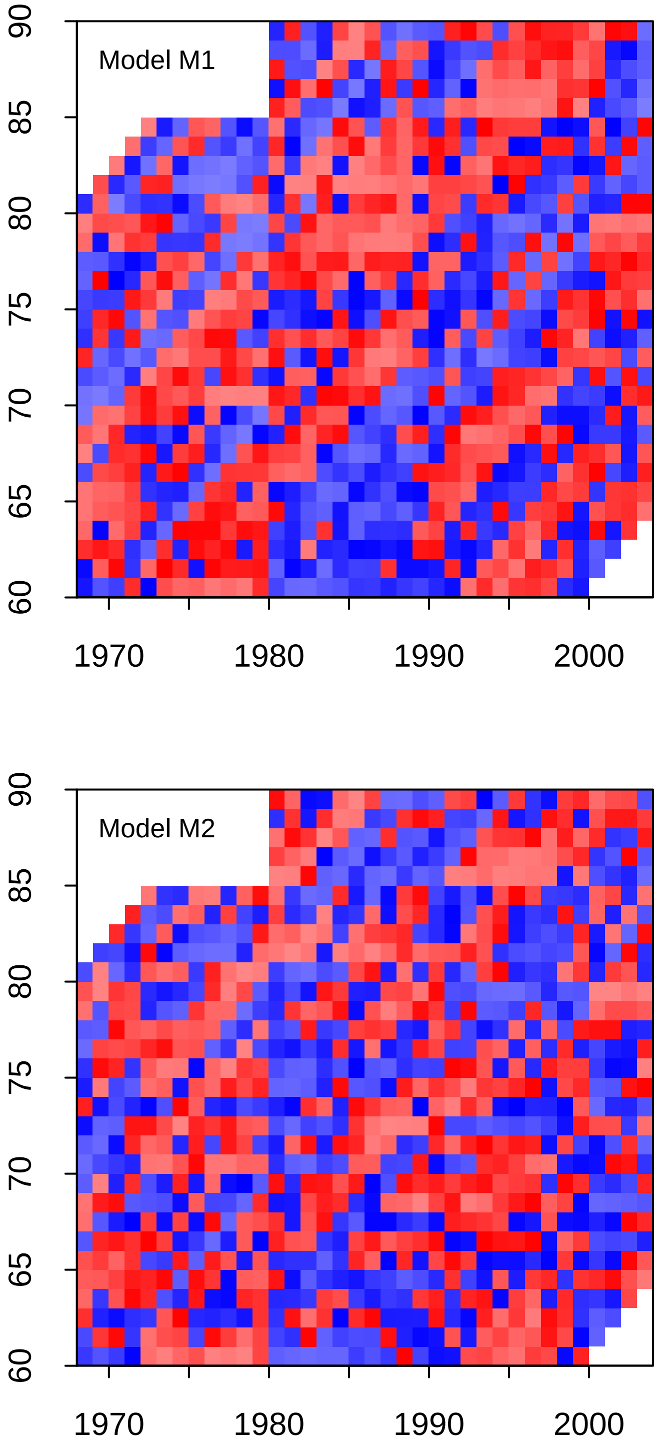


Figure 29: US data. Standardised residuals $Z(t, x)$ for Models M1 (top) and M2 (bottom). Red cells mean $Z(t, x) > 0$, blue means $Z(t, x) < 0$, white means the cell was excluded from the analysis. Light red and light blue cells indicate $Z(t, x)$ is far from zero. Dark red and blue indicate $Z(t, x)$ is closer to zero.

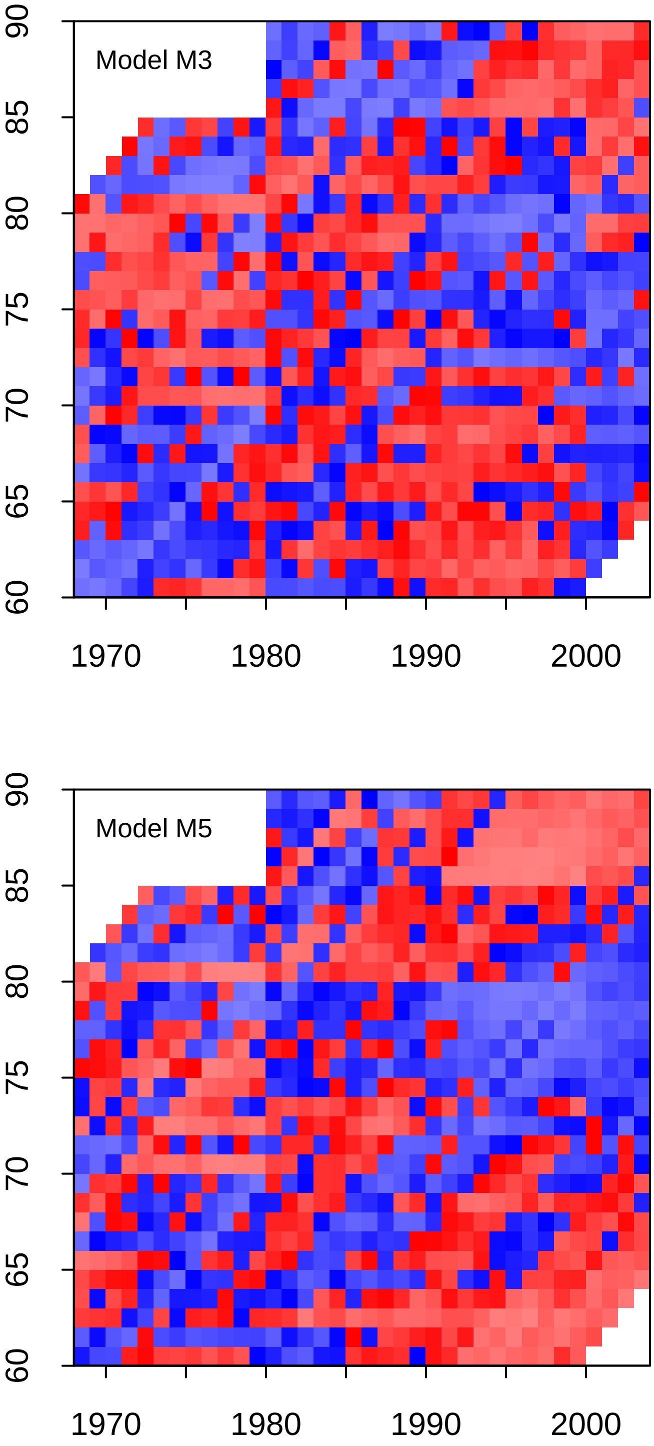
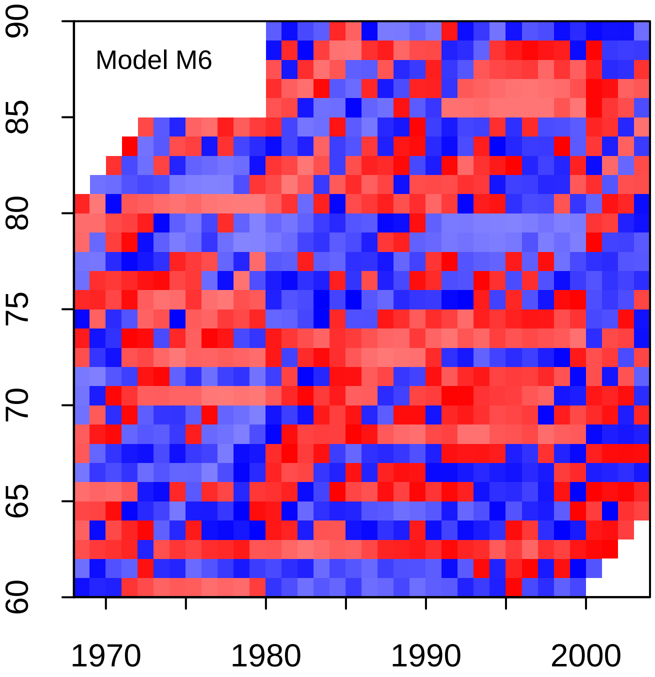


Figure 30: US data. Standardised residuals $Z(t, x)$ for Models M3 (top) and M5 (bottom). Red cells mean $Z(t, x) > 0$, blue means $Z(t, x) < 0$, white means the cell was excluded from the analysis. Light red and light blue cells indicate $Z(t, x)$ is far from zero. Dark red and blue indicate $Z(t, x)$ is closer to zero.



