Journal of the Royal Statistical Society Series C: Applied Statistics, 2023, **00**, 1–21
https://doi.org/10.1093/jrsssc/qlad021

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series C
Applied Statistics

Original Article

Bayesian model comparison for mortality forecasting

Jackie S.T. Wong[1], Jonathan J. Forster[2] and Peter W.F. Smith[3]

[1]Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, UK
[2]Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[3]Department of Social Statistics and Demography, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

Address for correspondence: Jackie S.T. Wong, Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK. Email: jw19203@essex.ac.uk

Abstract

Stochastic models are appealing for mortality forecasting in their ability to generate intervals that quantify uncertainties underlying the forecasts. We present a fully Bayesian implementation of the age-period-cohort-improvement (APCI) model with overdispersion, which is compared with the Lee–Carter model with cohorts. We show that naive prior specification can yield misleading inferences, where we propose Laplace prior as an elegant solution. We also perform model averaging to incorporate model uncertainty. Our findings indicate that the APCI model offers better fit and forecast for England and Wales data spanning 1961–2002. Our approach also allows coherent inclusion of multiple sources of uncertainty, producing well-calibrated probabilistic intervals.

Keywords: age-period-cohort-improvement (APCI), Laplace prior distribution, Lee–Carter, model averaging, mortality forecasting, overdispersion

1 Introduction

Mortality forecasting is crucial in the planning of social securities and in various life-related industries. A diversity of methodologies has been developed with the common aim to capture and project mortality trends, accompanied by well-calibrated uncertainty bands. For a comprehensive overview of mortality forecasting approaches, see Tabeau et al. (2001) or Booth and Tickle (2008). Particularly well known is the Lee–Carter (LC) mortality model developed by Lee and Carter (1992), which has been the backbone of many of the existing stochastic mortality forecasting approaches. Various modifications that emerged thereafter seek to improve particular aspects of the LC approach. Brouhns et al. (2002), among others, proposed a Poisson-equivalent version of the LC model which better describes the distribution of the number of deaths. Renshaw and Haberman (2003) enhanced the LC model using an extra bilinear term to explain a greater proportion of the variation in the data, which was then refined by Renshaw and Haberman (2006) to incorporate cohort effects. Czado et al. (2005) implemented a fully integrated Bayesian approach for fitting the Poisson LC model, which was further developed by Wong et al. (2018) to account for overdispersion.

The LC family of mortality models has been widely used, but also encountered some criticisms (see, for example, Cairns et al., 2007; Girosi & King, 2008). In this paper, we explore the alternative model proposed by Continuous Mortality Investigation Bureau (2016a), the age-period-cohort-improvement (APCI) model. The APCI model was originally developed as a deterministic targeting approach to projecting mortality rates. This model is structurally simple and has been found to possess numerous advantages (see Continuous Mortality Investigation Bureau, 2016a for more details). Continuous Mortality Investigation Bureau (2016b) illustrates the use of

Received: August 27, 2021. **Revised:** October 31, 2022. **Accepted:** February 1, 2023

the APCI model (together with the targeting method) on England and Wales male mortality data for ages 20–100, allowing implicitly for the presence of overdispersion using smoothing hyperparameters. Richards et al. (2019) present a stochastic implementation of the APCI model (without overdispersion) for UK male mortality data for ages 50–104 and demonstrate that the APCI model offers the best fit among the competing models. Hilton et al. (2019) present a Bayesian approach (with vague priors) to forecast mortality using a generalized additive model with the APCI structure for majority of the age range, except infant and old age mortality.

In this paper, we present a Bayesian implementation of the APCI model, explicitly accounting for overdispersion using additional parameters. We also show that this model is equally applicable to the entire age range even though it is more commonly used for ages above 20. The advantages of Bayesian mortality forecasting have been discussed in several articles (Czado et al., 2005; Pedroza, 2006; Wong et al., 2018, to name a few). The primary advantage being the coherent inclusion of various sources of uncertainties for better calibration of data signals and errors, which also provides a natural framework for model comparison using posterior model probabilities. The significance of incorporating cohort effects and overdispersion have been separately discussed (for the former, see Börger & Aleksic, 2014; Renshaw & Haberman, 2006; while for the latter, see Delwarde et al., 2007; Li et al., 2009; Wong et al., 2018). We illustrate in this paper that accounting for both cohort effects and overdispersion under a Bayesian paradigm leads to pronounced improvement in the calibration between data signals and errors.

Before fully implementing the Bayesian APCI model, we focus on analysing the impact of prior specification. This is vital because naive prior specification using conventional diffuse priors can yield misleading inference, particularly for model comparison using Bayesian quantities. To motivate a better choice of prior distributions, we initially ignore the cohort components and perform a pairwise comparison to establish parameter correspondence relationships between the resulting age-period-improvement (API) and LC models. The aim is to tune the hyperparameters so that the information from the priors is compatible under both models. We show through some theoretical analysis that this can be achieved using Laplace priors on relevant parameters of the API model.

The interest is then to compare the APCI model with the Bayesian LC model with cohorts (cohort-extended version of the approach by Wong et al., 2018). Similar comparison has been performed by Richards et al. (2019) within a frequentist paradigm, where they focused on the age range 50–104 and the crucial cohort component was not included for the LC model. Cairns et al. (2007) also compared several stochastic mortality models quantitatively based on non-Bayesian criteria. Barigou et al. (2022) illustrated the use of some model averaging techniques (hence, model comparison implicitly), including Bayesian model averaging (BMA) by marginal likelihoods, stacking, and pseudo-BMA methods, to combine several stochastic mortality models for mortality forecasting. However, they did not consider the APCI with overdispersion model and did not include the entire age range. Their study also did not focus on the impact of prior specification on Bayesian model comparison, where weakly informative priors were used throughout. In this paper, we explicitly focus on that impact in the context of mortality forecasting.

Using the properly tuned priors, we conclude that our results agree to a reasonable extent with Richards et al. (2019) that the APCI model fits the mortality data better, even after including the entire age range and cohort effect for LC model. We also consider the use of two time series models for projecting the time trend of mortality. Finally, the BMA technique is applied to combine several models rather than selecting the best one. This produces well-calibrated probabilistic mortality forecasts that incorporate model uncertainty, on top of other sources of uncertainty such as parameter uncertainty (in the form of priors) and forecast uncertainty.

We begin this paper by introducing the general formulation of our model with two specification for mortality rates (initially without cohort components) in Section 2. The relationships between the parameters of the two models are then illustrated. In Section 3, we present an extensive analysis of prior specification, with the aim of providing compatible prior information for the competing models. We tune both the form of the prior distribution (using Laplace priors for certain parameters of the API model) and the corresponding hyperparameters so that they lead to compatible implied priors on mortality rates. Section 4 briefly discusses related computations involved for posterior estimation and Bayesian model comparison. In Section 5, cohort parameters are introduced to form the APCI and LC with cohort models, which are then estimated and compared using the priors previously formulated. Model averaging and some numerical results are also presented.

1.1 Data and notation

We denote the observed number of deaths of age group x in year t by d_{xt}, where $x = x_1, x_2, \ldots, x_A$ and $t = t_1, t_2, \ldots, t_T$ represent a set of A different age groups and T years, respectively. Also we define e_{xt} and μ_{xt} to be the corresponding central exposed to risk and central mortality rate of age group x in year t.

The data chosen for illustrative purposes are the female death data and the corresponding exposures for England and Wales, extracted from the Human Mortality Database (HMD). They are classified by single year of age from 0 to 99, and years ranging from 1961 to 2002. Hence, we have $\{x_1, \ldots, x_A\} = \{0, \ldots, 99\}$ and $\{t_1, \ldots, t_T\} = \{1961, \ldots, 2002\}$ with $A = 100$ and $T = 42$. We intentionally held back the data for years 2003–2016 as a validation data set.

2 Models

Let D_{xt} be the random variable denoting the number of deaths age x year t. As in Brouhns et al. (2002), we impose a conditional Poisson distribution for D_{xt}, i.e.,

$$D_{xt} \mid \mu_{xt} \sim \text{Poisson}(e_{xt}\mu_{xt}) \tag{1}$$

The rate model is then defined as

$$\log \mu_{xt} = M_{xt} + \log v_{xt} \tag{2}$$

where M_{xt} is to be specified and v_{xt} characterize overdispersion. We incorporate overdispersion into the model for μ_{xt} to capture extra variabilities due to heterogeneity (which ensures that marginally $\mathbb{E}[D_{xt}] \neq \text{Var}[D_{xt}]$). As discussed in Wong et al. (2018), this prevents over-fitting and also avoid over-optimistic forecast intervals. Following the recommendation by Wong et al. (2018), we assume that

$$v_{xt} \mid \phi \overset{\text{ind}}{\sim} \text{Gamma}(\phi, \phi) \tag{3}$$

mainly for computational ease as the latent variable can be integrated out to give

$$D_{xt} \mid \phi \sim \text{Neg} - \text{Bin}\left(\phi, \frac{\phi}{e_{xt}\exp(M_{xt}) + \phi}\right) \tag{4}$$

Hence, we have

$$\mathbb{E}[D_{xt}] = e_{xt}\exp(M_{xt}) \quad \text{and} \quad \text{Var}[D_{xt}] = \mathbb{E}[D_{xt}] \times [1 + \mathbb{E}[D_{xt}]/\phi] > \mathbb{E}[D_{xt}] \tag{5}$$

where $1/\phi$ represents the magnitude of overdispersion.

2.1 Two rate models

We are first interested in comparing two models for μ_{xt}. The first one is the well-known LC model (Lee & Carter, 1992) given by

$$M_{xt} = \alpha_x + \beta_x \kappa_t \tag{6}$$

where α_x, β_x, κ_t are model parameters. For model identifiability, the constraints $\sum_x \beta_x = 1$ and $\sum_t \kappa_t = 0$ are adopted. We will refer this model as the negative-binomial LC model.

The second rate model is given by

$$M_{xt} = \alpha_x + \beta_x t + \kappa_t \tag{7}$$

where $\alpha_x, \beta_x, \kappa_t$ are model parameters. Here the constraints,

$$\sum_t \kappa_t = \sum_t t\kappa_t = 0 \tag{8}$$

are adopted for identifiability. This model has an advantage of being structurally simpler, and more easily interpretable, because μ_{xt} are log-linear with respect to the model parameters, in contrast to the log-bilinear specification in the LC model, which can cause computational instability. Also notice that

$$\log\left(\frac{\mu_{xt}}{\mu_{x\ t-1}}\right) = \beta_x + \kappa_t - \kappa_{t-1} + \log\left(\frac{v_{xt}}{v_{x\ t-1}}\right) = \beta_x + \kappa'_t + \log v'_{xt}$$

meaning the mortality improvement is essentially an age-period model. Hence, we will call this second model the negative-binomial API model. Note that the inclusion of cohort effect would lead to the model proposed by Continuous Mortality Investigation Bureau (2016a), the Bayesian implementation of which will be detailed in Section 5.

