Bayesian model averaging for mortality forecasting using leave-future-out validation*

K. Barigou P.-O. Goffard S. Loisel Y. Salhi

ISFA, Univ Lyon 1, UCBL, LSAF EA2429, F-69007, Lyon, France

Version: March 20, 2021

Abstract

Predicting the evolution of mortality rates plays a central role for life insurance and pension funds. Various stochastic frameworks have been developed to model mortality patterns taking into account the main stylized facts driving these patterns. However, relying on the prediction of one specific model can be too restrictive and lead to some well documented drawbacks including model misspecification, parameter uncertainty and over-fitting. To address these issues we first consider mortality modelling in a Bayesian Negative-Binomial framework to account for overdispersion and the uncertainty about the parameter estimates in a natural and coherent way. Model averaging techniques, which consists in combining the predictions of several models, are then considered as a response to model misspecifications. In this paper, we propose two methods based on leave-future-out validation which are compared to the standard Bayesian model averaging (BMA) based on marginal likelihood. Using out-of-sample errors is a well-known workaround for overfitting issues. We show that it also produces better forecasts. An intensive numerical study is carried out over a large range of simulation setups to compare the performances of the proposed methodologies. An illustration is then proposed on real-life mortality datasets which includes a sensitivity analysis to a Covid-type scenario. Overall, we found that both methods based on out-of-sample criterion outperform the standard BMA approach in terms of prediction performance and robustness.

Keywords: Mortality forecasting, Bayesian model averaging, Age-period-cohort, overdispersion, stacking.

*Contacts: Karim Barigou karim.barigou@univ-lyon1.fr (corresponding author), Pierre-Olivier Goffard pierre-olivier.goffard@univ-lyon1.fr, Stéphane Loisel stephane.loisel@univ-lyon1.fr and Yahia Salhi yahia.salhi@univ-lyon1.fr

1 Introduction

Apart from short epidemic shocks, most developed countries face unprecedented improvements in longevity that contribute to the aging of the population. As a consequence, pension funds, social security systems and life insurers face longevity risk, namely the risk that policyholders live longer than expected. These concerns have led to a large development of stochastic mortality models in the actuarial, demographic and statistical literature. The selection of a specific model is naturally subject to model risk, that is the risk of picking the wrong model. This paper considers a full Bayesian model averaging approach to mitigate this risk while taking into account the uncertainty in the value of the parameters due to the potential lack of fit of the mortality models to the data.

A major part of the literature on stochastic mortality modelling has developed from the seminal work of Lee and Carter (1992). It introduced a factor-based framework on which the mortality surface (on the logarithmic scale) is decomposed into the sum of an age-specific term representing the average mortality rate per age and a bilinear term including a single time-varying index, which represent the mortality trend and an age-specific component that characterizes the sensitivity to this trend at different ages. Several extensions were proposed in the literature. For example, Renshaw and Haberman (2006) proposed an extension of the Lee-Carter model with a cohort effect and Cairns et al. (2006) proposed a two-factor model for pensioners mortality often abbreviated as CBD. The CBD model was then extended by incorporating combinations of a quadratic age term and a cohort effect term in Cairns et al. (2009). Plat (2009) combined the features of existing models to come up with a model that covers the entire age range and takes into account cohort effects. For an overview of existing models, we refer to Hunt and Blake (2020). Mortality forecasts are usually obtained in a frequentist two-step procedure. In a first step, estimates of the parameters are obtained by Singular Value Decomposition or Maximum Likelihood Estimation, noticing that standard mortality models can be expressed as a generalized non-linear or linear model, see Currie (2016). In a second step, parameters are projected using time-series techniques.

In this paper, we consider mortality modelling in a Bayesian framework. When compared to the classical framework, the Bayesian approach offers two notable advantages. First, the estimation and forecasting steps go hand in hand, which leads to more consistent estimates, see Cairns et al. (2011b) and Wong et al. (2018) among others. Second, it better accounts for the different sources of uncertainty in a natural and coherent way. Within the literature on Bayesian mortality modeling, Czado et al. (2005) proposed a fully integrated Bayesian approach tailored to the Poisson Lee-Carter (LC) model. It was extended to the multi-population setting in Antonio et al. (2015). Pedroza (2006) performed mortality forecasting using a Bayesian state-space model using Kalman filters, that handle missing data. Kogure and Kurachi (2010) presented a Bayesian approach to pricing longevity risk under the LC framework. Finally, Venter and Şahın (2018) considered Bayesian shrinkage to obtain a parsimonious parameterization of mortality models.

To account for model uncertainty, we consider model averaging. Compared to using the predictions of one specific model, combining the forecasts of various models is more robust toward model mis-specification and is more likely to produce reliable point and interval forecasts. There are two standard approaches to model averaging: a frequentist approach based

on the Akaike Information Criterion (AIC) by Buckland et al. (1997) and a Bayesian approach known as Bayesian model averaging, see Hoeting et al. (1999), relying on the Bayes factor, see Kass and Raftery (1995). While both approaches received much attention in several areas such as ecology (Cade (2015)) or finance (Koop and Korobilis (2012)), there are only few papers in the context of demography and actuarial science. Shang (2012) combined mortality forecasts based on two weighting schemes, the first is based on out-of-sample forecast accuracy and the other relies on in-sample goodness-of-fit. Instead of choosing the optimal weights, Shang and Haberman (2018) considered trimming a set of models before equally averaging forecasts from the selected superior models. In the Bayesian setting, we only found Benchimol et al. (2018) who applied Bayesian Model Averaging (BMA) to combine four popular mortality models via their posterior probability. However, they did not show the mathematical details nor did they compare the BMA with the single-model forecasts.

In this paper, we propose a full Bayesian approach for mortality forecasting. We first sample from the posterior distribution of the mortality model parameters using Markov Chain Monte Carlo (MCMC) techniques. We then derive weights for each mortality model. The standard method for calculating the Bayesian model weights uses a marginal likelihood approximation. The latter characterizes the suitability of the model to the data used to train this very model. We therefore introduce two alternative model averaging methods based on the forecast accuracy measured on a validation data set (different from the training one). The validation set is made of the most recent years, hence the name *leave-future-out* validation. We refer to these method as *stacking* and *pseudo-BMA* because they follow from an adaptation of the model averaging strategies described in the work of Yao et al. (2018) based on leave-one-out validation. We show that stacking and pseudo-BMA outperform the standard averaging approach in terms of forecasting accuracy when applied to real as well as simulated mortality data. To the best of our knowledge, this is the first time that a Bayesian model averaging approach based on out-of-sample performance is considered for mortality forecasting.

The remainder of the paper is organized as follows. In Section 2, we introduce the Bayesian mortality modeling framework which accommodates a wide range of well-known mortality models. In Section 3, we discuss model aggregation strategies, starting with the standard method before moving on to the alternative methods designed to make predictions. In Section 4, an intensive numerical study is carried out accross a large range of simulation setups to provide a fair comparison of the proposed methodologies. Section 5 compares the prediction performance of the model averaging methodologies on real-life mortality datasets. Section 6 investigates the impact of a COVID-type effect on the mortality rate projections and Section 7 provides some concluding remarks and perspectives for future research work.

2 Bayesian Mortality Modeling

When studying human mortality, the data at hand consist of death counts $d_{x,t}$ and central exposures $e_{x,t}$, where $x = x_1, x_2, \ldots, x_A$ and $t = t_1, t_2, \ldots, t_N$ represent a set of A age groups and N calendar years respectively. We denote by $\mu_{x,t}$ the force of mortality at age x and calendar year t. A stochastic mortality model commonly relies on two assumptions:

1. The number of deaths is modelled by a counting random variable $D_{x,t}$ following either a Poisson or a binomial distribution.

2. The force of mortality has a log or logit link to the age and calendar year variables.

2.1 Negative-Binomial model

The data provided to mortality models are generally at the country level. Empirical studies have shown that life expectancy depends on socioeconomic status, individual income, education, marital status, among other factors. This heterogeneity within a given population tends to increase the variability of the underlying death counts, leading to overdispersion. To tackle this issue, we consider a classic extension of the Poisson distribution, namely a gamma mixture of Poisson distributions, which assumes that

$$D_{x,t} \mid \mu_{x,t} \overset{\text{ind}}{\sim} \mathsf{Poisson}(\mu_{x,t} e_{x,t}) \tag{2.1}$$

$$\log \mu_{x,t} = \alpha_x + \sum_{i=1}^{p} \beta_x^{(i)} \kappa_t^{(i)} + \beta_x^{(0)} \gamma_{t-x} + \log \nu_{x,t} \tag{2.2}$$

$$\nu_{x,t} \mid \phi \overset{\text{ind}}{\sim} \text{Gamma}(\phi, \phi), \tag{2.3}$$

where the average mortality rate within each age group relates to the α_x coefficient, while age specific patterns of mortality improvement over time are captured through the $\beta_x^{(i)}$ and $\kappa_t^{(i)}$ for $i = 1, \ldots, p$. The model can accomodate for an age-specific cohort effect with the product of $\beta_x^{(0)}$ by γ_{t-x} while overdispersion relates to the parameter ϕ. The expectation and variance of this model are given by

$$\mathbb{E}\left[D_{x,t}\right] = e_{x,t} \exp\left(\alpha_x + \sum_{i=1}^{p} \beta_x^{(i)} \kappa_t^{(i)} + \beta_x^{(0)} \gamma_{t-x}\right) \tag{2.4}$$

$$\text{Var}\left[D_{x,t}\right] = \mathbb{E}\left[D_{x,t}\right] \times \left[1 + \frac{\mathbb{E}\left[D_{x,t}\right]}{\phi}\right] > \mathbb{E}\left[D_{x,t}\right]. \tag{2.5}$$

This model has the same mean as the standard Poisson model but possesses a larger variance which depends on the value of ϕ. When $\phi \to \infty$, we recover the standard Poisson model. An important feature of this model is its equivalence to a Negative-Binomial (NB) model, in the sense that

$$D_{x,t} \mid \alpha_x, \beta_x, \kappa_t, \gamma_{t-x}, \phi \sim \text{Neg} - \text{Bin}\left(e_{x,t} \exp\left(\alpha_x + \sum_{i=1}^{p} \beta_x^{(i)} \kappa_t^{(i)} + \beta_x^{(0)} \gamma_{t-x}\right), \phi\right).$$

The NB model was considered in a frequentist framework by Delwarde et al. (2007) and in a Bayesian setting by Wong et al. (2018). We remark that Wong et al. (2018) compared the NB model with a Poisson model with normal random error $\nu_{x,t}$, and found that both specifications provide similar fits.