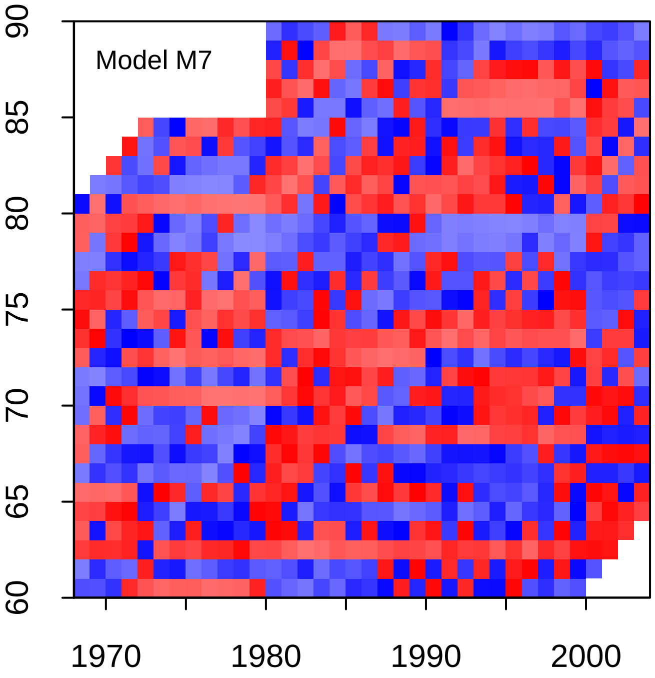


Figure 31: US data. Standardised residuals $Z(t,x)$ for Models M6 (top) and M7 (bottom). Red cells mean $Z(t,x) > 0$, blue means $Z(t,x) < 0$, white means the cell was excluded from the analysis. Light red and light blue cells indicate $Z(t,x)$ is far from zero. Dark red and blue indicate $Z(t,x)$ is closer to zero.

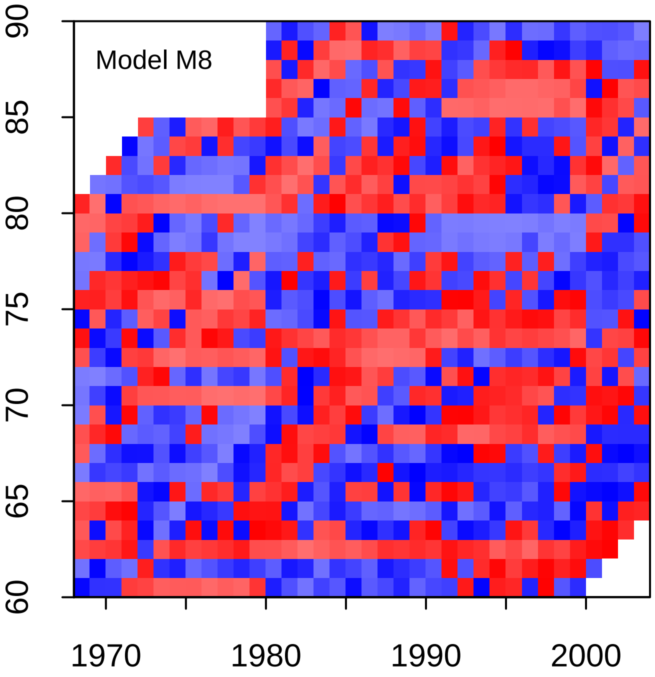


Figure 32: US data. Standardised residuals $Z(t,x)$ for Models M8. Red cells mean $Z(t,x) > 0$, blue means $Z(t,x) < 0$, white means the cell was excluded from the analysis. Light red and light blue cells indicate $Z(t,x)$ is far from zero. Dark red and blue indicate $Z(t,x)$ is closer to zero.

requirements of the test prevent us from comparing M2, M7 and M8.

One can compare Tables 4 and 7 to investigate the relative importance of specific model features. For example, compare M6 with M7. With the US data, the test statistic is larger than the EW test statistic with fewer degrees of freedom, and this indicates that the quadratic age-effect is more prominent in the US data. For the same reason, model M5 performs much more poorly with the US data relative to the other models.

We also compared M7 with a special case that constrained the cohort effect $\gamma_{t-x}^{(4)}$ to be zero. The unconstrained case was found to be very significantly better: that is, both the quadratic age effect and the cohort effect in M7 are significant.

7.6 Model risk: robustness of projections

It is interesting and informative to consider the differences between projections using different models. In Figure 33 (top) we have plotted the mean and 90% confidence intervals for the survivor index $S(t, 65)$: that is the proportion out of those alive and aged 65 in 2003 surviving to year $2003 + t$.[19] Models M2, M5 and M7 produce relatively similar projections although the confidence interval for M2 is narrower: a feature that is more obvious if we look at the variance of $\log S(t, 65)$ in the lower half of the figure.

M8 stands out as being substantially different, and is a consequence of the fitted cohort effect under M8. The steepening of $\gamma_{t-x}^{(3)}$ around 1920 in combination with the negative fitted values for $\beta_x^{(3)}$ imply that cohorts born after 1920 have increasingly poor mortality relative to the $\kappa_t^{(1)}$ improving trend. This form of cohort effect also appears in model M6, but not in any of the other models. As a consequence M8 relative to M2, M5 and M7 has substantially lower survivor rates in the 2003 age-65 cohort.

7.7 Model risk: annuity values

We calculated the value of a 25-year annuity payable to a male aged 65 in 2003. Values for models M2, M5, M7 and M8 based on 1968-2003 and 1980-2003 data are given in Table 8.

From this table we can see that the relative lack of robustness in parameter estimates for M2 and M8 means that values under those two models are moderately sensitive to changes in the period of data used. More noticeable is how relatively low the M8 values are and this reflects the adverse cohort effect discussed in the previous subsection.

[19] See appendix C for further details.