2.2 Projection models

The time-varying parameter κ_t in equation (6) typically demonstrates a linearly decreasing trend, where a random walk with drift has been empirically observed to provide adequate fit (see Tuljapurkar et al., 2000). Following the formulation of Czado et al. (2005), we set

$$\left.\begin{array}{ll} \kappa_t - \eta_t = \rho(\kappa_{t-1} - \eta_{t-1}) + \epsilon_t & \text{for } t = 2, 3, \ldots, T \\ \kappa_1 = \eta_1 + \epsilon_1 & \end{array}\right\} \tag{9}$$

where $\eta_t = \psi_1 + \psi_2 t$ denotes the linear drift and $\epsilon_t \overset{\text{ind}}{\sim} N(0, \sigma_\kappa^2)$ are random innovations. Then, depending on the value of ρ, this forms either a first-order autoregressive (AR(1)) model ($|\rho| < 1$) or a random walk model ($\rho = 1$) for κ_t. Equivalently, this model could be expressed multivariately (together with the constraints) as

$$\left.\begin{array}{l} \boldsymbol{\kappa}_{-1} \,|\, \rho, \boldsymbol{\psi}, \sigma_\kappa^2 \sim N(\boldsymbol{\mu}_\kappa, \sigma_\kappa^2 \boldsymbol{V}) \\ \kappa_1 = -\displaystyle\sum_{t=2}^{T} \kappa_t \end{array}\right\} \tag{10}$$

where $\boldsymbol{\kappa}_{-1} = (\kappa_2, \ldots, \kappa_T)^\top$, $\boldsymbol{\mu}_\kappa = (\boldsymbol{I}_{T-1} - \boldsymbol{B}_{21}\boldsymbol{B}_{11}^{-1}\boldsymbol{1}_{T-1}^\top) \times \boldsymbol{Y}_{-1}\boldsymbol{\psi}$, $\boldsymbol{Y}_{-1} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 2 & 3 & \cdots & T \end{pmatrix}^\top$, $\boldsymbol{1}_n$ and $\boldsymbol{I}_n$ denote, respectively, a length$-n$ vector of ones and an identity matrix of dimension $n \times n$, $\boldsymbol{\psi} = (\psi_1, \psi_2)^\top$, and $\boldsymbol{V}$ is chosen such that the marginal distribution implied for κ_t is equivalent to that of (9) (see Online Supplementary Material, Appendix A for details).

According to Richards et al. (2019), κ_t of the API model behaves like that of the LC model but with linear drift extracted, and hence, appear as (driftless) residuals. Therefore, an appropriate model for κ_t of the API model is the model in (9) without the linear drift, that is

$$\left.\begin{array}{ll} \kappa_t = \rho\kappa_{t-1} + \epsilon_t & \text{for } t = 2, 3, \ldots, T \\ \kappa_1 = \epsilon_1 & \end{array}\right\} \tag{11}$$

where $\epsilon_t \overset{\text{ind}}{\sim} N(0, \sigma_\kappa^2)$. Applying the constraints $\sum_t \kappa_t = \sum_t t\kappa_t = 0$ on the model in (11), we have (multivariately)

$$\left.\begin{array}{l} \boldsymbol{\kappa}_{-\{1,2\}} \,|\, \rho, \sigma_\kappa^2 \sim N(\boldsymbol{0}, \sigma_\kappa^2 \boldsymbol{W}) \\ \kappa_1 = \displaystyle\sum_{t=3}^{T} (t-2)\kappa_t \\ \kappa_2 = -\displaystyle\sum_{t=3}^{T} (t-1)\kappa_t \end{array}\right\} \tag{12}$$

where $\boldsymbol{\kappa}_{-\{1,2\}} = (\kappa_3, \ldots, \kappa_T)^{\top}$, and $\boldsymbol{W}$ is chosen such that the marginal distribution of κ_t as in (11) is maintained (see Online Supplementary Material, Appendix A).

2.3 Relationships between the LC and API model parameters

Here, we establish the relationships between the negative-binomial LC and API model parameters by performing a term-by-term comparison. This is useful for our prior specification later in Section 3. Where necessary, we use superscripts $^{\mathrm{LC}}$ and $^{\mathrm{API}}$ to denote parametrization under LC and API models, respectively.

For simplicity, we ignore the constraints of the autoregressive integrated moving average (ARIMA) models for κ_t^{LC} and κ_t^{API} momentarily, and consider their marginal distributions, given respectively by

$$\left.\begin{aligned} \kappa_t^{\mathrm{LC}} &= \psi_1^{\mathrm{LC}} + \psi_2^{\mathrm{LC}} t + \epsilon_t'^{\;\mathrm{LC}} \\ \kappa_t^{\mathrm{API}} &= \epsilon_t'^{\;\mathrm{API}} \end{aligned}\right\} \tag{13}$$

where

$$\epsilon_t'^{\;\mathrm{LC}} \sim N\left(0, \frac{(\sigma_\kappa^{\mathrm{LC}})^2}{1-(\rho^{\mathrm{LC}})^2}\right) \quad \text{and} \quad \epsilon_t'^{\;\mathrm{API}} \sim N\left(0, \frac{(\sigma_\kappa^{\mathrm{API}})^2}{1-(\rho^{\mathrm{API}})^2}\right)$$

are Gaussian errors. Hence, we can write

$$\left.\begin{aligned} \log \mu_{xt}^{\mathrm{LC}} &= \alpha_x^{\mathrm{LC}} + \beta_x^{\mathrm{LC}}\psi_1^{\mathrm{LC}} + \beta_x^{\mathrm{LC}}\psi_2^{\mathrm{LC}} t + \beta_x^{\mathrm{LC}}\epsilon_t'^{\;\mathrm{LC}} + \log v_{xt}^{\mathrm{LC}} \\ \log \mu_{xt}^{\mathrm{API}} &= \alpha_x^{\mathrm{API}} + \beta_x^{\mathrm{API}} t + \epsilon_t'^{\;\mathrm{API}} + \log v_{xt}^{\mathrm{API}} \end{aligned}\right\} \tag{14}$$

The following parameter correspondence can be established:

$$\alpha_x^{\mathrm{LC}} + \beta_x^{\mathrm{LC}}\psi_1^{\mathrm{LC}} \longleftrightarrow \alpha_x^{\mathrm{API}}, \quad \beta_x^{\mathrm{LC}}\psi_2^{\mathrm{LC}} \longleftrightarrow \beta_x^{\mathrm{API}}, \quad \beta_x^{\mathrm{LC}}\epsilon_t'^{\;\mathrm{LC}} \longleftrightarrow \epsilon_t'^{\;\mathrm{API}}, \quad v_{xt}^{\mathrm{LC}} \longleftrightarrow v_{xt}^{\mathrm{API}} \tag{15}$$

Expressions in (15) reiterate that β_x^{LC} and β_x^{API} are of reverse signs, since ψ_2^{LC} is the slope of the decreasing linear drift of κ_t^{LC}. Note that the correspondence relationship in (15) is not unique, but is chosen such that pairwise comparison is sensible in terms of parameter interpretation.

3 Prior specification

Although conventional diffuse priors are typically used when a data-driven inference is of interest, caution needs to be exercised when model comparison using Bayesian quantities is to be undertaken (Weakliem, 1999). Specifically, using overly diffuse priors has a higher tendency to induce Lindley–Bartlett Paradox (Bartlett, 1957). This may lead to unreliable model comparison procedure by inherently favouring one of the models. In particular, we demonstrate how a naive prior specification produces results that misleadingly favour the LC model over its counterpart in Section 3.1. We remedy this issue by first modifying the hyperparameters such that the implied priors for $\log \mu_{xt}$ are plausible in Section 3.2, which enables a clearer picture as to why the LC model is inherently favoured. Second, we propose the use of Laplace priors for relevant parameters, ensuring compatibility in terms of the prior information specified for both models.

3.1 Naive prior specification

Vague priors implemented in Wong et al. (2018) are first used. Briefly, posterior samples are generated using Markov chain Monte Carlo (MCMC) methods with a burn-in phase of 1,000 iterations and a thinning by 100, resulting in samples of size 10,000 for both models. Bridge sampling (Meng & Wong, 1996) is then applied to obtain estimated log marginal likelihoods of −23729.06 and −23769.77, respectively, for the LC and API models. This result is in contrast to that suggested by the Bayesian information criterion (BIC), where the API model (BIC: 47169.48) is found to be superior to the LC model (BIC: 47217.47). While it is mathematically plausible for the conclusion from Bayes factors to disagree with that of BIC, it potentially indicates that the diffuse priors have a strong influence on the comparison procedure.

3.2 Compatible prior specification

To understand the influence of the prior specification, we examine in detail the prior distribution implied for μ_{xt}, as μ_{xt} have identical interpretation under both models (and also because μ_{xt} are generally the determining factor for mortality projections). This is achieved by monitoring the implied prior distributions for μ_{xt}, using their kernel density estimates based on samples generated from the associated prior distributions. For instance, the generation of prior μ_{xt} for the API model proceeds in two steps:

1. Generate κ_t^{API} from equation (12), where samples from the priors of ρ^{API} and $(\sigma_\kappa^{\mathrm{API}})^2$ are substituted appropriately.
2. Generate μ_{xt} from $\mu_{xt} \sim \mathrm{Gamma}(\phi^{\mathrm{API}}, \phi^{\mathrm{API}}/\exp(\alpha_x^{\mathrm{API}} + \beta_x^{\mathrm{API}}\kappa_t^{\mathrm{API}}))$, where κ_t^{API} are from step 1 and $(\alpha_x^{\mathrm{API}}, \beta_x^{\mathrm{API}}, \phi^{\mathrm{API}})$ are samples from their corresponding priors.

The implied prior distributions for $\log\mu_{xt}$ under the naive prior specification (see Section 3.1) are presented in Figure 1. Clearly, the implied priors of $\log\mu_{xt}$ under both models are so diffuse that nonnegligible probabilities are given to unrealistic values of mortality rates. Hence, we first aim to tune the prior specification such that the implied priors for $\log\mu_{xt}$ have most of their densities within a reasonable range in the sense of realistic mortality rate, typically around $(-15, 0)$. This imposition of prior mortality knowledge enables a better visualization to assess properly the impact of prior specification on the Bayes factor.