Under the NB assumption, the full likelihood of the death records is given by

$$l(y \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \phi) = \prod_{x,t} \left\{ \frac{\Gamma\left(d_{xt} + \phi\right)}{\Gamma(\phi)\Gamma\left(d_{xt} + 1\right)} \left[\frac{e_{xt} \exp\left(\eta_{x,t}\right)}{e_{xt} \exp\left(\eta_{x,t}\right) + \phi}\right]^{d_{xt}} \left[\frac{\phi}{e_{xt} \exp\left(\eta_{x,t}\right) + \phi}\right]^{\phi} \right\}, \tag{2.6}$$

with

$$\eta_{x,t} = \alpha_x + \sum_{i=1}^{p} \beta_x^{(i)} \kappa_t^{(i)} + \beta_x^{(0)} \gamma_{t-x}. \tag{2.7}$$

In this section we are concerned with finding the parameters

$$\theta = (\alpha_x, \beta_x^{(0)}, \ldots, \beta_x^{(p)}, \kappa_t^{(1)}, \ldots, \kappa_x^{(p)}, \gamma_{t-x}),$$

in the set of possible parameters Θ, that best explains our data $y = (d_{x,t}, e_{x,t})$, for $x = x_1, \ldots, x_A$ and $t = t_1, \ldots, t_N$.

2.2 Bayesian analysis

Bayesian inference is based on the idea of updating our prior beliefs $p(\theta)$ over θ with the observed data at hand y to come up with posterior beliefs $p(\theta|y)$, see Gelman et al. (1995). By Bayes' theorem, we can determine the posterior distribution of the parameters given the data as follows

$$p(\theta|y) = \frac{l(y|\theta)p(\theta)}{\int_\Theta l(y|\theta)p(\theta)}, \tag{2.8}$$

which in turn allows us to build credible intervals as well as point estimates of the parameters by taking the mean or the mode of the posterior. The integral in the denominator of (2.8) is often analytically intractable due to the high dimension of the parameter space Θ. The usual workaround consists in sampling from the posterior distribution using a Markov Chain Monte Carlo (MCMC) simulation scheme:

$$\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)} \sim p(\theta|y) \propto l(y|\theta)p(\theta).$$

In this paper, we consider five standard mortality models, each implemented in the Negative-Binomial setting with the likelihood (2.6). In Table 1, we specify the predictor $\eta_{x,t}$ entering in the likelihood through (2.7). Hereafter, we discuss the prior distributions of the different parameters.

2.3 Prior distributions

For the choice of the prior distributions, there are essentially two common approaches. The first one specifies diffuse or weakly informative priors such that the posterior inference is dominated by the likelihood of the data, see e.g. Wong et al. (2018). The second one specifies prior distributions which depend on hyperparameters which are estimated by an empirical frequentist approach, see e.g. Czado et al. (2005) and Kogure and Kurachi (2010). In this paper, we follow the first approach.

Table 1: Model structures considered in this paper.

Mortality model	Predictor $\eta_{x,t}$
Lee-Carter (LC)	$\eta_{x,t} = \alpha_x + \beta_x \kappa_t^{(1)}$
Renshaw-Haberman (RH)	$\eta_{x,t} = \alpha_x + \beta_x \kappa_t^{(1)} + \gamma_{t-x}$
Age-Period-Cohort (APC)	$\eta_{x,t} = \alpha_x + \kappa_t^{(1)} + \gamma_{t-x}$
Cairns-Blake-Dowd (CBD)	$\eta_{x,t} = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)}$
M6	$\eta_{x,t} = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + \gamma_{t-x}$

2.3.1 Prior distribution for α_x, β_x and ϕ

Similar to Wong et al. (2018), we assign independent normal priors on α_x, i.e.

$$\alpha_x \sim N(\alpha_0, \sigma_\alpha^2),$$

with $\alpha_0 = 0$ and $\sigma_\alpha^2 = 100$. Because of the constraint $\sum_x \beta_x = 1$, we let the β_x's be Dirichlet distributed with

$$\beta_x \sim \text{Dirichlet}(1, \ldots, 1).$$

Since the model variance is measured by $1/\phi$, see Equation (2.5), the parameter ϕ is actually a concentration parameter for which a standard prior assumption is the half-normal distribution,

$$\frac{1}{\phi} \sim \text{Half-Normal}(0, 1),$$

see for instance Gelman et al. (2006).

2.3.2 Prior distributions for κ_t

For the period indexes we follow the standard actuarial science practice (Cairns et al. (2011b), Cairns et al. (2006), Haberman and Renshaw (2011), Lovász (2011)) and assume that the period indexes follow a multivariate random walk with drift. That is,

$$\boldsymbol{\kappa}_t = \boldsymbol{c} + \boldsymbol{\kappa}_{t-1} + \boldsymbol{\epsilon}_t^\kappa, \quad \boldsymbol{\kappa}_t = \begin{pmatrix} \kappa_t^{(1)} \\ \kappa_t^{(2)} \end{pmatrix}, \quad \boldsymbol{\epsilon}_t^\kappa \sim N(\mathbf{0}, \Sigma), \tag{2.9}$$

where $\boldsymbol{c}$ is a 2-dimensional vector of trend parameters and Σ is a 2×2 variance-covariance matrix of the multivariate white noise $\boldsymbol{\epsilon}_t^\kappa$. For models with a single period effect like LC, RH and APC, the dimension of Equation (2.9) shrinks to one. For the sake of identifiability, we impose $\kappa_1 = 0$ similar to Haberman and Renshaw (2011) and Wong et al. (2018). Under this constraint, the remaining κ_t quantify the mortality improvements relative to the first year while the first year log mortality rates are determined by the α_x's. To complete the model specifications on the $\boldsymbol{\kappa}_t$'s, we set independent normal priors over the regression coefficients $\boldsymbol{c} \sim N(0, 10)$. The variance-covariance matrix of the error term is defined by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_\Sigma \sigma_1 \sigma_2 \\ \rho_\Sigma \sigma_1 \sigma_Y & \sigma_2^2 \end{pmatrix}$$

where the variance coefficients are independent exponentials $\sigma_1, \sigma_2 \sim Exp(0.1)$ and the correlation parameter is uniform $\rho_\Sigma \sim U[-1, 1]$.

2.3.3 Prior distributions for γ_c

For the cohort effect, we consider a second order autoregressive process (AR(2)):

$$\gamma_c = \psi_1 \gamma_{c-1} + \psi_2 \gamma_{c-2} + \epsilon_t^\gamma, \quad \epsilon_t^\gamma \sim N(0, \sigma_\gamma), \tag{2.10}$$

which is in line with previous study conducted by Cairns et al. (2011a) and Lovász (2011). Several model specifications such as AR(1) or ARIMA$(1, 1, 0)$ can be seen as special cases of Equation (2.10). To ensure identifiability, the cohort component is constrained so that the first and last components are equal to 0:

$$\gamma_1 = 0, \quad \gamma_C = 0.$$

For the RH model, we also impose that the sum of effects over the whole range of cohorts is zero[1]:

$$\sum_{i=1}^{C} \gamma_i = 0,$$

where C corresponds to the most recent cohort. These constraints ensure that γ truly represents a cohort effect. Indeed, if the cohort effect presents a trend, this can be compensated by an adjustement to the age and period effects. We close the model specification by imposing some vague priors assumptions on the hyperparameters:

$$\psi_1, \psi_2 \sim N(0, 10), \quad \sigma_\gamma \sim Exp(0.1).$$

2.4 Hamiltonian Monte Carlo and Stan

To produce samples from the posterior distribution, we have implemented our stochastic mortality models using a programming language called *Stan*, see Carpenter et al. (2017). Stan performs an Hamiltonian Monte Carlo (HMC) sampling scheme through the No-U-TurnS (NUTS) algorithm. HMC is a MCMC algorithm where the Metropolis moves are informed by first order gradient information, see Neal (2011). HMC enhances the sampling efficiency and robustness for models with complex posteriors compared to the widely used Metropolis-Hasting within Gibbs sampling scheme.[2] The NUTS algorithm, introduced by Hoffman and Gelman (2014), cope with the difficult choice of the tuning parameters and makes possible the incorporation of the HMC routine into inferencial engines such as Stan. The latter software is gaining popularity among Bayesian statistics practitionners and actuarial scientists, see for instance the work of Gao et al. (2019) and Hilton et al. (2019) where Stan is used for claim reserving and mortality modeling, respectively.