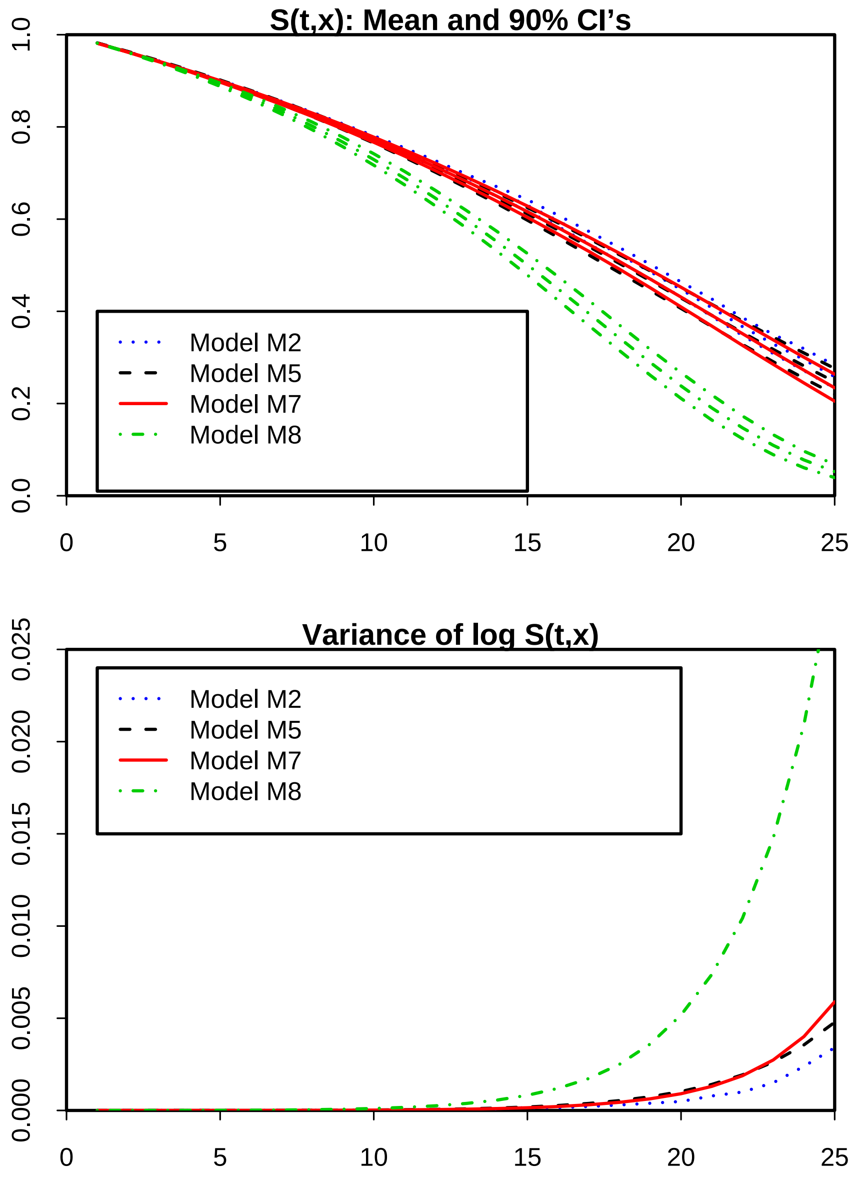


Figure 33: US data: Top: 90% confidence intervals for the survivor index $S(t, 65)$ for models M2, M7 and M8, with the mean of $S(t, 65)$ running down the middle. Bottom: $Var[\log S(t, 65)]$ for models M2, M7 and M8.

Model	M2	M5	M7	M8
1968-2003	11.213	11.101	11.107	9.962
1980-2003	11.524	11.101	11.119	9.685

Table 8: 25-year annuity values for 65-year-old US male calculated under different models and based on different periods of historical data.

8 Conclusions

We have attempted to explain mortality improvements for males aged 60 to 89 in England & Wales and the US using eight stochastic mortality models that decompose mortality improvements into one or more age-, period-, and cohort-related effects. No single model stands out as being best in all aspects. However, different models have different strengths. For example, the Lee-Carter class of models allows for greater flexibility in the age effects, $\beta_x^{(i)}$. One-dimensional P-splines can be exploited to smooth these age effects if the roughness of the $\beta_x^{(i)}$ is seen as a drawback. The CBD-Perks family of models by contrast imposes smoothness in the age effects as an assumption, but allows for richer period effects than the Lee-Carter class. We therefore need to balance up the pros and cons of each model in order to form a conclusion. To some extent it is up to the reader to decide the weights they place on the different selection criteria.

If the reader looks only at the BIC ranking criterion, then model M8 for the England & Wales data and model M2 for the US data dominate. However, if the reader takes into account the robustness of the parameter estimates, then model M7 is preferred for both data sets. This model fits both data sets well and the stability of the parameter estimates over time enables one to place some degree of trust in its projections of mortality rates. The lack of robustness in the other models means that we cannot wholly rely on projections produced by them.

Model M7 shows that mortality rates in both England & Wales and the US have the following features in common (see Figures 15 and 27):

- Mortality rates have been improving over time at all ages: the 'intercept' period term ($\kappa_t^{(1)}$) has been declining over time, so that the upward-sloping plot of the logit of mortality rates against age has been shifting downwards over time.

- These improvements have been greater at lower ages than at higher ages: the 'slope' period term ($\kappa_t^{(2)}$) has been increasing over time, so that the plot of the logit of mortality rates against age has been steepening as it shifts downwards over time. This phenomenon has been noted by the studies surveyed in Wong-Fupuy and Haberman (2004, secs. 5.2 and 5.3), for example.

- The changes over time in $\kappa_t^{(1)}$ and $\kappa_t^{(2)}$ have been approximately linear and such linear improvements have been noted in previous studies (e.g., Wong-Fupuy and Haberman (2004, sec. 5.1), for example).

- The logit of mortality rates plotted against age have a slight curvature over the 60 to 89 age range that can be modelled using a quadratic function of age. The quadratic period term ($\kappa_t^{(3)}$) is statistically significant, but very small. It is time varying and has generally been increasing over time: having started the data period negative, it ends the period positive. When combined with the quadratic age term, $((x-\bar{x})^2-\hat{\sigma}_x^2)$, its contribution to mortality rate dynamics is highly complex.

- There is a significant cohort effect ($\gamma_{t-x}^{(4)}$) in mortality improvements, although this is more prominent and systematic in the UK than the US.

To a large extent these commonalities are also features of the other models considered.

A good stochastic mortality model must take these features into account when forecasting mortality improvements and confidence intervals around these forecasts. This is important for quantifying longevity risk, for providing benchmarks for longevity-linked financial instruments, for pricing such instruments (Cairns et al. 2006a, b), and for designing longevity risk hedging positions, as suggested in recent studies such as Blake and Burrows (2001), Blake et al. (2006a,b), and Dowd et al (2006).

Acknowledgements

AC wishes to thank Iain Currie for his enlightening discussions of the P-splines model.

Appendices

A The cohort effect

Alternative ways of identifying the cohort effect in the England & Wales data are plotted in Figures 34 to 36. The cohort effect was clear enough in Figure 3 but there is a certain amount of noise between calendar years. The effect of the calendar year has been removed in Figure 34. This seems to smooth things out but also makes the cohort diagonals more prominent.

In Figure 35 we take a rather different approach. Here we take the annual improvements at each age x from year t to $t+1$ and then we deduct from that the improvement predicted by the Lee-Carter model (see Section 4).[20] The results are then smoothed before plotting. The cohort diagonals are still clear although the patterns of colour look somewhat different from Figures 3 and 34.

Finally in Figure 36 we plot the difference between actual death rates and the rates predicted by the Lee-Carter model. These average out at about zero, but the cohort effect is more evident here than in any of the other plots. The 1925 cohort (the white/blue diagonal just above the black diagonal) is more prominent in having actual death rates significantly higher than predicted rates. For other cohorts the effect seems to be less persistent over long periods of time, but blocks of colour still have a prominent diagonal pattern except at the lowest and highest ages.

Each of these plots gives clear evidence for a cohort effect. Out of the eight models that we examine here, six make allowance in some form for a cohort effect; the remaining two models do not. Our analysis includes a rigorous analysis of the quality of fit of the eight models. Amongst other things, therefore, we can investigate how significant the cohort effect is from a statistical perspective.

The equivalent plot for US data is given in Figure 37 and we can see that this looks quite different from the England & Wales data. Above age 60 the cohort effect is less prominent. At younger ages there are very strong deviations from the Lee-Carter model predictions: an issue that is discussed in more detail in the main text.

Black and white representations of the cohort effect are presented in Figures 38 (England & Wales) and 39 (US).

[20] We use the Lee-Carter model as the basis for these calculations because it is the oldest and best known out of the models we consider in this paper.