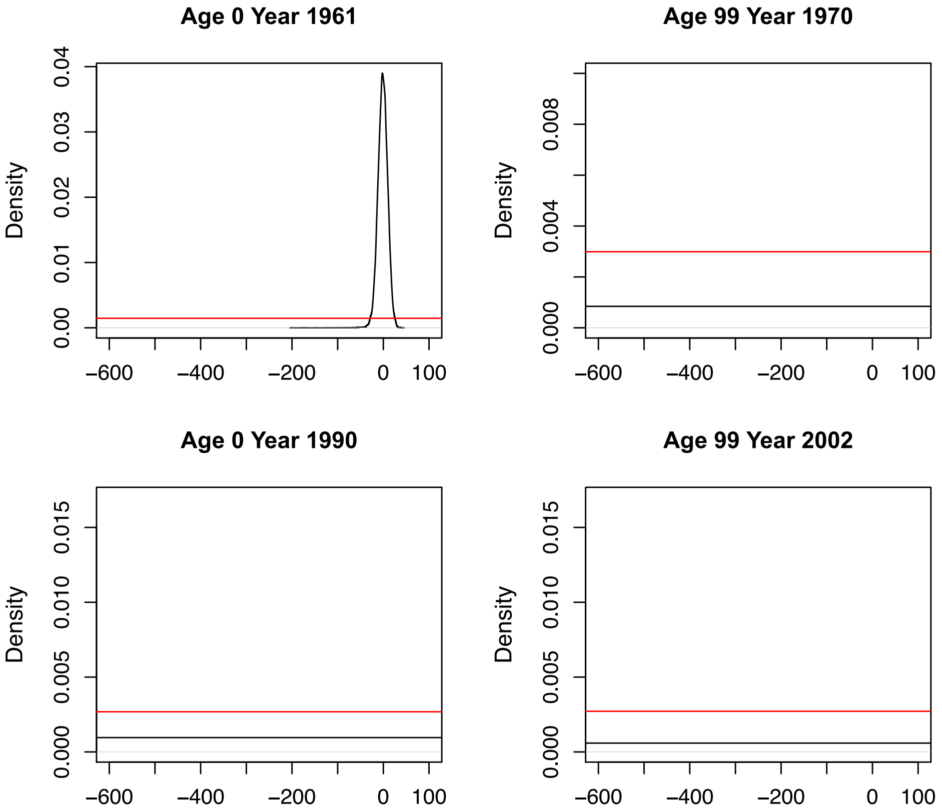

Figure 1. A plot of kernel density estimates of the implied priors of several chosen $\log\mu_{xt}$ for the Lee–Carter (black) and age-period-improvement (red) models, using the naive specification.

The priors chosen for the LC model are as follows:

$$\left.\begin{aligned}
\alpha_x^{\mathrm{LC}} &\overset{\text{ind}}{\sim} N(-5, 4) \\
\boldsymbol{\beta}_{-1}^{\mathrm{LC}} &\sim N\left(\frac{1}{A}\times\mathbf{1}_{A-1},\, 0.005\times\left(\boldsymbol{I}_{A-1}-\frac{1}{A}\boldsymbol{J}_{A-1}\right)\right) \\
(\sigma_\beta^{\mathrm{LC}})^2 &= 0.005 \\
\frac{\rho^{\mathrm{LC}}+1}{2} &\sim \mathrm{Beta}(3, 2) \quad \text{where } \rho^{\mathrm{LC}} \in (-1, 1) \\
(\sigma_\kappa^{\mathrm{LC}})^{-2} &\sim \mathrm{Gamma}(1, 0.0001) \\
\boldsymbol{\psi}^{\mathrm{LC}} &\sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2000 & 0\\ 0 & 2\end{pmatrix}\right) \\
\phi^{\mathrm{LC}} &\sim \mathrm{Gamma}(25, 0.05)
\end{aligned}\right\}\tag{16}$$

where $\boldsymbol{\beta}_{-1}^{\mathrm{LC}} = (\beta_2^{\mathrm{LC}}, \ldots, \beta_A^{\mathrm{LC}})^\top$, $\mathbf{1}_{A-1}$ is a vector of ones with length $A-1$, $\boldsymbol{I}_{A-1}$ and $\boldsymbol{J}_{A-1}$ are the identity matrix and matrix of ones, respectively, with dimension $(A-1)\times(A-1)$. Note that the transformed beta prior on ρ^{LC} implies that a stationary AR(1) model with drift is fitted to κ_t^{LC}, which has a computational advantage of avoiding $|\rho^{\mathrm{LC}}| > 1$ (that can result in explosive behaviour in projection and unstable MCMC computation). A random walk with drift model can be separately fitted to κ_t^{LC} by setting $\rho^{\mathrm{LC}} = 1$ deterministically. The two projection models for κ_t^{LC} can then be combined using model averaging (see Section 5.1). For the purpose of prior analysis here, we focus on the transformed beta prior since the case with $\rho^{\mathrm{LC}} = 1$ follows trivially.

To formulate prior distributions with compatible information for the API model, the following priors are chosen by matching the first two moments (see Online Supplementary Material, Appendix B for details) of the parameters based on the correspondence relationship in (15):

$$\left.\begin{aligned}
\alpha_x^{\mathrm{API}} &\overset{\text{ind}}{\sim} N(-5, 14) \\
\beta_x^{\mathrm{API}} &\overset{\text{ind}}{\sim} N(0, 0.01) \\
\frac{\rho^{\mathrm{API}}+1}{2} &\sim \mathrm{Beta}(3, 2) \text{ where } \rho^{\mathrm{API}} \in (-1, 1) \\
(\sigma_\kappa^{\mathrm{API}})^2 &\sim \mathrm{Gamma}(0.1, 5\times 10^{-7}) \\
\phi^{\mathrm{API}} &\sim \mathrm{Gamma}(25, 0.05)
\end{aligned}\right\}\tag{17}$$

It is evident from Figure 2 that the priors specified for both models are incompatible even after the moment-matching procedure. The priors of $\log\mu_{xt}$ under the LC model have sharper peaks around the region $(-15, 0)$, where the likelihood is expected to dominate. Being an integrated likelihood with respect to prior, there is the potential for the marginal likelihood to artificially favour the LC model. Conversely, the priors under the API model over-penalize the likelihood by allocating excessive weight to regions where likelihood is known a priori to be essentially negligible.

An investigation using quantile-quantile (Q-Q) plot also reveals that the priors of $\log\mu_{xt}$ under the LC model are more heavy-tailed due to a mismatch of family of distribution. For example, the relationship $\beta_x^{\mathrm{LC}}\psi_2^{\mathrm{LC}} \longleftrightarrow \beta_x^{\mathrm{API}}$ matches a product of two normally distributed random variables to a single normal random variable (similarly for α_x and κ_t). More precisely, suppose that $U \sim N(0, \sigma_u^2)$ and $V \sim N(0, \sigma_v^2)$ are two independent variables, then $X = UV$ has a probability density function given by

$$f(x) = \frac{1}{\pi\sigma}K_0\left(\frac{|x|}{\sigma}\right)\tag{18}$$

where $\sigma = \sigma_u\sigma_v$ and $K_0()$ is the modified Bessel function of the second kind of order zero (see Craig, 1936). Henceforth, a distribution with density function as in (18) is called a Bessel distribution with

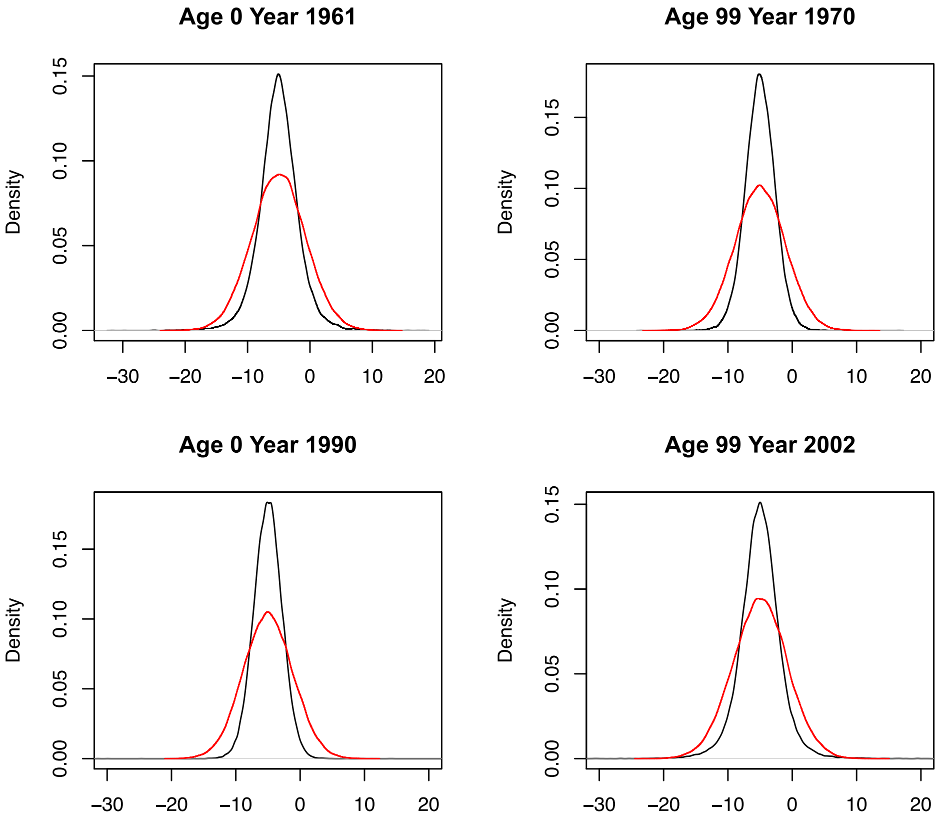

Figure 2. A plot of kernel density estimates of the implied priors of several chosen $\log \mu_{xt}$ for the Lee–Carter (black) and age-period-improvement (red) models, using the specification in (16) and (17).

parameter σ, Bess (σ). A Bessel distribution possesses a significantly heavier tail than a normal distribution (kurtosis of nine compared to three). Thus, to better satisfy the correspondence relationships in terms of prior specification, Bessel priors should be imposed on the relevant parameters of the API model.