[1] We started our analysis with the constraints on the first and last component and found convergence issues in the Renshaw-Haberman model during our simulation study. Adding the sum-to-zero constraint solved the convergence problem.

[2] Neal (2011) analyzes the scaling benefit of HMC with dimensionality. Hoffman and Gelman (2014) provide practical comparisons of **Stan**'s adaptive HMC algorithm with Gibbs, Metropolis, and standard HMC sample.

Implementation in R We have built our own R package `StanMoMo` which implements the mortality models of Table 1 under the Poisson and the Negative-Binomial setting. It can be downloaded from `https://kabarigou.github.io/StanMoMo`. The package provides high-level R functions to perform Bayesian mortality inference, model selection and model averaging while using `Stan` and HMC sampling in the background.

HMC sampling For each model, four parallel chains are constructed, each of length 4000. The first half of each chain is used as a warm-up round (during which `stan` tunes the algorithm to reflect the characteristics of the posterior) and discarded. Parallel chains are used to better assess the convergence toward the posterior distribution; we follow the diagnostic measure that was advocated by Vehtari et al. (2020) which indicated that all parameters have converged to an acceptable degree. The remainder of all the chains are then gathered and used for inference.

3 Bayesian mortality model averaging

Instead of choosing one model, model averaging stems from the idea that a combination of candidate models among a model list $\mathcal{M} = (M_1, \ldots, M_K)$ may perform better than one single model. The standard Bayesian approach, called *Bayesian model averaging* (BMA), consists in weighing each model by its posterior model evidence. This approach is discussed in subsection 3.1 but should be avoided for mortality forecasting for several reasons. Among them, BMA is very sensitive to prior choices and tends to select only one model asymptotically. Moreover, like the Bayes Information Criterion (BIC), BMA measures how well the model fits the past but not how well the model predicts the future.

We propose two alternative model averaging approaches, called *stacking* and *Pseudo-BMA*, based on leave-future-out and inspired from the work of Yao et al. (2018). These approaches, seemingly more suited for forecasting, are described in subsection 3.2.

3.1 Bayesian model averaging by marginal likelihoods

In the standard BMA approach, each model is weighted by its posterior probability. If y represents the observed data, then the posterior distribution for any quantity of interest Δ (e.g. mortality forecasts) is $p(\Delta \mid y) = \sum_{k=1}^{K} p\left(\Delta \mid M_k, y\right) p\left(M_k \mid y\right)$. In this averaging expression, the weight assigned to each model M_k is its posterior probability:

$$p\left(M_k \mid y\right) = \frac{p\left(y \mid M_k\right) p\left(M_k\right)}{\sum_{k=1}^{K} p\left(y \mid M_k\right) p\left(M_k\right)}, \tag{3.1}$$

which depends itself on the marginal likelihood (ML) under each model:

$$p\left(y \mid M_k\right) = \int_{\Theta} p\left(y \mid \theta_k, M_k\right) p\left(\theta_k \mid M_k\right) d\theta_k. \tag{3.2}$$

Since we typically assume equal prior model probabilities, i.e. $p\left(M_k\right) = \frac{1}{K}$, it remains to compute the MLs for each model. For this, we use bridge sampling as in Meng and Wong

(1996), which is a powerful method for approximating ratio of normalizing constants. The bridge sampling estimator of the marginal likelihood is given by[3]

$$p(y) = \frac{\mathbb{E}_{g(\theta)}[h(\theta)p(y \mid \theta)p(\theta)]}{\mathbb{E}_{p(\theta|y)}[h(\theta)g(\theta)]} \approx \frac{\frac{1}{n_2}\sum_{j=1}^{n_2} h\left(\tilde{\theta}_j\right) p\left(y \mid \tilde{\theta}_j\right) p\left(\tilde{\theta}_j\right)}{\frac{1}{n_1}\sum_{i=1}^{n_1} h\left(\theta_i^*\right) g\left(\theta_i^*\right)} \tag{3.3}$$

where $h(\theta)$ is called the bridge function and $g(\theta)$ denotes the proposal distribution. Here, the sequences $(\theta_i^*)_{i=1,\ldots,n_1}$ and $(\tilde{\theta}_i)_{i=1,\ldots,n_2}$ denote the samples of size n_1 and n_2 drawn, respectively, from the posterior distribution $p(\theta \mid y)$ and $g(\theta)$.

For the choice of the bridge function h, Meng and Wong (1996) showed that the optimal bridge function (in the sense of minimizing the relative mean-squared error of the estimator) is

$$h(\theta) \propto \frac{1}{s_1 p(y \mid \theta)p(\theta) + s_2 p(y)g(\theta)}$$

where $s_i = \frac{n_i}{n_1+n_2}$, for $i \in \{1, 2\}$. We can notice that the ML appears on both sides of Equation (3.3). Therefore, the authors suggested an iterative scheme where the initial guess of the ML is updated until convergence:

$$\hat{p}(y)^{(t+1)} = \frac{\frac{1}{n_2}\sum_{j=1}^{n_2} \frac{l_{2,j}}{s_1 l_{2,j} + s_2 \hat{p}(y)^{(t)}}}{\frac{1}{n_1}\sum_{i=1}^{n_1} \frac{1}{s_1 l_{1,i} + s_2 \hat{p}(y)^{(t)}}},$$

where $l_{1,i} = \frac{p\left(y|\theta_i^*\right)p\left(\theta_i^*\right)}{g\left(\theta_i^*\right)}$, and $l_{2,j} = \frac{p\left(y|\tilde{\theta}_j\right)p\left(\tilde{\theta}_j\right)}{g\left(\tilde{\theta}_j\right)}$.

For the choice of the proposal distribution, following Overstall and Forster (2010) and Gronau et al. (2017), the ideal distribution should resemble the posterior distribution to increase the efficiency of the estimator. Therefore, we choose a multivariate normal distribution with a mean vector and a covariance matrix that match the respective posterior samples quantities. The bridge sampling algorithm is readily available in the R package **bridgesampling**, see Gronau et al. (2020). Once the MLs are obtained for each model, weights are given by the posterior model probabilities in Equation (3.1).

3.2 Bayesian model averaging by stacking and Pseudo-BMA

Bayesian model averaging is flawed in a setting where the "true" data-generating process is not part of the model candidates, see Yao et al. (2018). Indeed, in this setting, BMA asymptotically selects the model in the list which is closest to the real model in the sense of Kullback - Leibler (KL) divergence. More importantly, as we can see from Equation (3.2), that the marginal likelihood is strongly sensitive to the specific prior choice $p\left(\theta_k \mid M_k\right)$ in each model, see Fernandez et al. (2001).

As an alternative approach, different authors considered model selection and averaging based on prediction performance on hold-out data. For instance, Geisser and Eddy (1979)

[3]We omit conditioning on the model for enhanced readibility but it should be kept in mind that this yields the estimate of the ML for a particular model.

proposed to replace marginal likelihoods $p(y \mid M_k)$ with a product of Bayesian leave-one-out cross-validation (LOO-CV) predictive densities $\prod_{i=1}^{n} p(y_i \mid y_{-i}, M_k)$ where y_{-i} is the data without the i-th-point. Roughly speaking, weights are chosen such that the averaged model has the best prediction performance according to a logarithm scoring rule.

In this section, we consider two bayesian model averaging techniques from Yao et al. (2018), namely *stacking* and *Pseudo-BMA*, but adapted to the problem of forecasting mortality. As pointed out by Bürkner et al. (2020), LOO-CV is problematic if the goal is to estimate the predictive performance for future time points. Leaving out only one observation at a time will allow information from the future to influence predictions of the past (i.e., data from times $t+1, t+2, \ldots,$ would inform predictions for time t). Instead, it is more appropriate to use leave-future-out validation. In our context of mortality forecasting, instead of leaving one point out, we leave the last M years of data out and evaluate the prediction performance over these M years.

More precisely, assume that the data for T years is split into a training set and a validation set as follows:

- $y_{1:N} = (d_{x,t}, e_{x,t})$ for all x's and $t = t_1, \ldots, t_N$ are the death and exposure counts of the first N years, used to fit the model.
- $y_{N+1:N+M} = (d_{x,t}, e_{x,t})$ for all x's and $t = t_{N+1}, \ldots, t_{N+M}$ are the death and exposure counts associated to the remaining M years, used to validate the model.

After fitting the NB model to $y_{1:N}$, we can obtain an empirical distribution of future $\mu_{x,t}$ for $t = t_{N+1}, \ldots, t_{N+M}$ based on MCMC samples. Combined with the exposures of the validation set, we can then obtain an empirical distribution of future deaths for each model M_k:

$$D_{x,t} \sim \mathsf{Poisson}(\mu_{x,t}^{k} \cdot e_{x,t})$$

where $\mu_{x,t}^{k}$ are the forecasted mortality rates under model M_k and $e_{x,t}$, for $t = t_{N+1}, \ldots, t_{N+M}$, are the exposures of the validation set. A good averaging approach should aggregate the models such that the resulting model maximizes the likelihood of the observed number of deaths on the validation set. This is the key idea of the stacking of predictive distributions.