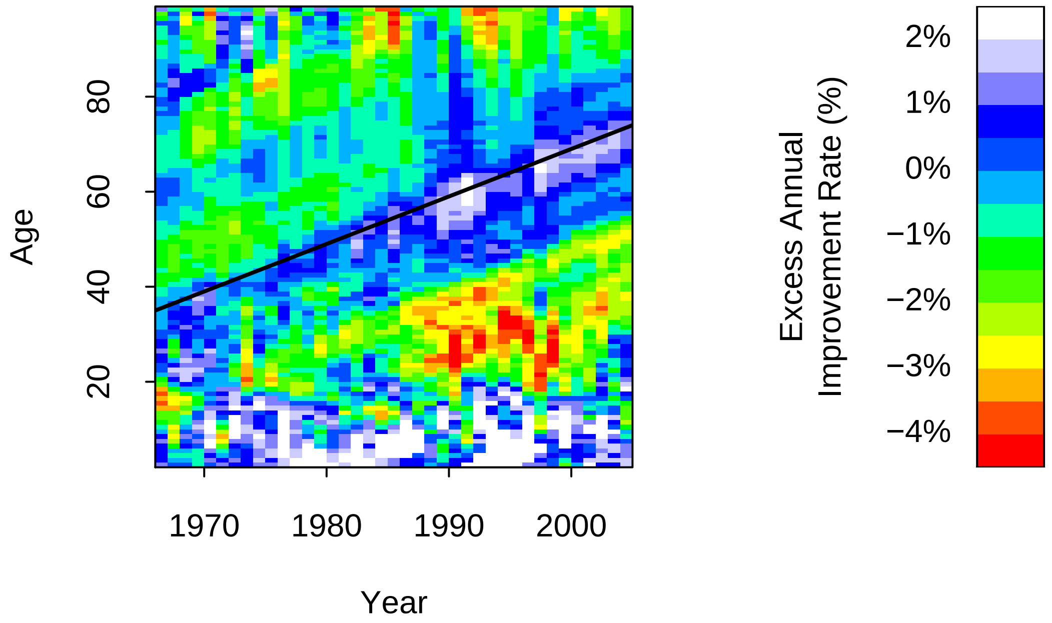


Figure 34: England & Wales data: As Figure 3 except that the average mortality improvement for each year has been deducted.

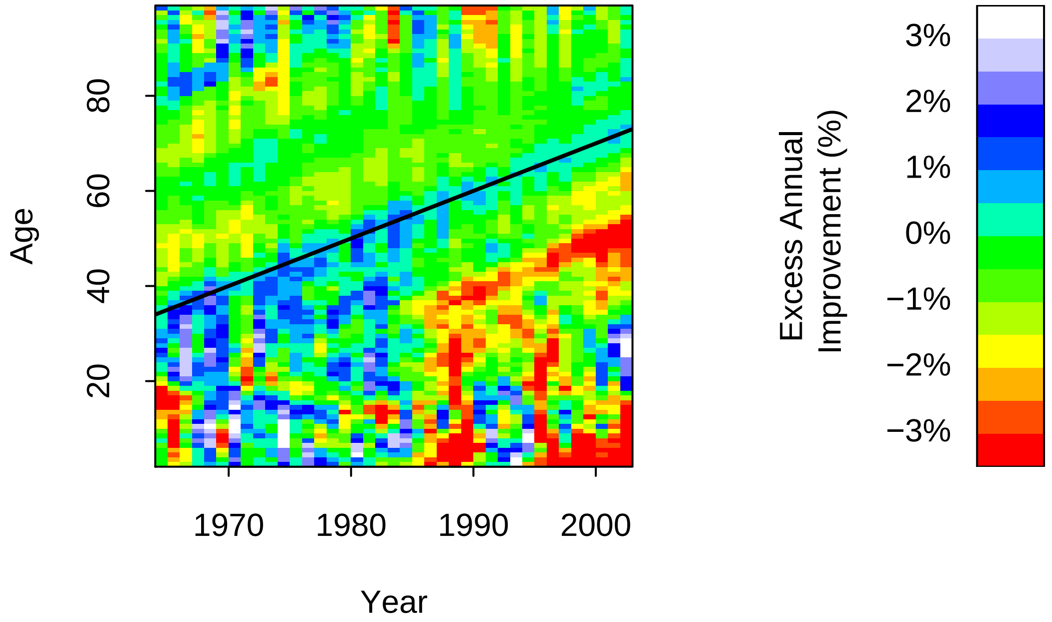


Figure 35: England & Wales data: As Figure 3 except that we plot the smoothed average mortality improvement in excess of the improvement predicted by the Lee-Carter model M1 (see Section 4).

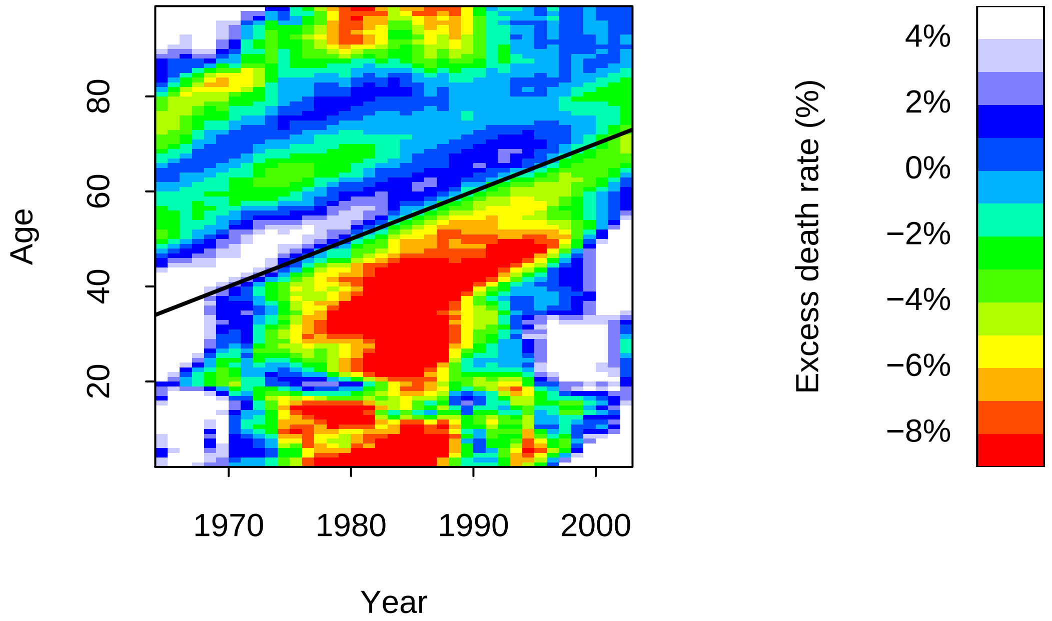


Figure 36: England & Wales data: As Figure 3 except that we plot the smoothed average of the difference between the actual death rates and the death rate predicted by the Lee-Carter model M1. Blue and white cells imply higher death rates than predicted; red cells lower.