3.3 Compatible prior specification: Laplace priors

Rather than using the structurally more complex Bessel distribution, we propose to use a Laplace (double-exponential) distribution with location parameter $-\infty < a < \infty$ and scale parameter $b > 0$. This is denoted by Laplace (a, b), the density of which is

$$f(x) = \frac{1}{2b} \exp\left(-\frac{|x-a|}{b}\right)$$

The Laplace distribution is a compound normal distribution, formed by specifying an exponential distribution on the variance of a normal random variable (see, for example, Johnson et al., 1995). That is, if

$$X \mid \sigma^2 \sim N(\mu, 2\sigma^2) \quad \text{with } \sigma^2 \sim \text{Exp}(\lambda) \tag{19}$$

where μ is a known constant, then marginally $X \sim \text{Laplace}(a = \mu, b = \lambda^{-(1/2)})$. Thus, the normal priors suggested previously can be easily modified into Laplace priors by allowing the variances to be hyperparameters with exponential distributions.

The Laplace distribution is sometimes used as a heavy-tailed replacement for the normal distribution (see Chen et al., 2012; Puig & Stephens, 2000). Our preliminary study also indicates that a Laplace distribution provides a reasonable approximation to the Bessel distribution in terms of matching the tail decay rate and the density around the peak. Consider three distributions which have the same variance $\sigma_u^2\sigma_v^2$, Bess$(\sigma_u\sigma_v)$, Laplace$(0, \sigma_u\sigma_v/\sqrt{2})$, and $N(0, \sigma_u^2\sigma_v^2)$, with density functions denoted, respectively, by f_B, f_L, f_N. The asymptotic expansion provided by Dempsey and Benson (1960) shows that $K_0(|x|/\sigma_u\sigma_v)$ is dominated by the term $(\pi\sigma_u\sigma_v/2|x|)^{1/2} \times \exp(-|x|/\sigma_u\sigma_v)$ for large values of $|x|$. We also know $f_L(x) \propto \exp(-\sqrt{2}|x|/\sigma_u\sigma_v)$ and $f_N(x) \propto \exp(-x^2/2\sigma_u^2\sigma_v^2)$ as $|x| \to \infty$. This indicates that f_B and f_L demonstrate similar tail decay rates; while f_N decays at a much faster rate. This phenomenon can be observed via the illustrative plots in Figure 3.

Our findings also suggest that a Laplace distribution is better at characterizing the dramatic peak around the centre of the Bessel distribution than the more commonly used Student *t*-distribution, which has difficulty matching sharpness of the peak and the heavy tail weight simultaneously (see also Balanda, 1987; Claus, 1968; Horn, 1983).

The priors we propose are $\alpha_x^{\text{API}} \overset{\text{ind}}{\sim} \text{Laplace}(a_\alpha, b_\alpha)$ and $\beta_x^{\text{API}} \overset{\text{ind}}{\sim} \text{Laplace}(a_\beta, b_\beta)$, where we set $a_\alpha = -5$ and $a_\beta = 0$ by directly matching the modes, while b_α and b_β are chosen on the basis of quantile-matching (because moment-based comparison is unable to recognize the dominant features of a Laplace distribution). Specifically, b_α is such that $b_\alpha = -(-5 - L_{\alpha;0.05})/\log(2 \times 0.05)$, where $L_{\alpha;0.05}$ is the (lower) 5th percentile of $\alpha_x^{\text{LC}} + \beta_x^{\text{LC}}\psi_1^{\text{LC}}$; while $b_\beta = -(0 - L_{\alpha;0.05})/\log(2 \times 0.05)$, where $L_{\beta;0.05}$ is the (lower) 5th percentile of $\beta_x^{\text{LC}}\psi_2^{\text{LC}}$. The numerically determined values are given by $b_\alpha = 2.5$ and $b_\beta = 0.03$.

Similarly, the projection model for κ_t^{API} can be extended to be

$$\kappa_t^{\text{API}} \mid \kappa_{t-1}^{\text{API}}, \rho^{\text{API}}, (\sigma_\kappa^{\text{API}})^2 \sim N(\rho^{\text{API}}\kappa_{t-1}^{\text{API}}, 2(\sigma_\kappa^{\text{API}})^2)$$
$$(\sigma_\kappa^{\text{API}})^2 \mid \lambda^{\text{API}} \sim \text{Exp}(\lambda^{\text{API}})$$

where λ^{API} is a hyperparameter to be given a prior distribution. This model could also be expressed marginally as

$$\kappa_t^{\text{API}} \mid \kappa_{t-1}^{\text{API}}, \rho^{\text{API}}, \lambda^{\text{API}} \sim \text{Laplace}(\rho^{\text{API}}\kappa_{t-1}^{\text{API}}, (\lambda^{\text{API}})^{-\frac{1}{2}}) \tag{20}$$

which is an AR(1) model with Laplace innovations as described in Wolf and Gastwirth (1967).

A summary of the compatible priors we propose for the API model is as follows:

$$\left.\begin{aligned}
\alpha_x^{\text{API}} &\overset{\text{ind}}{\sim} \text{Laplace}(-5, 2.5)\\
\beta_x^{\text{API}} &\overset{\text{ind}}{\sim} \text{Laplace}(0, 0.03)\\
\boldsymbol{\kappa}_{-\{1,2\}}^{\text{API}} \mid \rho^{\text{API}}, (\sigma_\kappa^{\text{API}})^2 &\sim N(\mathbf{0}, 2(\sigma_\kappa^{\text{API}})^2\mathbf{W})\\
\frac{\rho^{\text{API}}+1}{2} &\sim \text{Beta}(3, 2) \quad \text{where } \rho^{\text{API}} \in (-1, 1)\\
(\sigma_\kappa^{\text{API}})^2 &\sim \text{Exp}(\lambda^{\text{API}})\\
\lambda^{\text{API}} &\sim \text{Gamma}(1, 2.5 \times 10^{-7})\\
\phi^{\text{API}} &\sim \text{Gamma}(25, 0.05)
\end{aligned}\right\} \tag{21}$$

The kernel density estimates of the implied priors for several $\log\mu_{xt}$ under specifications (16) and (21) are illustrated in Figure 4. The resulting priors of $\log\mu_{xt}$ for the two models are now practically the same, particularly for the region of interest $(-15, 0)$ (they cannot be exactly the same because of the naturally distinct model structures and constraints). Marginal likelihoods (and hence Bayes factors) computed can now serve as a fair model comparison criterion. Note that despite the major impact on marginal likelihoods, the change in prior distributions is not consequential in the estimation of the parameters, given the size of our mortality data.

4 Computation

4.1 Posterior sampling

MCMC methods are employed in posterior sample generation, from which subsequent inferences are drawn. The MCMC algorithm we adopt is the variable-at-a-time Metropolis–Hastings (MH)

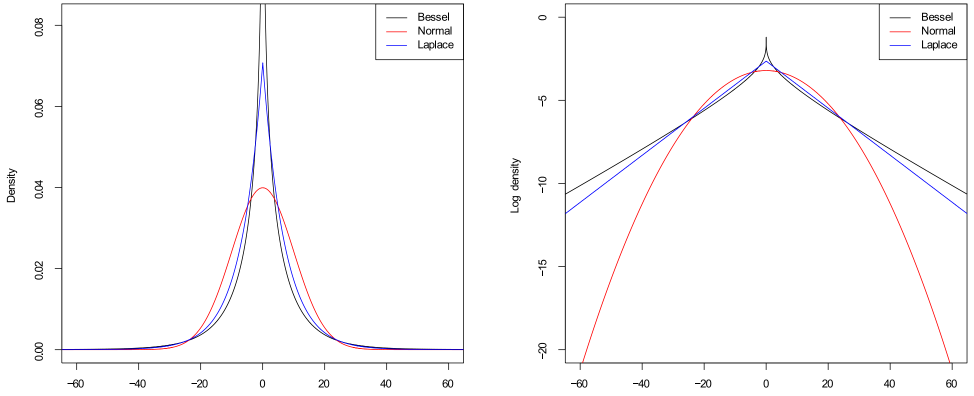

Figure 3. Plots of density (left) and log density (right) functions of Bess($\sigma_u\sigma_v$), $N(0, \sigma_u^2\sigma_v^2)$, and Laplace$(0, \sigma_u\sigma_v/\sqrt{2})$, where $\sigma_u^2 = \sigma_v^2 = 10$.

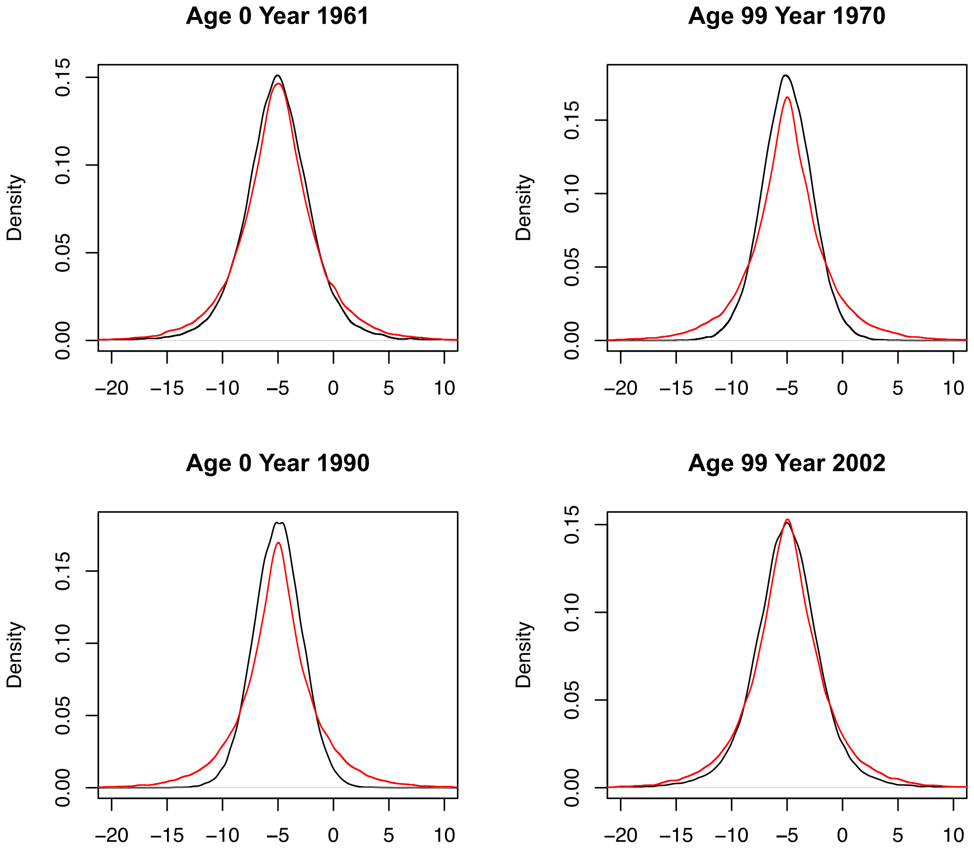

Figure 4. A plot of kernel density estimates of the implied priors for several chosen $\log \mu_{xt}$ under the Lee–Carter (black) and age-period-improvement (red) models, using the compatible priors.

updating as described in Hastings (1970), where each component of the parameters is updated sequentially in each iteration, conditional on the rest. Where conditional posterior distributions are tractable, the Gibbs sampling algorithm is used; where they are intractable, the random walk MH algorithm is used.