3.2.1 Stacking of predictive distributions

The first quantity to determine is the posterior predictive density of future deaths given the training data, i.e. $p(d_{x,j}|y_{1:N})$ for all validation years $j = t_{N+1}, \ldots, t_{N+M}$. These quantities can be computed with the help of the posterior distribution $p(\theta \mid y_{1:N})$ of the parameters θ conditionally to the training dataset for each model M_k. Formally, we have

$$p(d_{x,j} \mid y_{1:N}, M_k) = \int p(d_{x,j} \mid y_{1:N}, \theta, M_k)\, p(\theta \mid y_{1:N}, M_k)\, \mathrm{d}\theta. \tag{3.4}$$

The density (3.4) is analytically intractable but can be approximated based on MCMC samples. Having obtained S draws $\left(\theta^{(1)}, \ldots, \theta^{(S)}\right)$ from the posterior distribution $p(\theta \mid y_{1:N}, M_k)$, we

simply approximate $p\left(d_{x,j} \mid y_{1:N}, M_k\right)$ by

$$p\left(d_{x,j} \mid y_{1:N}, M_k\right) \approx \frac{1}{S} \sum_{s=1}^{S} p\left(d_{x,j} \mid y_{1:N}, \theta^{(s)}, M_k\right).$$

The goal of stacking a set of K predictive distributions built from the models $\mathcal{M} = (M_1, \ldots, M_K)$ is to find the distribution in the convex hull $\mathcal{C} = \left\{\sum_{k=1}^{K} w_k \times p\left(\cdot \mid M_k\right) : \sum_k w_k = 1, w_k \geq 0\right\}$ that is optimal according to some given criterion. In this paper, we follow the approach of Yao et al. (2018) and use a logarithm scoring rule to define the optimality criterion. The weights w_i, $i = 1, \ldots, K$, associated to each mortality model $\mathcal{M}_i \in \mathcal{M}$ follows from solving the optimization problem

$$\max_{w \in \mathcal{S}_1^K} \sum_{x=x_1}^{x_n} \sum_{j=t_{N+1}}^{t_{N+M}} \log \sum_{k=1}^{K} w_k p\left(d_{x,j} \mid y_{1:N}, M_k\right),$$

where

$$\mathcal{S}_1^K = \left\{ w \in [0,1]^K : \sum_{k=1}^{K} w_k = 1 \right\}.$$

The combined predictive distribution is then given by

$$p\left(d_{x,j} \mid y_{1:N}\right) = \sum_{k=1}^{K} w_k p\left(d_{x,j} \mid y_{1:N}, M_k\right).$$

By construction, this averaged distribution maximizes the log likelihood of the observed number of deaths in the validation set among all distributions in the convex hull $\mathcal{C}$.

3.2.2 Pseudo-BMA

As an alternative approach, we consider an AIC-type weighting scheme using leave-future-out validation. To compare the different models, we use the expected log predictive density for each model M_k (elpd^k) as a measure of predictive accuracy, see Vehtari et al. (2017). The elpd^k is defined as follows:

$$\mathrm{elpd}^k = \sum_{x=x_1}^{x_n} \sum_{j=t_{N+1}}^{t_{N+M}} \log p\left(d_{x,j} \mid y_{1:N}, M_k\right). \tag{3.5}$$

Hence, elpd^k is the sum of the point-wise posterior predictive densities over all held-out data points, namely observed deaths for all ages x and all validation years $j = t_{N+1}, \ldots, t_{N+M}$. We can interpret elpd^k as an aggregate measure of how well the model M_k predicts the observed deaths in the validation set. The Pseudo-BMA weight for model M_k is given by

$$w_k = \frac{\exp\left(\mathrm{elpd}^k\right)}{\sum_{k=1}^{K} \exp\left(\mathrm{elpd}^k\right)}.$$

4 Simulation study

A simulation experiment is carried out in order to better understand the behaviour of the selection methods described in Section 3. We take the Belgian mortality data for calendar years from 1959 to 2019 and people aged 50 to 90. A mortality model is fitted to these data and the draws from the posterior distribution are then used to generate 80 synthetic mortality data sets. The dimensions of the synthetic data corresponds exactly to the original mortality data. The various mortality models are fitted to the synthetic datasets, the last ten calendar years of which have been set aside for the evaluation of the out-of-sample forecast error. We want to measure the ability of the selection method to choose the most suitable model. We do this by counting the number of times each model is selected, that is, which model is associated to the highest weight. We also compare the predictive power of the model averaging strategies by calculating the mean absolute error (MAE), based on the test data set, averaged over all ages as

$$\text{MAE} = \frac{1}{40}\sum_{x=50}^{90}\frac{1}{10}\sum_{t=2009}^{2018}|d_{x,t} - e_{x,t}\widehat{\mu}_{xt}|. \tag{4.1}$$

The depth of the data history ranges from 20 up to 50 calendar years. For the pseudo-BMA and stacking approach, we have considered validation sets containing $1, 5$ and 10 calendar years. The split of the data between training, validation and test sets is summarized in Table 2. We

	Fitting	Validation	Prediction
BMA	1989-2009		2010-2019
	1979-2009		-
	1969-2009		-
	1959-2009		-
pseudo-BMA / Stacking	1989-2008	2009	2010-2019
	1989-2004	2005-2009	-
	1989-1999	2000-2009	-
	1979-2008	2009	2010-2019
	1979-2004	2005-2009	-
	1979-1999	2000-2009	-
	1969-2008	2009	2010-2019
	1969-2004	2005-2009	-
	1969-1999	2000-2009	-
	1959-2008	2009	2010-2019
	1959-2004	2005-2009	-
	1959-1999	2000-2009	-

Table 2: Simulation experiments and time-line assumptions

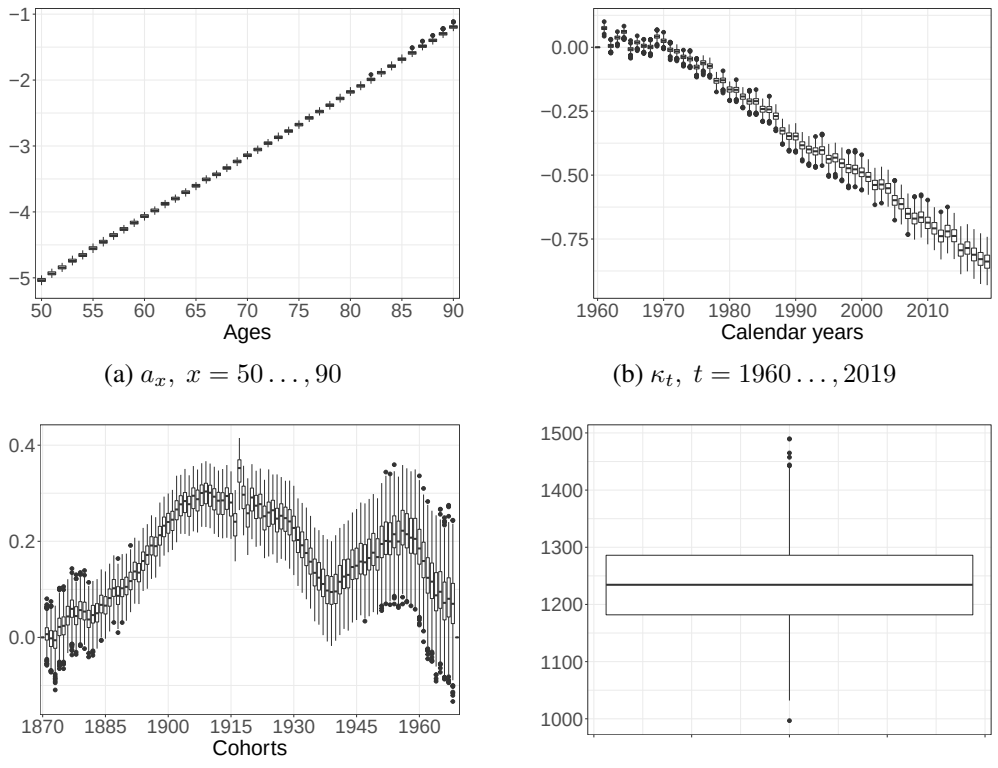
have noticed that a validation set containing only one calendar year is insufficient, hence the results are not reported for brevity. The difference between having 5 or 10 years in the validation

set is so small that we only report the results associated with a validation set containing 10 years of data. Note that it is consistent with the size of the test dataset. We consider two cases:

- In the first one, the data is generated by an Age-Period-Cohort model. The true model is among the competing models and the results are given in subsection 4.1.
- In the second case, the numbers of death result from taking the average of death counts drawn from a Cairns-Blake-Dowd model and from a Renshaw-Haberman model. The true model is not among the competing models, which brings us closer to a real situation. The results are discussed in subsection 4.2.