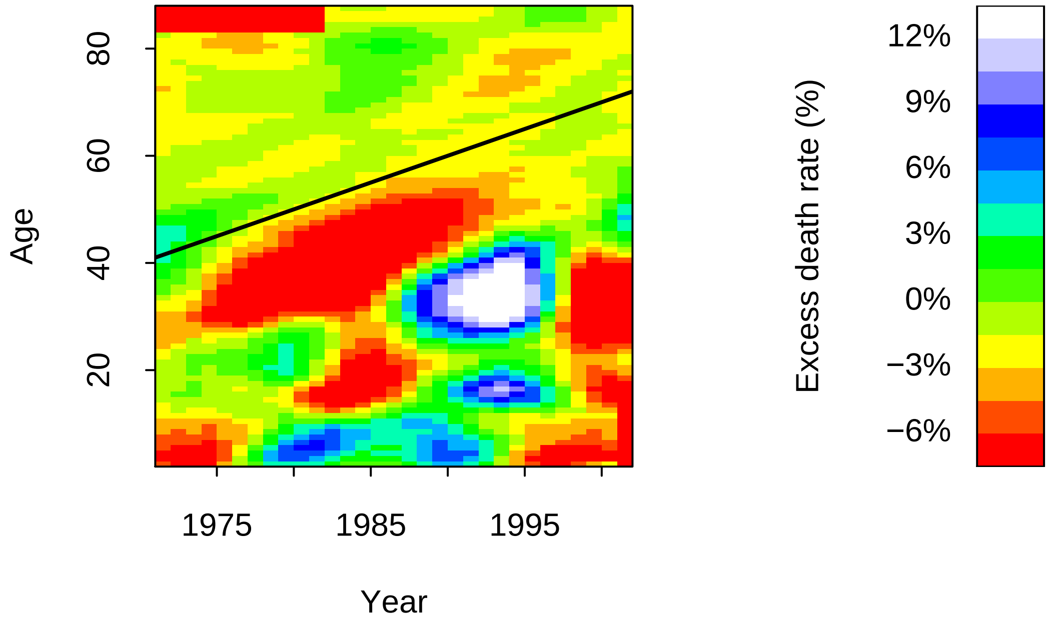


Figure 37: US data: As Figure 3 except that we plot the smoothed average of the difference between the actual death rates and the death rate predicted by the Lee-Carter model M1. Blue and white cells imply higher death rates than predicted; red cells lower.

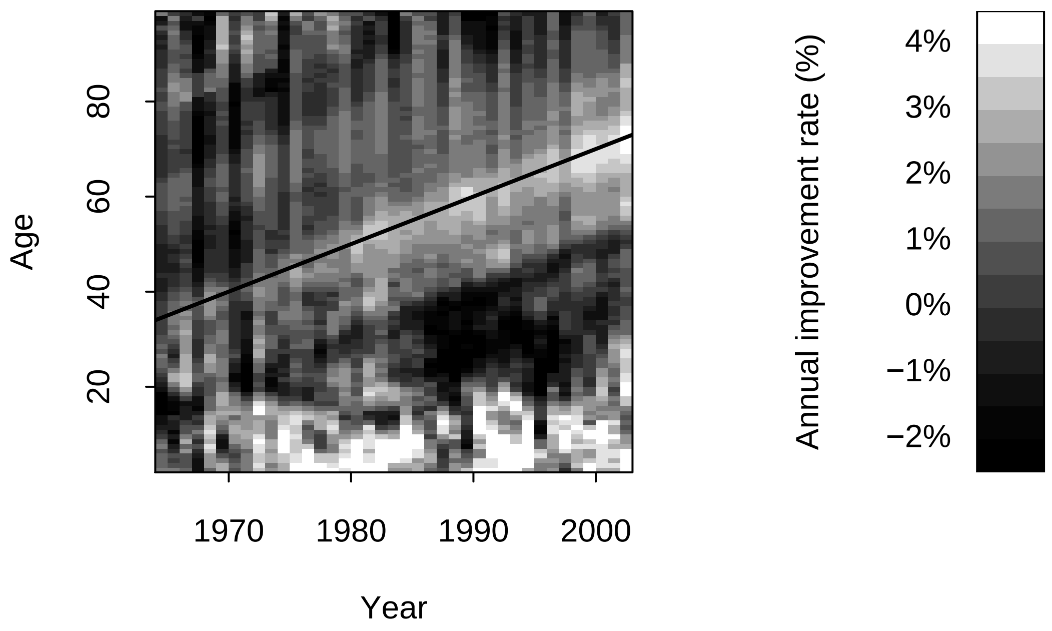


Figure 38: England & Wales data: Improvement rates in mortality by calendar year and age relative to mortality rates at the same age in the previous year. Dark cells imply that mortality is deteriorating; light grey small rates of improvement, and white strong rates of improvement. The black diagonal line follows the progress of the 1930 cohort.

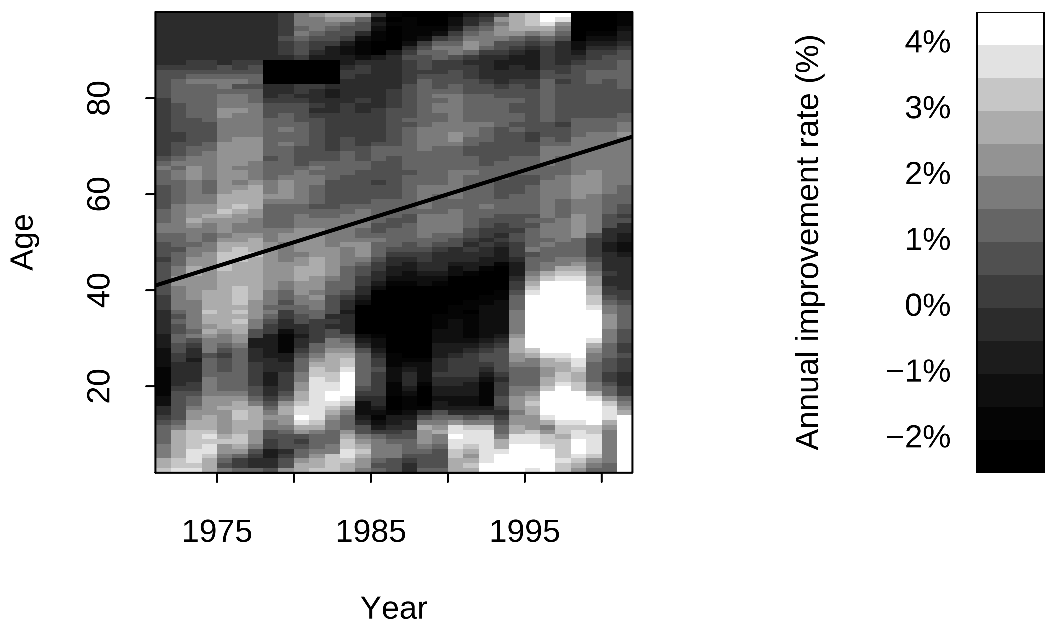


Figure 39: US data: Improvement rates in mortality by calendar year and age relative to mortality rates at the same age in the previous year. Dark cells imply that mortality is deteriorating; light grey small rates of improvement, and white strong rates of improvement. The black diagonal line follows the progress of the 1930 cohort.

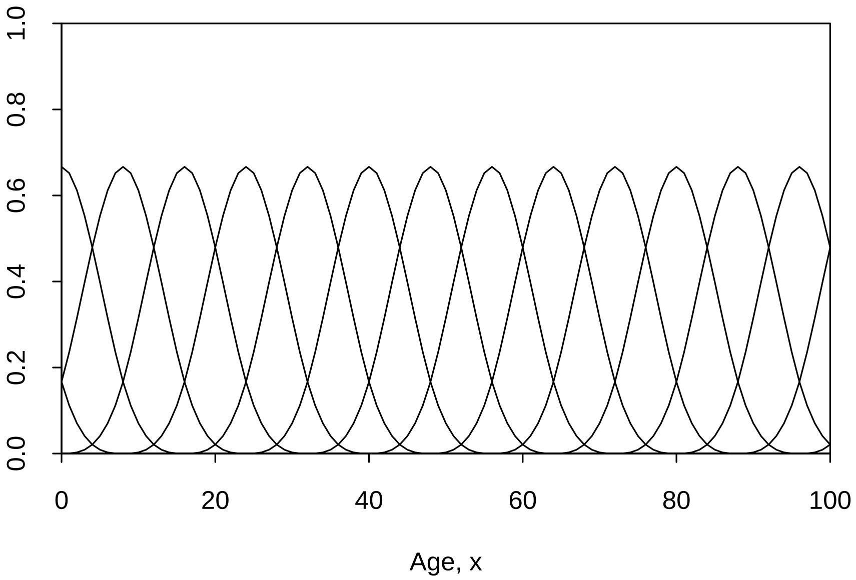


Figure 40: A set of one-dimensional B-splines with a distance of 8 between knots.