For the API model, we will primarily apply the random walk MH updating as priors are mostly nonconjugate (see Online Supplementary Material, Appendix C). For the LC model, the MCMC updating scheme as developed by Wong et al. (2018) can be implemented with appropriate modification to account for the different prior specification and constraint. Online Supplementary Material, Appendix D provides some technical details to highlight the changes made, specifically for $(\sigma_\kappa^{\mathrm{LC}})^2$, ρ^{LC}, and $\boldsymbol{\kappa}_{-1}^{\mathrm{LC}}$.

Note that the constraints on κ_t induce (time-varying) correlation among the κ_t for both models, the effect of which is more notable for the API model due to the additional constraint $\sum_t t\kappa_t = 0$. Hence, we update κ_t in blocks rather than univariately to facilitate posterior exploration (again, see Online Supplementary Material, Appendices C and D for details).

The initial values recommended by Wong et al. (2018) are adopted for both models, with an additional $\lambda^{\mathrm{API}} = 1$. A burn-in phase of 10,000 iterations is applied, with a posterior sample thinning (collecting one realization every 500 iterations) for each parameter. A sample of size 10,000 is obtained for each of the models. Before making any inferential comparisons, trace and autocorrelation plots are checked to ensure sample trajectories demonstrate proper mixing.

4.2 Bayesian model comparison

Bridge sampling, developed by Meng and Wong (1996), is adapted to compute the marginal likelihood of each model. Following the empirical investigation of Wong et al. (2020), the cross-splitting approach is implemented within the bridge sampling algorithm to improve the accuracy of our estimates (see Online Supplementary Material, Appendix E for details). Table 1 gives the estimated log marginal likelihoods of four of the models fitted and the associated posterior model probabilities (assuming equal prior model probabilities). Here, we let API-AR1 and API-RW be abbreviations for the API model with AR(1) and random walk with drift models on κ_t, respectively. LC-AR1 and LC-RW are defined analogically. Overall, the conclusion from this table is now in agreement with that of BIC, that the API models outperform the LC models by a considerable margin. Rather than selecting the best model (API-RW), the models can be combined using BMA. However, this is not the focus now but will become relevant in Section 5.1.

5 Cohort models

Cohort or year-of-birth effects refer to the phenomenon where individuals born in the same time period exhibit similar health characteristics due to common exposures to factors such as smoking behaviours, diets, socio-economic factors, etc. (Wadsworth, 1991). The existence of cohort effects is a prominent feature of the UK mortality data (thoroughly discussed by Willets, 2004), and has been found to be more significant than period effects (Kermack et al., 1934; Richards et al., 2006). These generational effects have strong explanatory and predictive potentials for mortality patterns (Cairns et al., 2007; Willets, 1999), and hence, should not be ignored.

The API model can be easily extended to account for cohort effects. Mathematically, we fit equations (1)–(3) with

$$M_{xt} = \alpha_x + \beta_x t + \kappa_t + \gamma_c \tag{22}$$

where γ_c are cohort parameters, $c = t - x \in \{1, \ldots, C\}$ is the cohort index representing cohorts born in $\{1861, \ldots, 2001\}$, and $C = A + T - 1$. For model identifiability, the following constraints,

$$\sum_t \kappa_t = \sum_t t\kappa_t = \sum_c \gamma_c = \sum_c c\gamma_c = \sum_c c^2\gamma_c = 0 \tag{23}$$

Table 1. The log marginal likelihoods of each model approximated by bridge sampling, and the corresponding posterior model probabilities

Model	Log marginal likelihood	Posterior model probability
API-AR1	−23,690.51	0.4110
API-RW	−23,690.15	0.5890
LC-AR1	−23,800.20	0.0000
LC-RW	−23,798.61	0.0000

Note. API = age-period-improvement; AR1 = autoregressive; RW = random walk; LC = Lee–Carter.

are adopted. We call this model the negative-binomial APCI model.

For γ_c, we use an ARIMA(1,1,0) as adopted by Villegas et al. (2018), i.e.,

$$\left.\begin{aligned} (\gamma_c - \gamma_{c-1}) &= \rho_\gamma(\gamma_{c-1} - \gamma_{c-2}) + \epsilon_c^\gamma \text{for } c = 3, \ldots, C \\ \gamma_2 - \gamma_1 &= \frac{1}{\sqrt{1 - \rho_\gamma^2}}\epsilon_2^\gamma \\ \gamma_1 &= 100\epsilon_1^\gamma \end{aligned}\right\} \tag{24}$$

where $\epsilon_c^\gamma \overset{\text{ind}}{\sim} N(0, \sigma_\gamma^2)$ for $c = 1, \ldots, C$, and ρ_γ and σ_γ^2 are hyperparameters. Applying the constraints $\sum_c \gamma_c = \sum_c c\gamma_c = \sum_c c^2\gamma_c = 0$ on (24), we write

$$\boldsymbol{\gamma}' \sim N_{C-3}(\mathbf{0}, \sigma_\gamma^2 \boldsymbol{W}_\gamma) \tag{25}$$

where $\boldsymbol{\gamma}' = (\gamma_2, \ldots, \gamma_{71}, \gamma_{73}, \ldots, \gamma_{C-1})^\top$, and $\boldsymbol{W}_\gamma$ is chosen such that the model structure in (24) with the constraints is maintained (see Online Supplementary Material, Appendix A). For computational stability, $\{\gamma_1, \gamma_{72}, \gamma_C\}$ are removed from the parameter space (rather than $\{\gamma_1, \gamma_2, \gamma_3\}$) and can be derived as

$$\gamma_1 = \frac{1}{71 \times (C-1)} \sum_{c \neq 1,72,C} (c - 72)(C - c)\gamma_c$$

$$\gamma_{72} = -\frac{1}{69 \times 71} \sum_{c \neq 1,72,C} (C - c)(c - 1)\gamma_c$$

$$\gamma_C = \frac{1}{69 \times (C-1)} \sum_{c \neq 1,72,C} (c - 1)(72 - c)\gamma_c$$

Similarly, the LC model can be extended by including the cohort parameters, i.e.,

$$M_{xt} = \alpha_x + \beta_x \kappa_t + \gamma_c \tag{26}$$

along with the constraints in (23). We call this the negative-binomial Lee-Carter with cohorts (LCC) model. We remark that the choice of constraints was to ensure compatibility between the competing models. When using the different constraints and the priors below, we did not encounter any convergence issues as reported by, for example, Hunt and Villegas (2015) and Cairns et al. (2007).

To facilitate Bayesian model comparison, the model in (25) is used for the cohort parameters under the LCC model. The compatible priors established in Section 3.3 are also imposed for relevant parameters under the APCI and LCC models. For complete model specification, we set priors

$$\rho_\gamma \sim N(0, 1) \quad \text{and} \quad \sigma_\gamma \sim \text{Uniform}(0.1)$$

The uniform prior on σ_γ was motivated by Gelman (2006) to avoid computational issues. Specifically, our pilot study indicates that the use of the conditionally conjugate inverse gamma prior on σ_γ^2 causes convergence issues in the MCMC runs, and the results are also very sensitive to the parameters chosen for the prior.

For posterior sample generation, the MCMC schemes described in Section 4.1 are used, with the additional updating steps for $\boldsymbol{\gamma}'$, σ_γ^2, and ρ_γ provided in Online Supplementary Material, Appendix F.

Table 2. The log marginal likelihoods (estimated using bridge sampling) and posterior model probabilities of the cohort models

Model	Log marginal likelihood	Posterior model probability
LCC-AR1	−22,629.20	0.0000
LCC-RW	−22,634.78	0.0000
APCI-AR1	−22,414.12	0.9364
APCI-RW	−22,416.81	0.0636

Note. AR1 = autoregressive; RW = random walk; APCI = age-period-cohort-improvement.

5.1 Bayesian model comparison and averaging

As before, we fit both stationary AR(1) and random walk models on κ_t, forming the LCC-AR1, LCC-RW, APCI-AR1, and APCI-RW models correspondingly. The estimated marginal likelihoods and the associated posterior model probabilities (assuming equal prior model probabilities) of all the cohort models are provided in Table 2. This indicates that the APCI models offer a substantially better fit in the prior predictive sense to the data than the LCC models. Hence, we eliminate the LCC models and only consider the APCI models hereon.

BMA (see, for example, Hoeting et al., 1999) is then applied to combine the two projection models to form the model-averaged APCI model. From a sampling perspective, this can be achieved by combining the posterior samples using the posterior model probabilities in Table 2. Specifically, a sample of size 10,000 for the model-averaged APCI model can be obtained by combining 9,364 and 636 posterior samples, ,respectively from the APCI-AR1 and APCI-RW models (readily available from the MCMC output). A similar technique can be applied for the API-AR1 and API-RW models to form the model-averaged API model using the probabilities in Table 1. Model uncertainty in light of the projection models is then incorporated into the model-averaged results. Another advantage of model averaging in this context is the ability to generate sensible projections, which avoid the explosive behaviour by random walk model and also overly optimistic prediction intervals by the stationary AR(1) model. In what follows, results illustrated are based on the model-averaged outcome.

5.2 Results

In presenting the results of the APCI model, we include those of the API model to illustrate the importance of including cohort effects. Where relevant, results of the classical-APCI (c-APCI) model by Richards et al. (2019) are also included to highlight the difference due to overdispersion and Bayesian methods.

5.2.1 Estimated parameters

Figure 5 depicts the posterior medians of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\kappa}$, and $\boldsymbol{\gamma}$, accompanied by 95% credible intervals. For the c-APCI model, point estimates (no parameter uncertainty) are computed using the maximum likelihood algorithm by Richards et al. (2019), where the projected values with 95% intervals of $\boldsymbol{\kappa}$ and $\boldsymbol{\gamma}$ are, respectively, given by extrapolating equations (11) and (24). Notice that the estimated $\boldsymbol{\gamma}$ are slightly smoother under the APCI model, with narrower prediction intervals. Smoother $\boldsymbol{\gamma}$ is advantageous for avoiding major fluctuations for projected death rates as cohort effects propagate into the future (see the next subsection). This is sensible because cohorts born in nearby periods are expected to possess similar generational characteristics, hence similar mortality experiences. Moreover, any irregular jumps, notably the cohort effects around 1918 and 1945, where a strong dip is immediately followed by a sudden surge, are mildly mitigated. These irregularities are often associated with major world events such as 1918 Influenza Pandemic and World Wars, which were identified by Cairns et al. (2016) as a consequence of miscalibrated exposures due to uneven birth patterns. We do not address this issue here. Instead, we hope that the mild smoothing manage to weaken the impact of the anomalies on subsequent projections.