4.1 Data generated by an Age-Period-Cohort model

The APC model is fitted to the Belgian mortality data for calendar years from 1959 to 2019 and people aged 50 to 90, the posterior distribution of the parameters is provided on Figure Figure 1. Based on the posterior draws, 80 synthetic mortality datasets are generated to which are fitted



(a) $a_x,\ x = 50 \ldots, 90$

(b) $\kappa_t,\ t = 1960 \ldots, 2019$

(c) $g_{t-x}, x = 50 \ldots, 90; t = 1960 \ldots, 2019$

(d) ϕ

Figure 1: Posterior distribution of the parameters of the APC model that generated the synthetic data.

the mortality models including LC, CBD, APC, RH and M6. The synthetic data provided to the mortality models only contain the calendar years from 1960 to 2010, the remaining ten years

are kept as a test set to assess the predictive power through the mean absolute error defined in (4.1). The validation set for the pseudo-BMA and stacking methods contains 10 years of data. A mortality model is deemed selected if it gets the highest weight according to a given model averaging method. Figure 2 shows how many times each mortality model got selected according to each selection method as a function of the number of calendar years in the training dataset. We note that all the methods discard the CBD and LC models as they do not account for

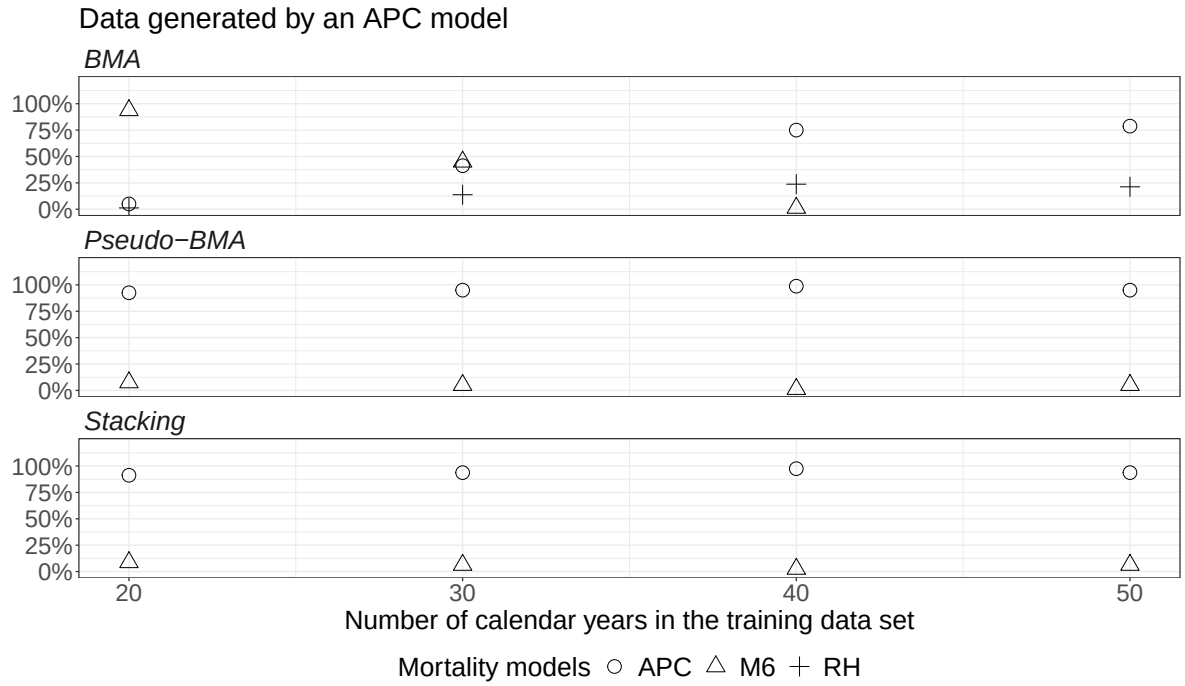


Figure 2: How many times each mortality model is selected, depending on the method, out of 80 synthetic data sets generated by an Age-Period-Cohort model as a function of the number of calendar years in the training set.

the cohort effect. The pseudo-BMA and stacking methods clearly favor the APC model, it takes a few more calendar years for the standard BMA approach to choose it. Figure 3 displays the mean absolute error of the prediction resulting from the mortality models and their combination via the different methods of model aggregation as a function of the number of calendar years in the training dataset. As expected, the best prediction is provided by the APC model while the predictions made by the CBD and LC models are quite flawed. Since stacking and pseudo-BMA tend to always choose the APC model, their use leads to a slight improvement in predictions over the BMA approach. We note that the prediction error is slightly higher when 50 calendar years are included in the training data set. This might seem counterintuitive as one would expect that the more data there is, the better the prediction. This is not true when studying mortality. Years too far from the projection horizon are irrelevant for understanding current trends in mortality and, more importantly, taking them into account appears to degrade the forecast. If 20 calendar years seem sufficient to make reasonable predictions, taking 30 or 40 calendar years widens the gap between the prediction errors resulting from models that encapsulate a cohort effect and those that do not. The case studied in this section corresponds to a situation where

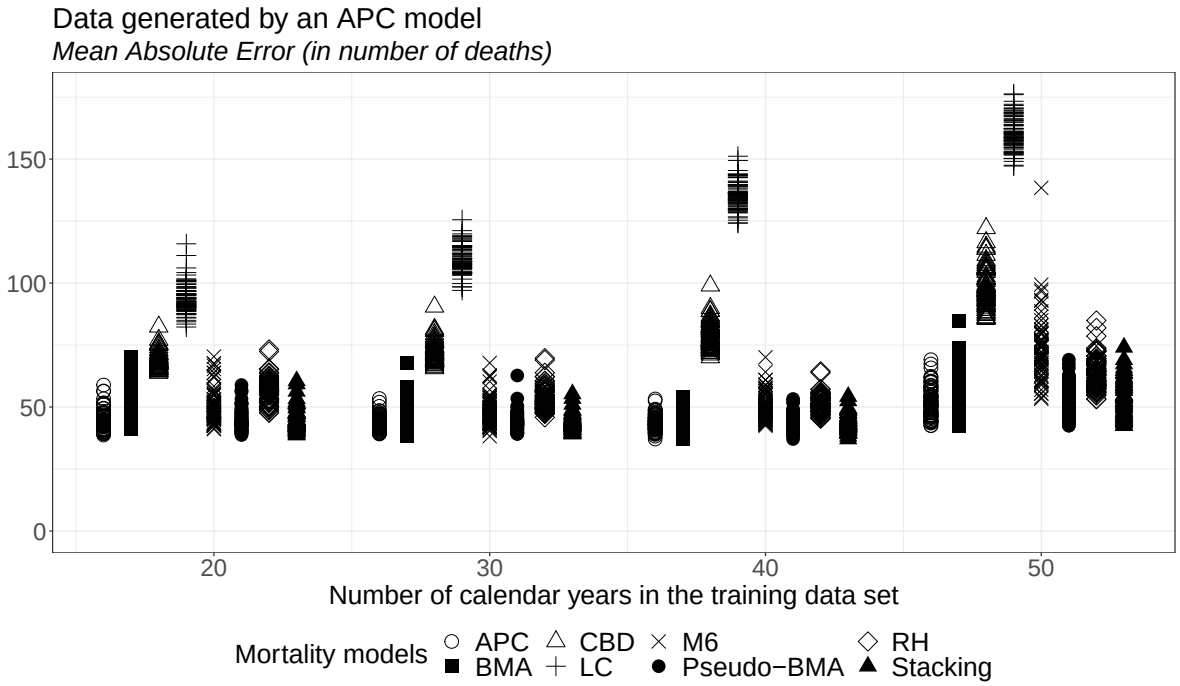


Figure 3: Mean absolute errors calculated over 80 simulated datasets from an APC model as a function of the number of calendar years in the training dataset.

the model is well specified because the APC model belongs to the competing models. The next section will allow us to see whether these results are also valid in a miss-specified case.

4.2 Data generated by a mixture between a CBD and RH model

The same Belgian mortality data set is used to fit the CBD and RH models. Each model is used to generate 80 synthetic mortality data sets. The synthetic data sets are combined in pairs by taking the average number of deaths. We then fit the mortality models to these hybrid mortality data (without the last ten years that will be used to measure the out-of-sample error) and apply the different models averaging strategies. Figure 4 shows how many times each mortality model got selected according to each selection method as a function of the number of calendar years in the training dataset. The BMA approach favors the M6 model but also chooses from time to time the RH and APC models. The stacking and pseudo-BMA techniques clearly side for the APC model and sometime select the M6 model. Let us see what it means for the prediction errors. Figure 5 displays the mean absolute error of the prediction resulting from the mortality models and their combination via the different methods of model aggregation as a function of the number of calendar years in the training dataset. The APC model returns the smallest prediction error and the same goes for stacking and pseudo-BMA approaches which tend to give the APC model a lot of credibility. Again, taking 50 calendar years is detrimental to the accuracy of the forecast. This study demonstrates the good behavior of the Bayesian model averaging methods in a controlled environment (the data generation process being specified by