B B-splines

We will begin this appendix by looking at B-splines (or Basis-Splines) in one dimension. Suppose that we have data in the range ξ_0 to ξ_n. B-splines are a set of basis functions each of which depends on the placement of a set of "knot" points. For B-splines we require knots outside the range ξ_0 to ξ_n to provide full coverage of this range.

Thus we have knots at $\xi_{-3} < \xi_{-2} < \ldots < \xi_{n+2} < \xi_{n+3}$.

For $k = -3, \ldots, n-1$ spline k is defined as

$$B_k(x) = \sum_{i=k}^{k+4} \prod_{j=k, j \neq i}^{k+4} \frac{1}{(\xi_j - \xi_i)} (x - \xi_i)_+^3$$

where $(x - \xi_i)_+$ is equal to $\max\{x - \xi_i, 0\}$.

Note that $B_k(x) = 0$ outside the range $\xi_k < x < \xi_{k+4}$, and that it has continuous first and second derivatives for all $-\infty < x < \infty$.

Typically the knots for B-splines are evenly spaced. Where this is the case, with $\xi_{k+1} - \xi_k = \delta$ we have for all k:

$$B_k(x) = \delta^{-4} \left\{ \frac{1}{24}(x - \xi_k)_+^3 - \frac{1}{6}(x - \xi_{k+1})_+^3 + \frac{1}{4}(x - \xi_{k+2})_+^3 - \frac{1}{6}(x - \xi_{k+3})_+^3 + \frac{1}{24}(x - \xi_{k+4})_+^3 \right\}.$$

A typical set of one-dimensional B-splines is plotted in Figure 40.

Defining B-splines in 2-dimensions is straightforward. We have two dimensions: age (a), x, and year (y), t. In the age dimension we have knots at $\xi^a_{-3} < \ldots < \xi^a_{m_a+3}$, and in the year dimension we have knots at $\xi^y_{-3} < \ldots < \xi^y_{m_y+3}$. The knots in the

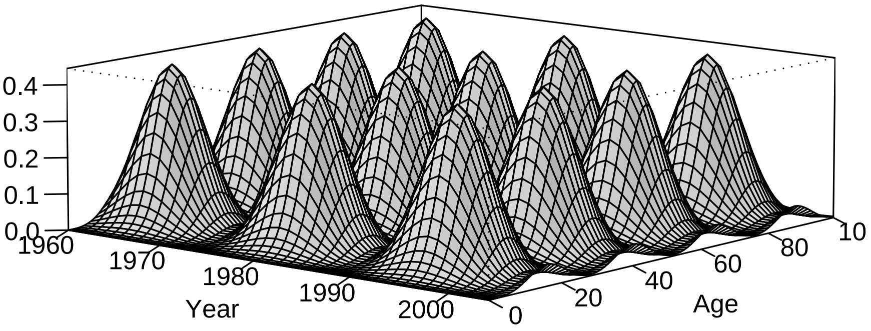


Figure 41: A subset of two-dimensional B-splines. The surface plotted is the sum of every third B-spline in each direction.

year and age dimensions give rise respectively to $B_k^y(t)$ for $k = -3, \ldots, m_y - 1$ and B-splines $B_l^a(x)$ for $l = -3, \ldots, m_a - 1$.

The two-dimensional B-splines are then

$$B_{kl}(x,t) = B_k^y(t) B_l^a(x)$$

for $k = -3, \ldots, m_y - 1$ and $l = -3, \ldots, m_a - 1$. Each B-spline can be thought of as an individual, small hill, and the complete set of B-splines as a set of hills arranged symmetrically. A subset of these hills is plotted in Figure 41.

If we use a large number of knots in the year and age dimensions then we can use B-splines to fit a set of data extremely accurately. However, this tends to result in a rather lumpy fit that might not accurately reflect what we believe to be the underlying reality. Often in cases like this, if the existing data have been overfitted, forecasts become less reliable.

P-splines (or Penalised Splines) is the term we use when B-splines are fitted to set of data where the likelihood or regression function is adjusted by a penalty function. The penalty function smoothes out the lumpiness mentioned above. The inclusion of a penalty with appropriate weight means that we can increase the number of knots without radically altering the smoothness of the fit.

If our fitted function is $f(x,t) = \sum_{k,l} \theta_{kl} B_{kl}(x,t)$ then possible 2nd-order smoothing penalties can be calculated separately in each dimension. Thus smoothing across ages:

$$P^a = \lambda_a \sum_{l=-3}^{m_y-1} \sum_{k=-1}^{m_a-1} (\theta_{k,l} - 2\theta_{k-1,l} + \theta_{k-2,l})^2 = \lambda_a \sum_{l=-3}^{m_y-1} \sum_{k=-1}^{m_a-1} (\Delta_a^2 \theta_{kl})^2$$

where $\Delta_a \phi_{kl} = \phi_{kl} - \phi_{k-1,l}$ is the difference operator on the k (age) dimension.

Similarly, in the year dimension:

$$P^y = \lambda_y \sum_{k=-3}^{m_a-1} \sum_{l=-1}^{m_y-1} (\theta_{k,l} - 2\theta_{k,l-1} + \theta_{k,l-2})^2 = \lambda_y \sum_{k=-3}^{m_a-1} \sum_{l=-1}^{m_y-1} (\Delta_y^2 \theta_{kl})^2$$

where $\Delta_y \phi_{kl} = \phi_{kl} - \phi_{k,l-1}$ is the difference operator on the l (year) dimension.

Finally, if we choose to apply a penalty across cohorts then we have:

$$P^c = \lambda_c \sum_{k=-2}^{m_a-2} \sum_{l=-2}^{m_y-2} (\theta_{k+1,l-1} - 2\theta_{k,l} + \theta_{k-1,l+1})^2.$$

In the analyses that follow we will use age-cohort penalties $P_a + P_c$ as suggested in CMI (2005) working paper 15.

C Simulation model

For projection of the survivor index, $S(t,x)$ we need to take the fitted parameter values illustrated in Figures 7 to 13 and 22 to 28 and use these to develop a stochastic projection model.

For model M5 we use the method described in Cairns, Blake and Dowd (2006) (CBD): thus we fitted a 2-dimensional random-walk model to $(\kappa_t^{(1)}, \kappa_t^{(2)})$ using the the final 21 years of data (that is, 20 observations of the change in $(\kappa_t^{(1)}, \kappa_t^{(2)})$. The form of $\beta_x^{(2)}$ in this paper is different from the original CBD paper so parameter estimates are different.

In the main body of this paper we report on simulation results for M2, M5, M7 and M8. For M2, M7 and M8 the $\beta_x^{(i)}$ effects are fixed. For each of M2, M7 and M8 we adopt the same principles for simulation of the $\kappa_t^{(i)}$ and $\gamma_{t-x}^{(i)}$. For model M7, for example, we take the following approach for England & Wales data from 1961 to 2004.