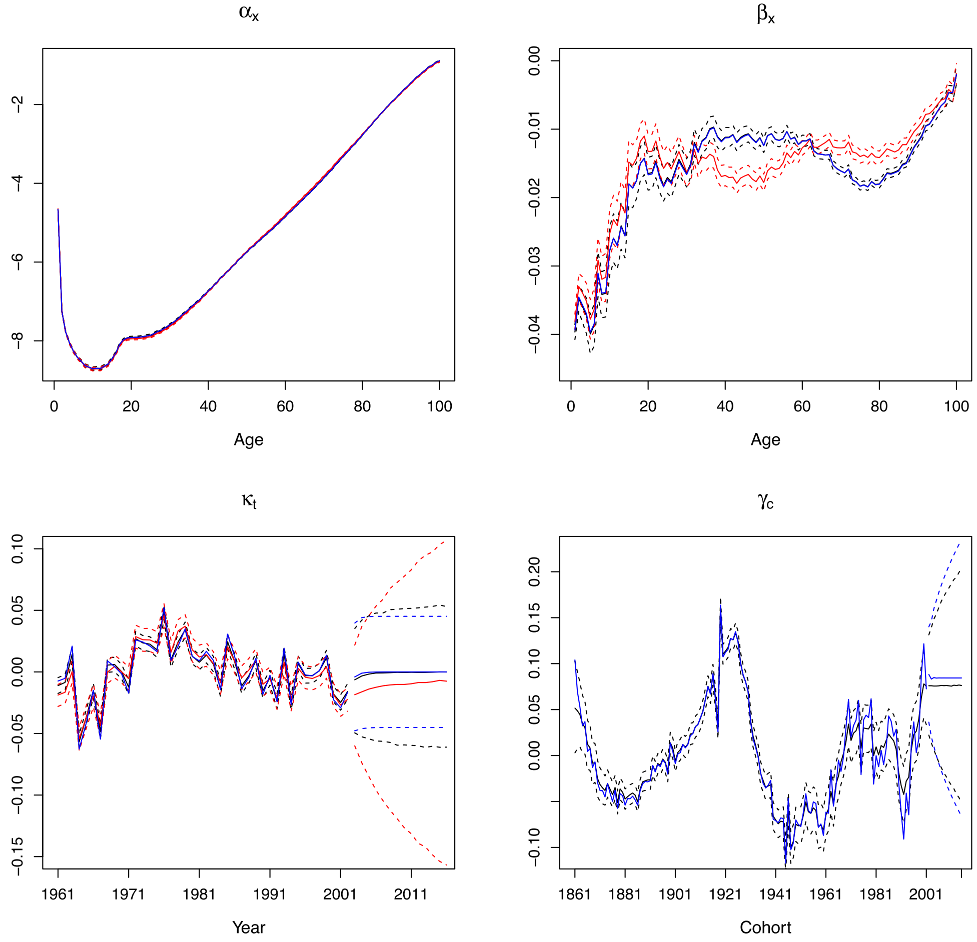

Figure 5. Plot of the medians (solid lines) and 95% intervals (dotted lines) of fitted and projected model parameters under the age-period-cohort-improvement [APCI (black)], age-period-improvement (red), and classical-APCI models (blue).

Estimated hyperparameters under the APCI and API models are presented in Figure 6. The marginal posterior of ρ is a combination between the continuous distribution in the region $\rho \in (0, 1)$ (a stationary AR(1) model) and the discrete peak at $\rho = 1$ (a random walk model), where the peaks are weighted according to the posterior model probabilities. This result has a resemblance with the bimodal marginal posterior for ρ by Wong et al. (2018), where their 'proxy' peak at $\rho = 1$ corresponds to the MCMC algorithm accepting values close to one. Our representation here has the advantage of explicitly distinguishing the peak at exact value of $\rho = 1$ from the peak for $\rho \in (0, 1)$ with appropriate proportions. Here, a smaller weight of 0.0636 is allocated for the peak at $\rho = 1$ under the APCI model as compared to that of the API model (0.5890), suggesting that a stationary AR(1) projection model for κ_t is preferred over the random walk model. This attributes to the relatively narrower prediction intervals for $\boldsymbol{\kappa}$ (see Figure 5) and other mortality quantities under the APCI model (see later).

The level of overdispersion implied by the APCI model is smaller than the API model, as indicated by posterior medians of approximately 3.1×10^{-4} and 1.4×10^{-3} for $1/\phi$ respectively under the APCI model and its counterpart. We also note from equation (5) that the term $\mathbb{E}[D_{xt}]/\phi$ can be viewed as relative increase in the variance of D_{xt} with respect to its mean, which measures the degree of overdispersion. As a quantitative illustration, substituting the mean observed number of deaths as $\mathbb{E}[D_{xt}]$ and the posterior median of ϕ gives $2846.95/3228.55 \approx 0.88$ and

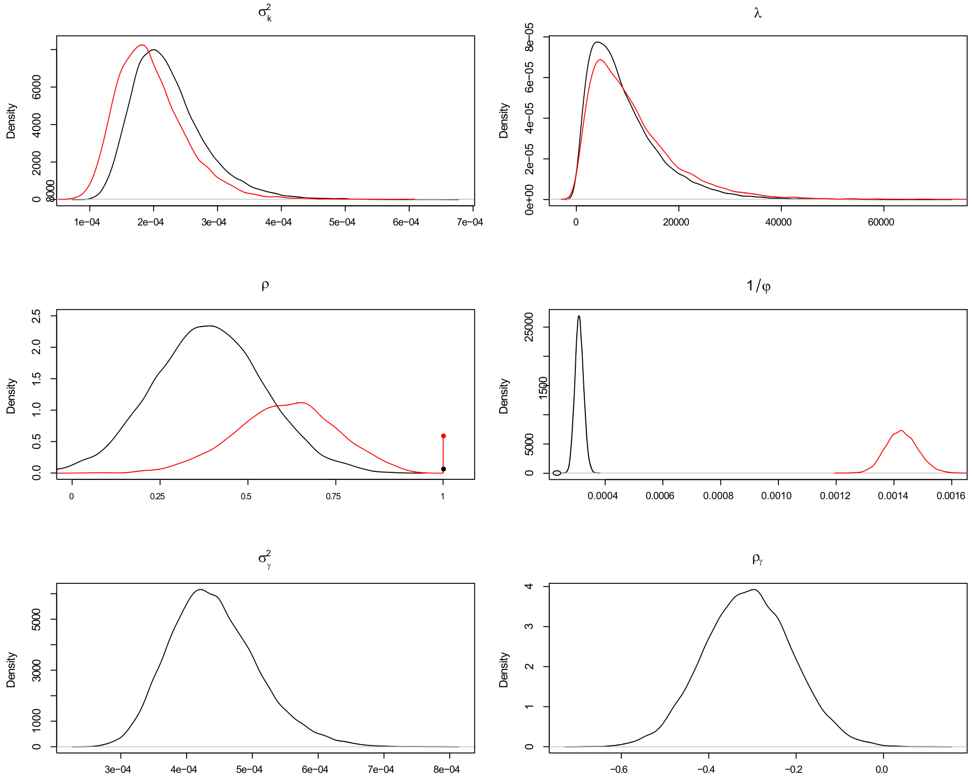

Figure 6. Kernel density plots of the hyperparameters under the age-period-cohort-improvement (black) and age-period-improvement (red) models.

$2846.95/701.28 \approx 4.06$, respectively, for the APCI and API models, confirming a smaller magnitude of overdispersion for the APCI model. This is expected because in the absence of cohort components, uncaptured mortality trends are misregarded as extra variations which overestimates overdispersion. By preventing the model from misidentifying the cohort effect as a form overdispersion, the dispersion parameter fitted under the APCI model provides a more precise description of the mortality heterogeneity present in the data.

5.2.2 Fitted and projected crude mortality rates

We also assess the performances of the models using crude mortality rates (observable) rather than the underlying mortality rates, μ_{xt} (unobservable). To obtain the fitted crude mortality rates, the fitted number of deaths, D^F_{xt}, are first generated through equation (4), where joint posterior samples of $\alpha_x, \beta_x, \kappa_t, \gamma_c$, and ϕ are substituted as appropriate. The fitted crude mortality rate for each $x = 1, \ldots, A$ and $t = 1, \ldots, T$ is then computed as $\hat{\mu}_{xt} = D^F_{xt}/e_{xt}$.

For mortality projection, we note that the posterior predictive distribution of 1-year ahead number of deaths for each age (with age parameters held fixed), under the APCI model for instance, is

$$\begin{aligned} f(D_{x\,T+1} \mid \boldsymbol{d}) = \int f(D_{x\,T+1} \mid \alpha_x, \beta_x, \kappa_{T+1}, \phi) \times f(\kappa_{T+1} \mid \kappa_T, \rho, \sigma^2_\kappa) \\ \times f(\gamma_{T+1-x} \mid \gamma_{T-x}, \gamma_{T-x-1}, \rho_\gamma, \sigma^2_\gamma) \\ \times f(\alpha_x, \beta_x, \kappa_T, \rho, \sigma^2_\kappa, \gamma_{T+1-x}, \rho_\gamma, \sigma^2_\gamma, \phi \mid \boldsymbol{d})\, \mathrm{d}\alpha_x \ldots \mathrm{d}\phi \end{aligned} \tag{27}$$

where $f(\alpha_x, \beta_x, \kappa_T, \rho, \sigma^2_\kappa, \gamma_{T+1-x}, \rho_\gamma, \sigma^2_\gamma, \phi \mid \boldsymbol{d})$ is the joint posterior distribution. This suggests the following generation procedure for the projected crude mortality rates:

1. Generate κ_{T+1} from equation (11) where joint posterior samples of $(\rho, \kappa_T, \sigma^2_\kappa)$ are substituted appropriately.
2. Generate γ_T from equation (24) where joint posterior samples of $(\gamma_{T-1}, \gamma_{T-2}, \rho_\gamma, \sigma^2_\gamma)$ are substituted. Other 'observed' cohort components $(\gamma_{T-99}, \ldots, \gamma_{T-3})$ are also available from the posterior samples.
3. Generate for each x the forecasted number of deaths using (4), i.e.,

$$D^F_{x\,T+1} \sim \text{Neg-Bin}\left(\phi, \frac{\phi}{e_{x\,T+1}\exp\left(\alpha_x+\beta_x(T+1)+\kappa_{T+1}+\gamma_{T+1-x}\right)+\phi}\right)$$

   where $e_{x\,T+1}$ are holdout central exposed to risks, joint posterior samples of $(\alpha_x, \beta_x, \phi)$ are substituted, κ_{T+1} and γ_{T+1-x} are, respectively, from steps 1 and 2.
4. Compute for each x the projected crude mortality rates as $\hat{\mu}_{x\,T+1} = D^F_{x\,T+1}/e_{x\,T+1}$.