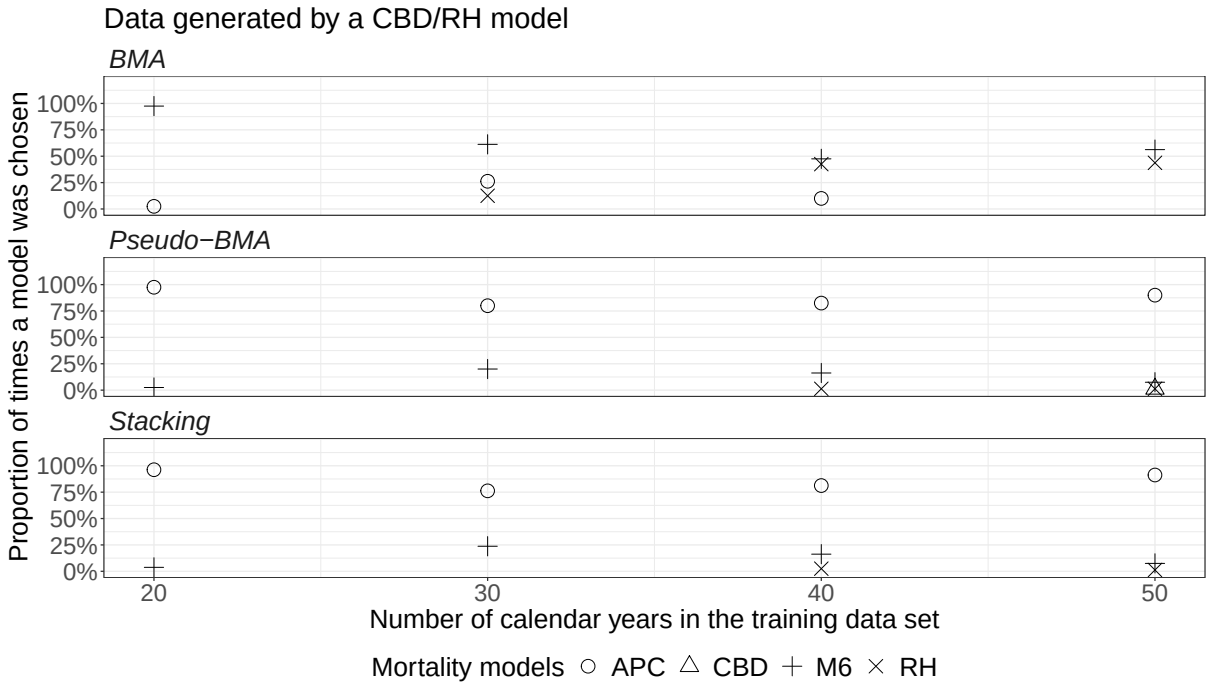


Figure 4: How many times each mortality model is selected, depending on the method, out of 80 synthetic data sets generated by a a mixture of a CBD model and a RH model as a function of the number of calendar years in the training set.

us). The following section is devoted to the application to actual mortality data sets.

5 Application to real mortality data

In this section we apply the three model averaging approaches discussed in Section 3 to mortality data from France, UK, USA and Japan. The data chosen for illustrative purposes are the male death data and the corresponding exposures of these four countries, for ages $50-90$ and the last 40 years of data available (1979-2018) extracted from the Human Mortality Database (HMD)[4]. To assess the prediction performance, we split the data into two parts: the first 30 years are used for the weights selection (1979-2008) and the last 10 years (2009-2018) are used to compare the weighted forecasts. For the calculation of the stacking and pseudo-BMA weights, the data is then divided into two parts: the first 20 years are used as a training set while the remaining 10 years are used for validation. The size of the leave-future-out validation set is consistent with the findings of Section 4. The data partitions associated to each model averaging method are given in Table 3.

[4]See www.mortality.org.

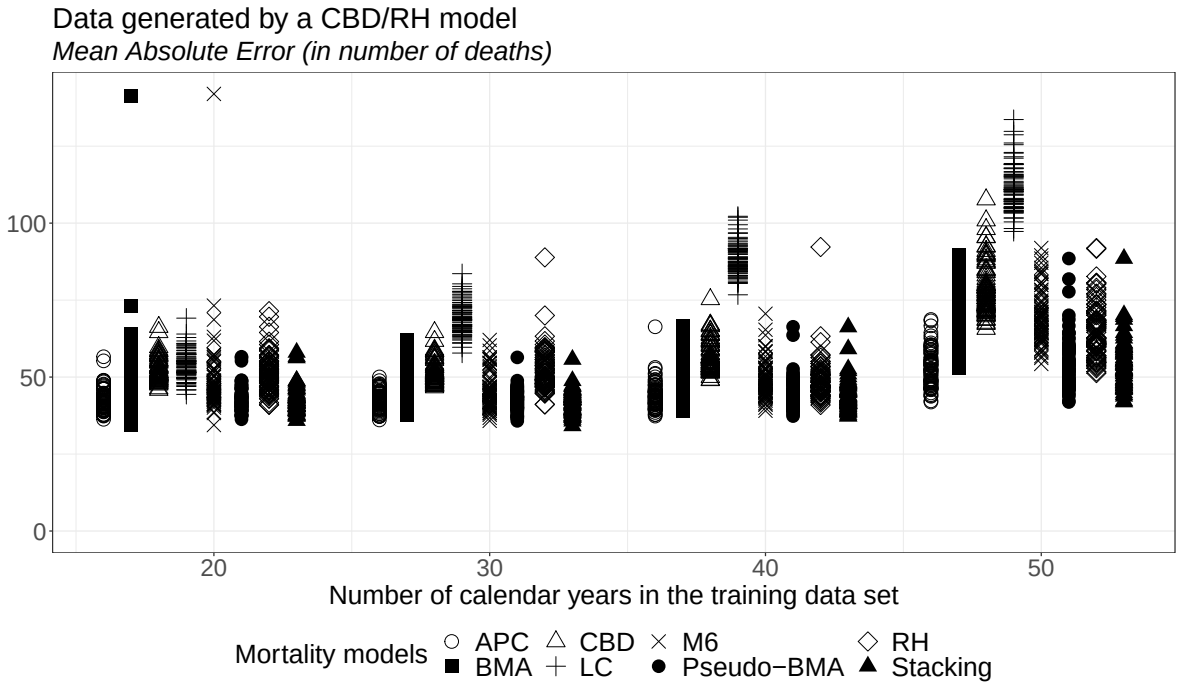


Figure 5: Average absolute errors calculated over 80 data sets simulated from a mixture of a CBD model and a RH model as a function of the number of calendar years in the training data set.

Table 3: Fitting, validation and prediction periods for the three model averaging approaches.

	1979-1998	1999-2008	2009-2018
BMA	Fitting		Prediction
Stacking	Fitting	Validation	
Pseudo-BMA	Fitting	Validation	

5.1 Model Weights

Table 4 provides the weights obtained via standard BMA (marginal likelihood), stacking and pseudo-BMA for France, UK, USA and Japan. The BMA and pseudo-BMA approaches tend to only select one model. This was expected given the size of the dataset (see Yao et al. (2018) and the references therein). On the other hand, the stacking approach selects two models for UK and USA and three models for France and Japan. Overall, we observe a certain agreement between the stacking and pseudo-BMA approaches based on validation while the BMA and pseudo-BMA do not always select the same model. We also note that the standard BMA approach favors either the RH model or the M6 model[5]; this is in agreement with the frequentist literature in which the RH model or the CBD with cohort effect have been often identified as the best

[5]We note that the model selection via BMA is sensitive to the sample period used to fit the models. For a 20-year fitting period (1979-2008), we found that the M6 model was selected for France and UK, and the APC

Table 4: Model Weights for France, UK, USA and Japan via BMA, stacking and pseudo-BMA.

	France			UK		
	BMA	Stacking	Pseudo-BMA	BMA	Stacking	Pseudo-BMA
LC	0	0.093	0	0	0	0
RH	1	0.750	1	0	0.342	0
APC	0	0.157	0	0	0	0
CBD	0	0	0	0	0	0
M6	0	0	0	1	0.658	1
	USA			Japan		
	BMA	Stacking	Pseudo-BMA	BMA	Stacking	Pseudo-BMA
LC	0	0	0	0	0.292	0
RH	1	0.548	0.982	1	0.239	0
APC	0	0.452	0.018	0	0.468	1
CBD	0	0	0	0	0	0
M6	0	0	0	0	0	0

candidate model when model selection is based on the BIC or AIC criterion, see Cairns et al. (2009) and Haberman and Renshaw (2011) among others.

5.2 Prediction performance

To assess the prediction performance of the three Bayesian model averaging approaches, we first compute the 95% credible intervals of the projected log death rates for age $x = 65, 75, 85$ as a function of time, 10 years into the future, along with the observed crude death rates as shown in Figure 6. An ideal credible interval should be sufficiently large to contain the observed death rates of the next 10 years but not too wide to avoid overconservative credible intervals. We note the following:

- For France, the three methods provide reasonable and similar credible intervals at age 85. However, at age 75 and age 65, whatever the approach, the intervals seem to be too narrow as the last death rates tend to fall outside the confidence bands.
- For the UK, we also observe that the intervals are too narrow at age 85 while the observed death rates fall right inside the intervals at age 65 and 75.
- For the USA, the model averaging methods fail to match the observed death rates at age 75 as they lie outside the confidence interval.
- For Japan, we observe that for the ages considered, the stacking approach better projects the future death rates. Indeed, the credible intervals of the stacking approach encompass

model for USA and Japan. The sensitivity of mortality models to the sample period has been extensively studied and we refer to Cairns et al. (2011a) among others.

the observed death rates. This example shows that a linear combination of three mortality models can outperform a single model in terms of 10-year prediction performance.

We now study the performance of the models when estimating mortality indicators that aggregate all ages. A common quantity is the life expectancy at birth but it would require the full age range. Since we focus on the age range $50 - 90$, we instead compute a 40-year period survival probability of a person of age $x = 50$ for any year t:

$$ {}_{40}p_{50,t} = \prod_{i=0}^{39} p_{50+i,t} = \prod_{i=0}^{39} \exp\left(-\mu_{50+i,t}\right). \tag{5.1} $$

It corresponds to the probability that a 50 years old person to live for more than 40 years given the mortality conditions at year t. On Figure 7, we have plotted the 95% credible intervals of the period survival probabilities for the 10-year period $2009 - 2018$, along with the observed quantities. For France, the holdout survival probabilities lie within the 95% prediction intervals of the three model averaging approaches. However, for the UK, the stacking approach overestimates the survival probabilities while the BMA and Pseudo-BMA approaches manage to get the observed quantities in their prediction intervals. For Japan, the intervals obtained by stacking seem to be slightly too narrow as the first holdout points lie outside the prediction intervals. For the four countries considered, the BMA and Pseudo-BMA slightly outperfoms the stacking approach by providing wider confidence intervals for the survival probability.