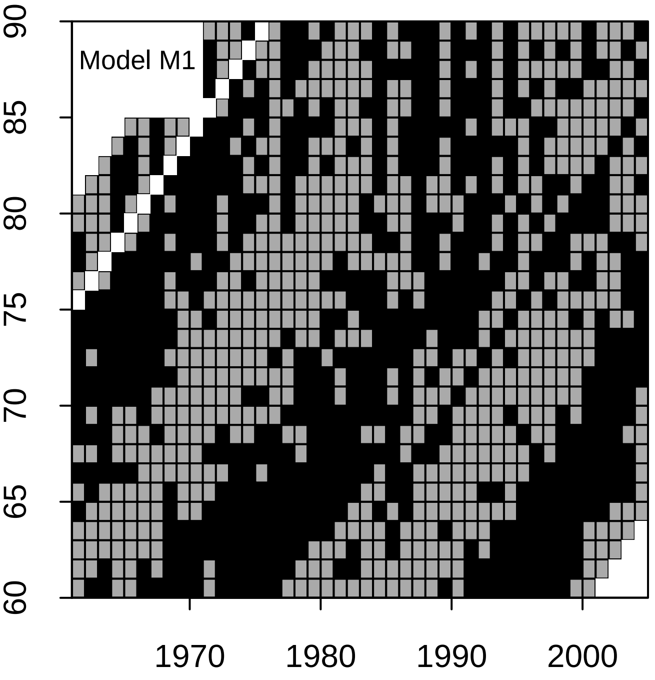
- Fit the $\beta_x^{(i)}$, $\kappa_t^{(i)}$ and $\gamma_{t-x}^{(i)}$ to the full set of data from age 60 to 89.
- Then take $\kappa_t^{(1)}$, $\kappa_t^{(2)}$ and $\kappa_t^{(3)}$ for years 1984 to 2004 inclusive, and fit a 3-dimensional random walk with drift.
- For the cohorts aged 65, 70 and 75 in 2004 we already have an estimate of the cohort effect, $\gamma_{1939}^{(4)}$, $\gamma_{1934}^{(4)}$, and $\gamma_{1929}^{(4)}$, so no model is required for these values.
- For the cohort aged 60 in 2004 we need to project the estimated $\gamma_{t-x}^{(4)}$ series. Our results clearly suggest that a random walk model is inappropriate, but development of a model for $\gamma_{t-x}^{(4)}$ is beyond the scope of this paper. In the context of projecting $S(t,60)$ and calculating an annuity value it suffices for us to specify a single value for $\gamma_{1944}^{(4)}$. Based on the historical development of $\gamma_{t-x}^{(4)}$ we try out two values for $\gamma_{1944}^{(4)}$ (one high and one low) to cover what we feel is the likely range of values that might be taken by $\gamma_{1944}^{(4)}$.

Once we have our simulation model for $S(t,x)$ we can calculate annuity values according to the formula

$$a_x(2004) = \sum_{t=1}^{90-x} 1.04^{-t} E[S(t,x)].$$

D Standardised residuals: black and white

Black and white versions of Figures 17 to 20 are presented in Figures 42 to 45. Figures 42 to 45



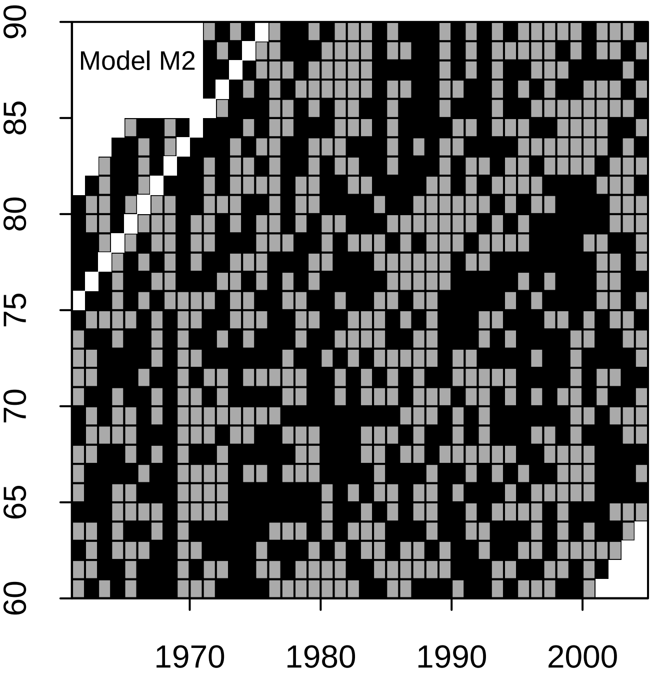
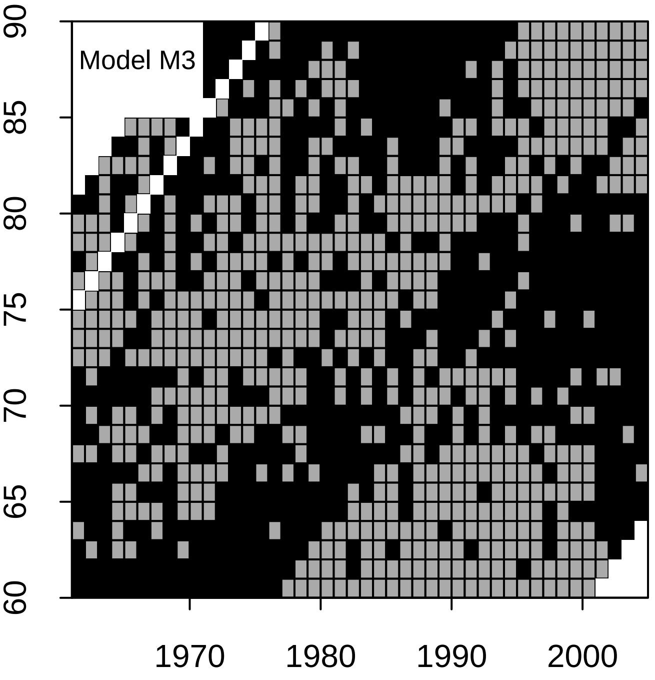


Figure 42: England & Wales males. Standardised residuals $Z(t, x)$ for Models M1 (top) and M2 (bottom). Grey cells mean $Z(t, x) > 0$, black means $Z(t, x) < 0$, white means the cell was excluded from the analysis.



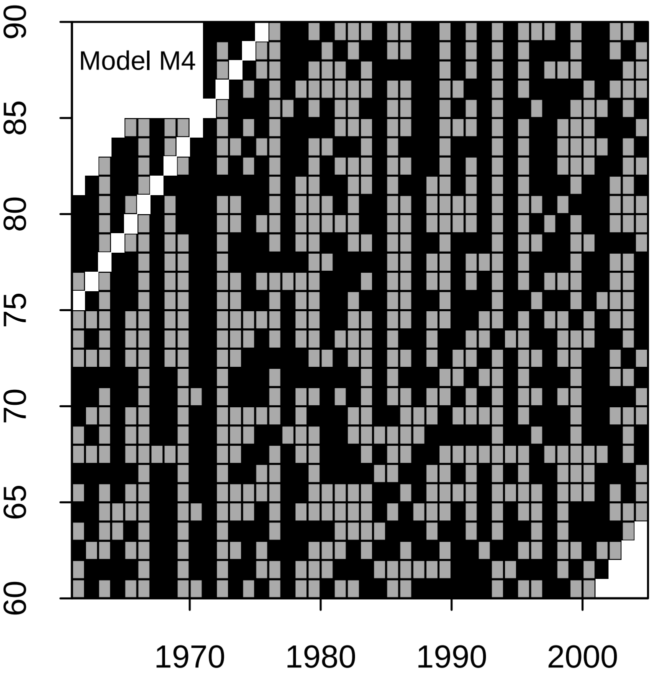
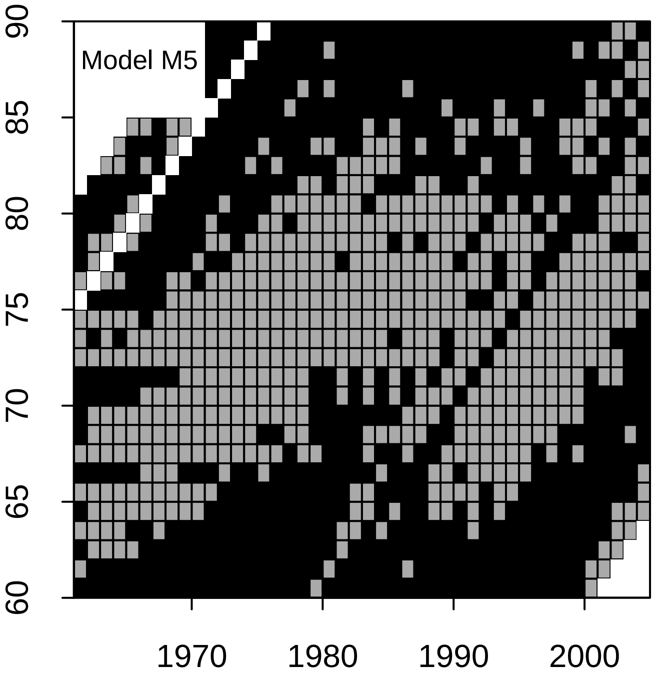


Figure 43: England & Wales males. Standardised residuals $Z(t, x)$ for Models M3 (top) and M4 (bottom). Grey cells mean $Z(t, x) > 0$, black means $Z(t, x) < 0$, white means the cell was excluded from the analysis.