By analogy, h-year ahead projections can be obtained by recursive implementation of the above algorithm. Having generated samples for $\hat{\mu}_{xt}$ for $x = 1, \ldots, A$ and $t = 1, \ldots, T+h$, sample percentiles can be constructed to form median and 95% intervals.

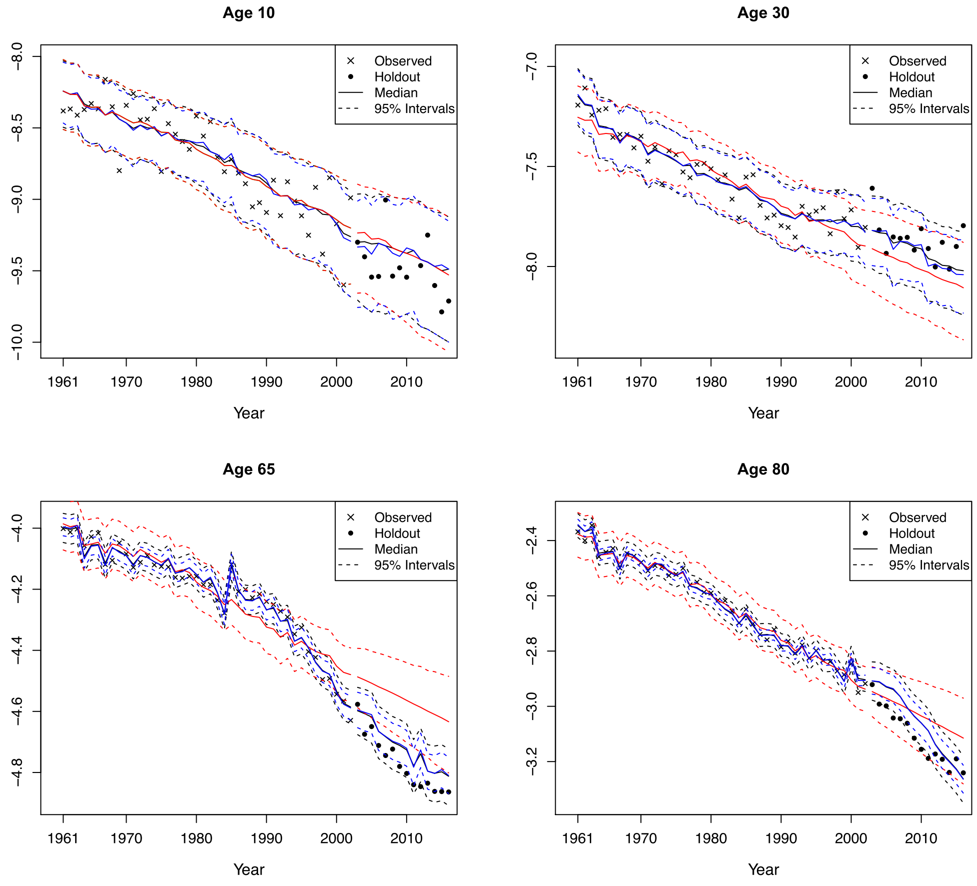

Figure 7. Plots of observed and holdout log crude death rates, fitted log crude death rates, and the associated 14-year ahead projection of the crude log death rates for several chosen ages under the age-period-cohort-improvement [APCI (black)], age-period-improvement (red), and classical-APCI (blue) models, accompanied by 95% intervals.

Under the c-APCI model, we generate $D^F_{xt} \sim \text{Poisson}(e_{xt} \exp(\alpha_x + \beta_x t + \kappa_t + \gamma_c))$ using maximum likelihood estimates of relevant parameters, and then compute the fitted crude mortality rates as $\hat{\mu}_{xt} = D^F_{xt}/e_{xt}$. Similar to above, projection for the c-APCI model proceeds as:

1. Generate κ_{T+1} and γ_T, respectively, from equations (11) and (24) using maximum likelihood estimates of relevant parameters.
2. Generate for each x the forecasted number of deaths as

$$D^F_{x\,T+1} \sim \text{Poisson}(e_{x\,T+1} \times \exp(\alpha_x + \beta_x(T+1) + \kappa_T + \gamma_{T+1-x}))$$

where $e_{x\,T+1}$ are holdout central exposed to risks; maximum likelihood estimates are substituted for α_x and β_x; κ_{T+1} and γ_{T+1-x} are from step 1.

3. Compute for each x the projected crude mortality rates as $\hat{\mu}_{x\,T+1} = D^F_{x\,T+1}/e_{x\,T+1}$.
4. Repeat steps 1–3 recursively to generate h-year ahead projections.

Note that the above generation procedure of fitted and projected mortality rates for the c-APCI model only allows uncertainties due to Poisson random variation and forecast uncertainty. This highlights the difference due to parameter uncertainty and the presence of overdispersion in the Bayesian models.

Figure 7 shows the fitted and projected crude death rates for several chosen ages when $h = 14$. Also included are the observed and holdout crude mortality rates. The incorporation of cohort effects generally leads to significantly better description of the underlying mortality trend for both fitted and projected rates. Crucially, the 95% prediction intervals of the (Bayesian) APCI model offer realistic coverage rates for including the holdout rates, despite their relatively narrow widths. By contrast, the API model generally fails to capture (and hence project) the mortality trends due to the lack of the crucial cohort components, necessitating wider prediction intervals to achieve acceptable coverage rates. This indicates that it is not sufficient to only include overdispersion when trend components are not adequately captured, since unexplained signals will be misidentified as model residuals, resulting in an overestimation of the degree of overdispersion which in turn, generates unnecessarily wide intervals. Appropriately incorporating cohort components improve the calibration between signals and errors, reducing the bias in the mortality projection with adequately wide intervals to characterize uncertainty.

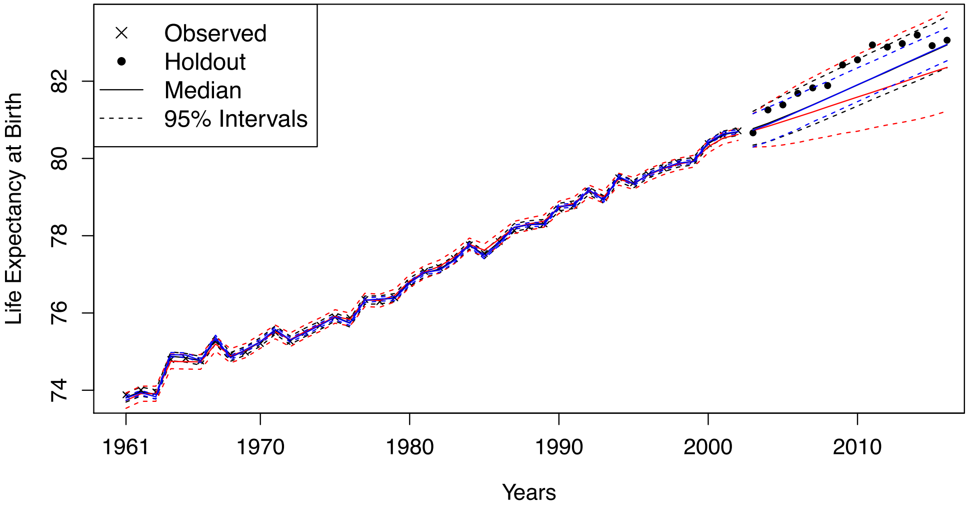

Figure 8. Plots of observed and holdout life expectancy at birth and the associated 14-year ahead projection under the age-period-cohort-improvement [APCI (black)], age-period-improvement (red), and classical-APCI (blue) models, accompanied by the 95% intervals.

Another key thing to highlight is that median crude death rates (both fitted and projected) under APCI are smoother than the c-APCI model. This is a consequence of both the estimated $\boldsymbol{\gamma}$ being smoother with less dramatic fluctuations (see Figure 5), as well as the presence of dispersion parameter to relax the rigid structural assumption of the model. More importantly, 95% prediction intervals of the APCI model are wider than the c-APCI model by construction, allowing most holdout rates to be included. The extra widths of the APCI intervals originate from the additional variation due to overdispersion, parameter uncertainty (through priors), and model uncertainty in the models for $\boldsymbol{\kappa}$. The c-APCI model yields intervals that are noticeably too narrow (with poor coverages) to adequately represent the variabilities of the holdout data. The difference between the predictive capabilities of the models can also be quantified by evaluating their mean absolute errors (e.g., Barigou et al., 2022), logarithmic scores (e.g., Gneiting & Raftery, 2007), and continuous ranked probability scores (Matheson & Winkler, 1976). All three measures, computed using the R package *scoringRules* (Jordan et al., 2019), indicated that the APCI model generates better out-of-sample forecasts than the c-APCI model.

Similarly, in terms of life expectancy at birth as depicted in Figure 8, the APCI model produces the best prediction (with smallest bias and highest coverage rate) when validated against the holdout data. The API model underestimates the mortality improvement by a substantial margin, and is accompanied by excessively wide prediction intervals. The c-APCI model yields a median projection that aligns well with the APCI model, but fails to produce uncertainty bands that correctly represent the future variation by not having sufficiently wide intervals.

6 Conclusion

In this paper, a Bayesian implementation of the APCI model with overdispersion was presented and illustrated on UK mortality data for the entire age range. The interest was to compare this model with the widely used LCC model using posterior model probabilities. For the comparison procedure to be meaningful, initial analysis focused on specifying priors that are compatible for the competing models. Some theoretical investigations revealed that the Laplace distribution provides a good approximation to the Bessel prior distribution formed from the multiplication of two normal priors. This led to our proposed choice of Laplace priors for relevant parameters under the APCI model. The modified prior specification should not affect the estimation of posterior distribution given the size of our mortality data, but has been shown to be consequential in Bayesian model comparison. The prior specification we have developed is recommended if users are interested in carrying out comparison of the same family of models, unless other models are to be considered where similar analytical work should be undertaken to ensure prior information is compatible under all models. After ensuring that the prior information is compatible, the posterior model probabilities computed are strongly in favour of the APCI model. In addition to fitting the data substantially better, the APCI model is also favourable on grounds of structural simplicity, and computational advantages over the LCC model.