To close, we also assess the predictive performance by age for each country through the Mean Absolute Error (MAE) over the years in the test set:

$$ \text{MAE}_x = \frac{1}{10} \sum_{t=2009}^{2018} |d_{x,t} - e_{x,t}\widehat{\mu}_{xt}|, \quad x = 50, \ldots, 90, $$

where $\widehat{\mu}_{xt}$ is the posterior mean of the forecasted death rates. Figure 8 shows the MAE by age for France, the UK, the USA and Japan according to the standard BMA, stacking and pseudo-BMA. Shifting from BMA to stacking or pseudo-BMA, a large improvement in the forecasts accuracy is obtained, especially for the ages 70 to 90. In particular, for France and USA, the MAE levels clearly lie below the MAE levels of BMA and Pseudo-BMA. For the UK, the performances of the three methods are close for the ages 50 to 80 but the stacking approach leads to better MAEs after age 80. For Japan, the comparison is not obvious. However, we did compute the overall MAE across ages and years and found for Japan:

$$ \text{MAE (BMA)} = 579.30.7629, \quad \text{MAE (stacking)} = 489.74, \quad \text{MAE (Pseudo-BMA)} = 758.76 $$

Hence, at the aggregate level, stacking still provides a better forecast performance than BMA, even for Japan.

Overall, this validation exercise shows that the Bayesian model averaging methods based on 10-year future-out validation (namely stacking and Pseudo-BMA) tend to outperform the standard BMA approach in terms of the ability to predict 10-year ahead for the four countries considered here. Intuitively, this means that a model which provided good forecasts for the last 10 years has a good chance to perform well for the following 10 years. On the other hand, a model that fits well the mortality data has no a priori reason to be good at *forecasting* mortality data. Methods based on future-out validation (stacking and pseudo-BMA) should therefore be preferred to methods based on goodness-of-fit (standard BMA) for forecasting purposes.

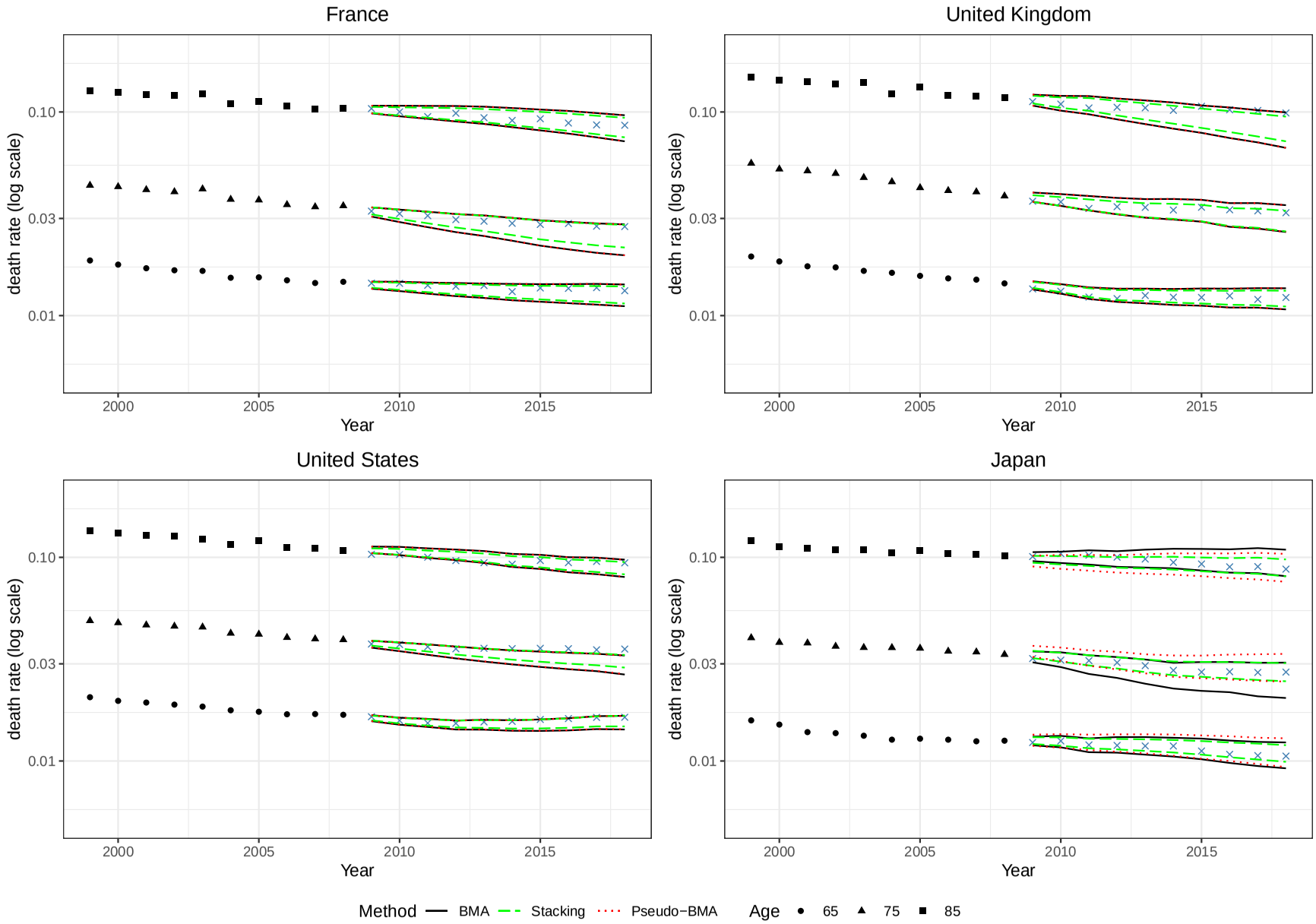


Figure 6: 95% prediction intervals for the death rates for age $x = 65, 75, 85$ via the three model averaging approaches along with the observed crude deaths rates from France, UK, USA and Japan for the 10-year period 2009-2018.

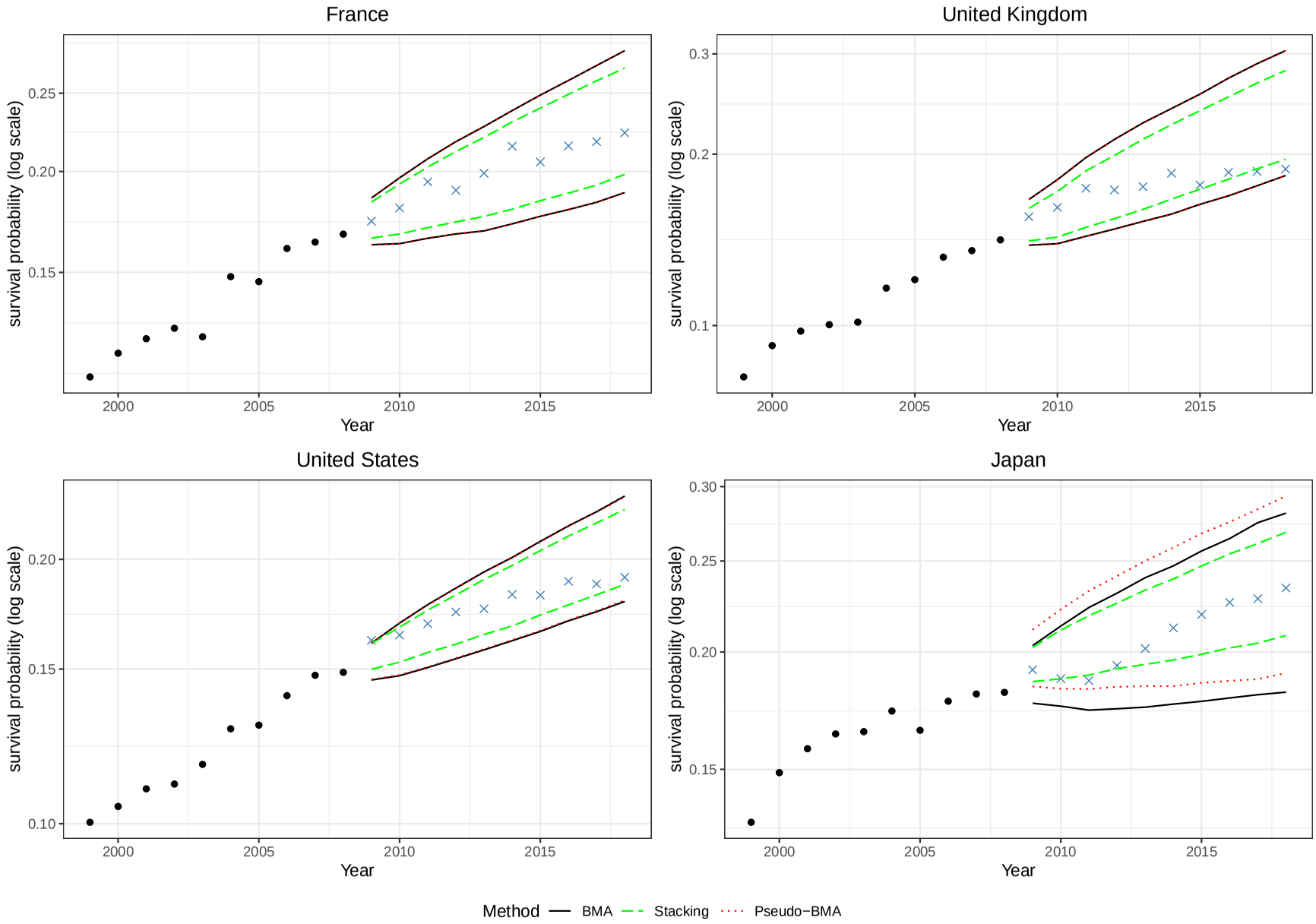


Figure 7: 95% prediction intervals for period survival probability at age 50 until 90 via the three model averaging approaches along with the observed period survival probabilities from France, UK, USA and Japan for the 10-year period 2009-2018.