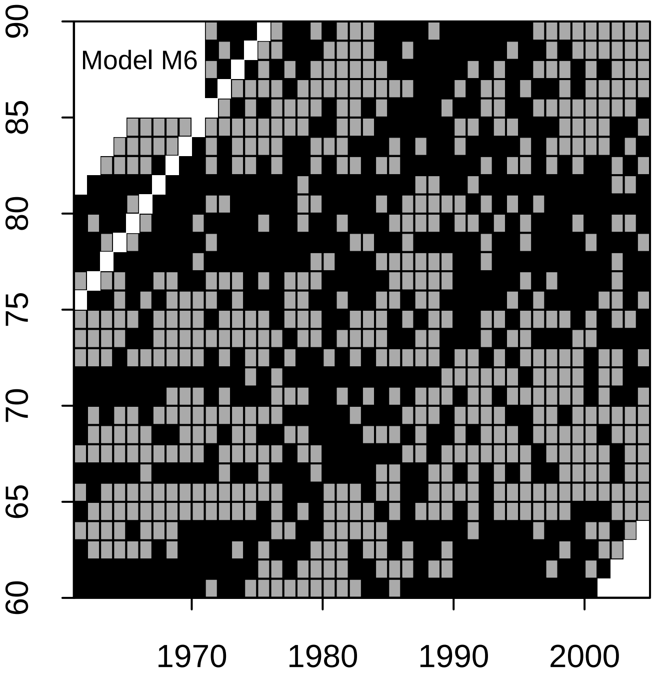
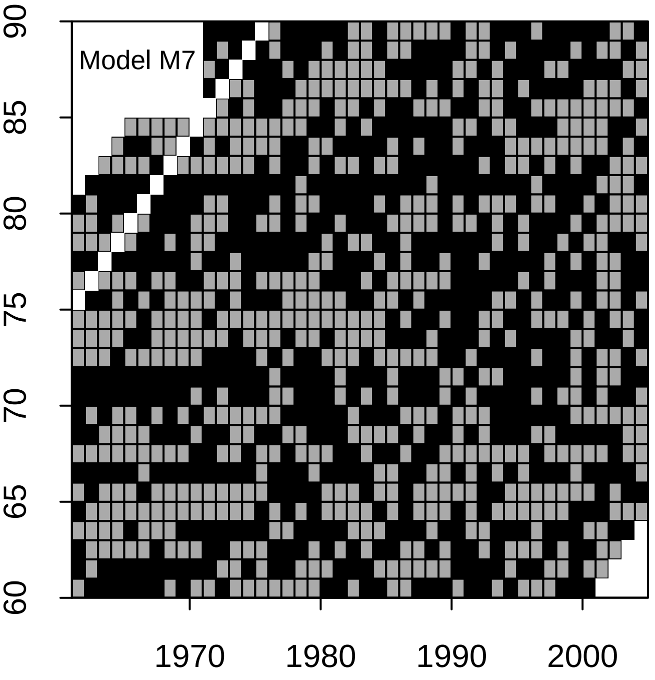


Figure 44: England & Wales males. Standardised residuals $Z(t, x)$ for Models M5 (top) and M6 (bottom). Grey cells mean $Z(t, x) > 0$, black means $Z(t, x) < 0$, white means the cell was excluded from the analysis.



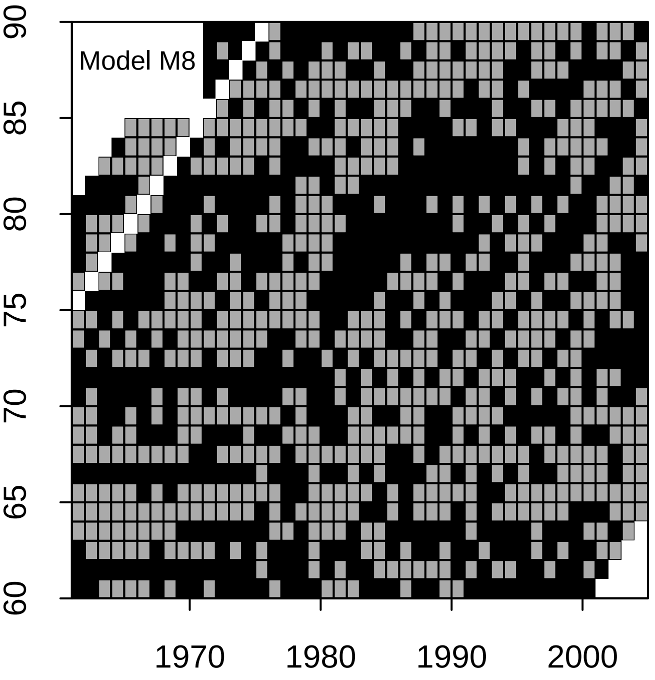


Figure 45: England & Wales males. Standardised residuals $Z(t, x)$ for Models M7 (top) and M8 (bottom). Grey cells mean $Z(t, x) > 0$, black means $Z(t, x) < 0$, white means the cell was excluded from the analysis.

E References and further reading

Blake, D., and Burrows, W. (2001) "Survivor bonds: Helping to hedge mortality risk," *Journal of Risk and Insurance,* 68: 339-348.

Blake, D., Cairns, A.J.G., and Dowd, K. (2006a) "Living with mortality: Longevity bonds and other mortality-linked securities," *British Actuarial Journal*, 12: 153-197.

Blake, D., Cairns, A.J.G., Dowd, K., and MacMinn, R. (2006b) "Longevity Bonds: Financial Engineering, Valuation & Hedging," *Journal of Risk and Insurance,* 73: 647-72.

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. (2000) "A discussion of parameter and model uncertainty in insurance," *Insurance: Mathematics and Economics,* 27: 313-330.

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.

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

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

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

Currie, I.D. (2006) "Smoothing and forecasting mortality rates with P-splines," Talk given at the Institute of Actuaries, June 2006.
See http://www.ma.hw.ac.uk/∼iain/research/talks.html

Dowd, K., Blake, D., Cairns, A.J.G., and Dawson, P. (2006) "Survivor Swaps," *Journal of Risk and Insurance,* 73: 1-17.

Hayashi, F. (2000) *Econometrics.* Princeton University Press: Princeton.

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

Perks, W. (1932) "On some experiments in the graduation of mortality statistics," *Journal of the Institute of Actuaries,* 63: 12-57.

Renshaw, A.E., and Haberman, S. (2003) "Lee-Carter mortality forecasting with age-specific enhancement," *Insurance: Mathematics and Economics,* 33: 255-272.

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

nomics, 38: 556-570.

Richards, S.J., Kirkby, J.G., and Currie, I.D. (2006) "The importance of year of birth in two-dimensional mortality data," *British Actuarial Journal*, 12: 5-38.

Willets, R.C. (1999) "Mortality in the next millennium," Paper presented to the Staple Inn Actuarial Society.

Willets, R.C. (2004) "The cohort effect: Insights and explanations," *British Actuarial Journal,* 10: 833-877.

Wong-Fupuy, C., and Haberman, S. (2004) "Projecting mortality trends: Recent developments in the UK and the US," *North American Actuarial Journal,* 8: 56-83.