To increase the robustness against misspecification due to time series models, we considered two ARIMA models (random walk and AR(1)) for the time-varying parameter κ_t under both APCI and LCC models. The model averaging technique was applied to combine the models to generate probabilistic forecasts that include model uncertainty. The model for $\boldsymbol{\gamma}$, however, was assumed to be ARIMA(0,1,1). Future work could consider averaging across multiple time series models for $\boldsymbol{\gamma}$ instead of using the default model.

Subsequent analysis then focused on illustrating, using holdout data, the significance of Bayesian methods, together with the simultaneous inclusion of dispersion and cohort components. Results suggested that our proposed approach enabled the APCI model to better calibrate between essential mortality data signals and random variations, leading to a noticeable improvement in the qualitative fit and adequately wide prediction intervals. Models that ignore cohort effects are more prone to overestimate the degree of overdispersion, producing unnecessarily wide prediction intervals (as in Wong et al., 2018). Comparing our results with Richards et al. (2019) also showed that allowing for overdispersion using Bayesian methods produces mortality projections with prediction intervals of appropriate width. Even though the degrees of overdispersion implied by all models are less than that claimed by Wong et al. (2018) due to the ‘contamination’ effect (disregarding uncaptured mortality trends as random variations) by the cohort effects, the dispersion parameter is required both intuitively for representing heterogeneity and quantitatively for better predictive properties.

Acknowledgments

The authors greatly appreciate the comments from the editor, associate editor, and all reviewers involved, who helped improving the quality of the paper. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest: The authors have no conflict of interest to declare.

Data availability

The data and code that support the findings of this study are available from the following GitHub repository: https://github.com/jstw1g09/JRSS-C-Bayesian-APCI.

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series C* online.

References

Balanda K. P. (1987). Kurtosis comparisons of the Cauchy and double exponential distributions. *Communications in Statistics-Theory and Methods*, *16*(2), 579–592. https://doi.org/10.1080/03610928708829388

Barigou K., Goffard P.-O., Loisel S., & Salhi Y. (2022). Bayesian model averaging for mortality forecasting using leave-future-out validation. *International Journal of Forecasting*. https://doi.org/10.1016/j.ijforecast.2022.01.011

Bartlett M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*(3–4), 533–534. https://doi.org/10.1093/biomet/44.3-4.533

Booth H., & Tickle L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science*, *3*(1–2), 3–43. https://doi.org/10.1017/S1748499500000440

Börger M., & Aleksic M. C. (2014). *Coherent projections of age, period, and cohort dependent mortality improvements*. A paper presented at Living to 100 Symposium.

Brouhns N., Denuit M., & Vermunt J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, *31*(3), 373–393. https://doi.org/10.1016/S0167-6687(02)00185-3

Cairns A. J. G., Blake D., Dowd K., Coughlan G. D., Epstein D., Ong A., & Balevich I. (2007). A quantitative comparison of stochastic mortality models using data from England & Wales and the United States. *North American Actuarial Journal*, *13*(1), 1–35. https://doi.org/10.1080/10920277.2009.10597538

Cairns A. J. G., Blake D., Dowd K., & Kessler A. R. (2016). Phantoms never die: Living with unreliable population data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *179*(4), 975–1005. https://doi.org/10.1111/rssa.12159

Chen G., Saad Z. S., Nath A. R., Beauchamp M. S., & Cox R. W. (2012). FMRI group analysis combining effect estimates and their variances. *NeuroImage*, *60*(1), 747–765. https://doi.org/10.1016/j.neuroimage.2011.12.060

Claus S. (1968). Zwet, W. R. van: Convex transformations of random variables. Mathematica centre tracts 7. Mathematisch Centrum Amsterdam, 1964, 116 Seiten. *Biometrische Zeitschrift*, *10*(1), 95–95. https://doi.org/10.1002/bimj.19680100134

Continuous Mortality Investigation Bureau (2016a). *CMI mortality projections model consultation* (CMI working paper no. 90). Institute and Faculty of Actuaries.

Continuous Mortality Investigation Bureau (2016b). *CMI mortality projections model consultation* (CMI working paper no. 91). Institute and Faculty of Actuaries.

Craig C. C. (1936). On the frequency function of *xy*. *Annals of Mathematical Statistics*, *7*(1), 1–15. https://doi.org/10.1214/aoms/1177732541

Czado C., Delwarde A., & Denuit M. (2005). Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics*, *36*(3), 260–284. https://doi.org/10.1016/j.insmatheco.2005.01.001

Delwarde A., Denuit M., & Partrat C. (2007). Negative binomial version of the Lee–Carter model for mortality forecasting. *Applied Stochastic Models in Business and Industry*, *23*(5), 385–401. https://doi.org/10.1002/asmb.679

Dempsey E., & Benson G. C. (1960). Note on the asymptotic expansion of the modified Bessel function of the second kind. *Mathematics of Computation*, *14*(72), 362–365. https://doi.org/10.1090/S0025-5718-1960-0120401-1

Gelman A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533. https://doi.org/10.1214/06-BA117A

Girosi F., & King G. (2008). *Demographic forecasting*. Princeton University Press.

Gneiting T., & Raftery A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Hastings W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Hilton J., Dodd E., Forster J. J., & Smith P. W. F. (2019). Projecting UK mortality by using Bayesian generalized additive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *68*(1), 29–49. https://doi.org/10.1111/rssc.12299

Hoeting J. A., Madigan D., Raftery A. E., & Volinsky C. T. (1999). Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. *Statistical Science*, *14*(4), 382–417. https://doi.org/10.1214/ss/1009212519

Horn P. S. (1983). A measure for peakedness. *The American Statistician*, *37*(1), 55–56. https://doi.org/10.1080/00031305.1983.10483090

Human Mortality Database (HMD). *Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France)*. www.mortality.org

Hunt A., & Villegas A. M. (2015). Robustness and convergence in the Lee–Carter model with cohort effects. *Insurance: Mathematics and Economics*, *64*(3), 186–202. https://doi.org/10.1016/j.insmatheco.2015.05.004

Johnson N. L., Kotz S., & Balakrishnan N. (1995). *Continuous univariate distributions* (3rd ed., Vol. 2). Wiley series in probability and mathematical statistics.

Jordan A., Krüger F., & Lerch S. (2019). Evaluating probabilistic forecasts with scoring rules. *Journal of Statistical Software*, *90*(12), 1–37. https://doi.org/10.18637/jss.v090.i12

Kermack W. O., McKendrick A. G., & Mckinlay P. L. (1934). Death-rates in great Britain and Sweden: Some general regularities and their significance. *The Lancet*, *223*(5770), 698–703 (Originally published as Volume 1, Issue 5770). https://doi.org/10.1016/S0140-6736(00)92530-3

Lee R. D., & Carter L. R. (1992). Modelling and forecasting U.S. mortality. *Journal of the American Statistical Association*, *87*(419), 659–671. https://doi.org/10.2307/2290201

Li S. H., Hardy M. R., & Tan K. S. (2009). Uncertainty in mortality forecasting: An extension to the classical Lee–Carter approach. *ASTIN Bulletin*, *39*(1), 137–164. https://doi.org/10.2143/AST.39.1.2038060

Matheson J. E., & Winkler R. L. (1976). Scoring rules for continuous probability distributions. *Management Science*, *22*(10), 1087–1096. https://doi.org/10.1287/mnsc.22.10.1087

Meng X.-L., & Wong W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, *6*(4), 831–860. https://www.jstor.org/stable/24306045

Pedroza C. (2006). A Bayesian forecasting model: Predicting U.S. male mortality. *Biostatistics*, *7*(4), 530–550. https://doi.org/10.1093/biostatistics/kxj024

Puig P., & Stephens M. A. (2000). Tests of fit for the Laplace distribution, with applications. *Technometrics*, *42*(4), 417–424. https://doi.org/10.1080/00401706.2000.10485715

Renshaw A. E., & Haberman S. (2003). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, *33*(2), 255–272 (Papers presented at the 6th IME conference, Lisbon, 15–17 July 2002). https://doi.org/10.1016/S0167-6687(03)00138-0

Renshaw A. E., & Haberman S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, *38*(3), 556–570. https://doi.org/10.1016/j.insmatheco.2005.12.001

Richards S. J., Currie I. D., Kleinow T., & Ritchie G. P. (2019). A stochastic implementation of the APCI model for mortality projections. *British Actuarial Journal*, *24*, E13. https://doi.org/10.1017/S1357321718000260

Richards S. J., Kirkby J. G., & Currie I. D. (2006). The importance of year of birth in two-dimensional mortality data. *British Actuarial Journal*, *12*(1), 5–38. https://doi.org/10.1017/S1357321700004682

Tabeau E., van den Berg Jeths A., & Heathcote C. (2001). *Forecasting mortality in developed countries: Insights from a statistical, demographic and epidemiological perspective*. (Vol. 9). Kluwer Academic Publishers.

Tuljapurkar S., Li N., & Boe C. (2000). A universal pattern of mortality decline in the G7 countries. *Letters to Nature*, *405*(6788), 789–792. https://doi.org/10.1038/35015561

Villegas A. M., Kaishev V. K., & Millossovich P. (2018). StMoMo: An R package for stochastic mortality modeling. *Journal of Statistical Software*, *84*(3), 1–38. https://doi.org/10.18637/jss.v084.i03

Wadsworth M. E. J. (1991). *The imprint of time: Childhood, history and adult life*. Clarendon Press.

Weakliem D. L. (1999). A critique of the Bayesian information criterion for model selection. *Sociological Methods and Research*, *27*(3), 359–397. https://doi.org/10.1177/0049124199027003002

Willets R. C. (1999). *Mortality in the next Millennium*. Staple Inn Actuarial Society (SIAS).

Willets R. C. (2004). The cohort effects: Insights and explanations. *British Actuarial Journal*, *10*(4), 833–877. https://doi.org/10.1017/S1357321700002762

Wolff S., Gastwirth J., & Rubin H. (1967). The effect of autoregressive dependence on a nonparametric test (corresp.). *IEEE Transactions on Information Theory*, *13*(2), 311–313. https://doi.org/10.1109/TIT.1967.1053997

Wong J. S. T., Forster J. J., & Smith P. W. F. (2018). Bayesian mortality forecasting with overdispersion. *Insurance: Mathematics and Economics*, *83*(2018), 206–221. https://doi.org/10.1016/j.insmatheco.2017.09.023

Wong J. S. T., Forster J. J., & Smith P. W. F. (2020). Properties of the bridge sampler with a focus on splitting the MCMC sample. *Statistics and Computing*, *30*(4), 799–816. https://doi.org/10.1007/s11222-019-09918-5