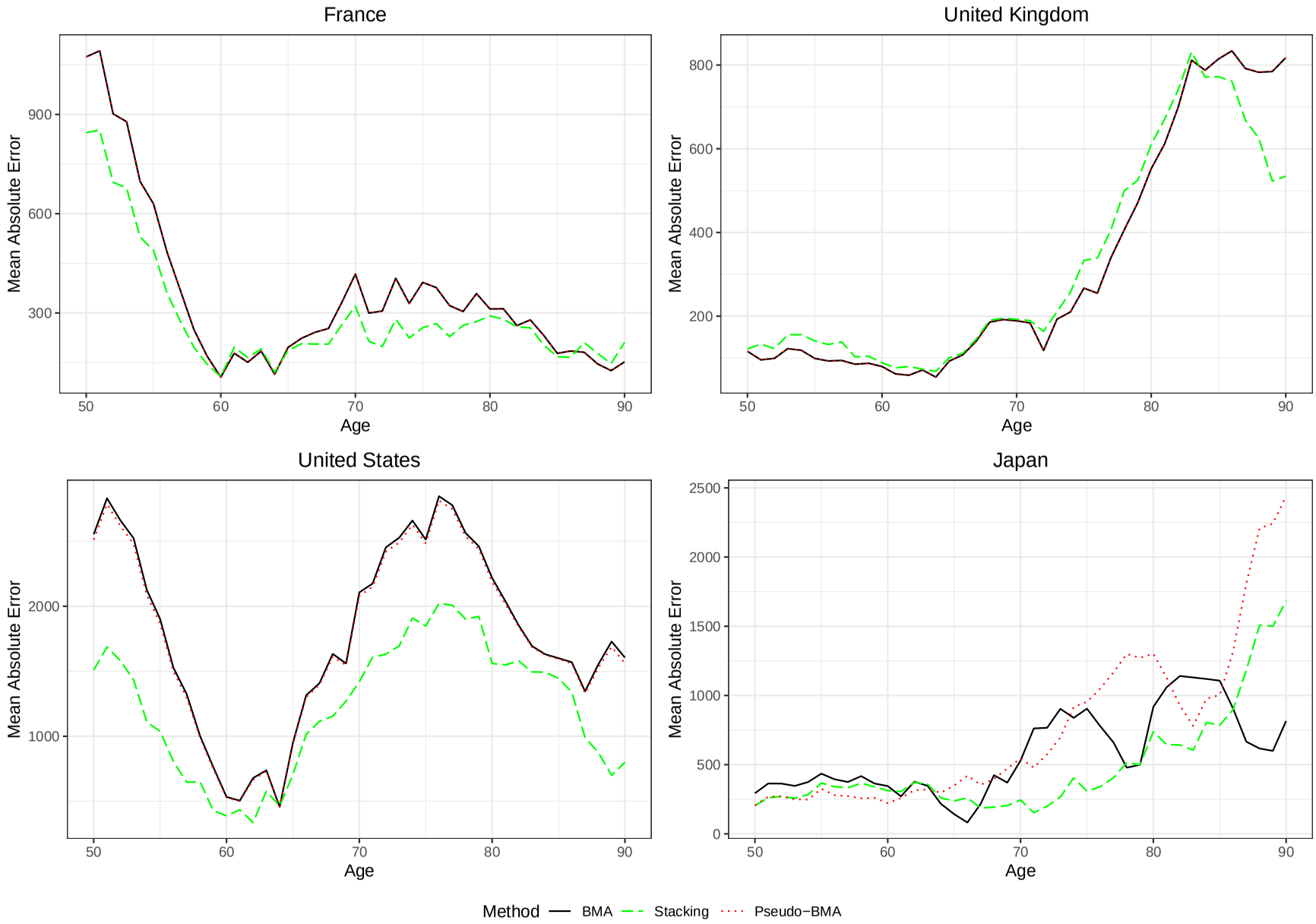


Figure 8: Mean Absolute Error per age (50-90) averaged across years (2009-2018).

6 Impact of Covid-type effect on mortality forecasting

In the context of the recent Covid-19 pandemic, it is important to determine how mortality models and forecasts react to a pandemic shock. In the following, we have perturbed the French male data with two years of excess mortality followed by one year of lower mortality, and assessed the impact in terms of model averaging weights and life expectancy. This pandemic scenario is in the spirit of Cairns et al. (2020) who proposed an accelerated deaths model to explore the impacts of the pandemic on life expectancy. The authors argue that "*many of those who die from coronavirus would have died anyway in the relatively near future due to their existing frailties or co-morbidities. Therefore, the life expectancy of the surviving population might slightly increase compared to their pre-pandemic levels*". For this reason, we do compensate two years of excess mortality by a slight decrease in mortality in the third year.

We take the male death data for France until year 2018 from the Human Mortality Database, and perturb the death counts associated to the remaining three years as follows:

- For the years 2016 and 2017, we assume that there is a uniform death increase of 5% across all ages:
$$d_{x,t}^{\text{new}} = (1 + \beta(x))d_{x,t},$$
with $\beta(x) = 0.05$ for all $x \in \{50, \ldots, 90\}$, with $t = 2016, 2017$.
- The increase in deaths is then compensated with a year of lower mortality. We assume a death decrease of 2% across ages:
$$d_{x,t}^{\text{new}} = (1 - \beta(x))d_{x,t},$$
with $\beta(x) = 0.02$ for all $x \in \{50, \ldots, 90\}$, with $t = 2018$.

First, we derive the weights associated to the standard BMA (marginal likelihood), stacking and pseudo-BMA approaches based on 40 years of data (1979-2018) including 10 validation years (2009-2018) with and without perturbations. Different observations can be drawn from the results in Table 5. For BMA and Pseudo-BMA, the perturbations do not affect the weights: the Renshaw-Haberman model is chosen by the BMA approach and the APC model is favored by the Pseudo-BMA approach. For the stacking approach, we observe some slight changes in the weights. With the perturbations, some weight is given to the Lee-Carter model and the stacking approach therefore averages over three models (LC, RH and APC). Moreover, we note that the weights obtained in Table 5 are different from the ones obtained in Table 4 in the previous section since the validation and calibration periods are different. For instance, for Pseudo-BMA, RH was chosen for the validation period 1999-2008 while APC was the selected model for the validation period 2009-2018.

To measure the effect on life expectancy, and since we focus on the age range $50 - 90$, we compute the life expectancy at age 50 truncated at age 90 for the next 10 years (2019-2028), that is

$$e_{50:\overline{40}|,t} = \sum_{k=1}^{40} {}_kp_{50,t} \tag{6.1}$$

Table 5: Model Weights for France with and without Covid-type effect.

	BMA		Stacking		Pseudo-BMA	
Perturbations	Without	With	Without	With	Without	With
LC	0	0	0	0.153	0	0
RH	1	1	0.21	0.14	0	0
APC	0	0	0.79	0.698	1	1
CBD	0	0	0	0	0	0
M6	0	0	0	0	0	0

where ${}_kp_{50,t}$ is the k-year survival probability at year t just like in Equation (5.1). We note that (6.1) can be interpreted as the average number of payments of a life annuity at age 50 that ends at age 90 since

$$e_{50:\overline{40|}} = \mathbb{E}\left[\min\left(K_{50}, 40\right)\right] \tag{6.2}$$

where K_{50} is the number of years lived by a person aged 50 (see for instance Section 2.6 in Dickson et al. (2013)). In Figure 9, we plot the life expectancies, observed and predicted, from 2009 to 2018, according to each model averaging method. In order to better assess the impact of the perturbations on the overall uncertainty, we show the predictions of the Lee-Carter model *without* perturbations. First, we observe that the perturbed data via all three approaches produce larger confidence intervals compared to the baseline LC model without perturbations as one would expect. Indeed, the perturbations increase the volatility of the period effects $\boldsymbol{\kappa}_t$ and therefore the uncertainty in future life expectancies. Second, we note that the leave-future-out based methods (stacking and Pseudo-BMA) forecast an increase in life expectancy (which is consistent with historical trend) while the BMA approach (and therefore the RH model) tends to forecast a rather constant and even decreasing life expectancy.

The median life expectancy with and without Covid-type effect is plotted on Table 6. With the perturbations, the median life expectancy increases between the years 2019 and 2028 by more than a year for stacking and pseudo-BMA but only by 0.34 for BMA. More surprisingly, the BMA approach (i.e. the RH model in this case) forecasts a decreasing life expectancy *in case of no perturbations*. To determine the reason of this downward trend, we plot the forecast life expectancy in Figure 10. The BMA approach (which only selects the RH model) leads to implausible forecasts, in contradiction with historical data. We found that these forecasts are the consequence of the erratic behaviour of the cohort effect in the RH model. Indeed, Figure 11 depicts the 95% prediction intervals for the RH and APC cohort effects, models chosen by BMA and pseudo-BMA, respectively. We get implausible forecast values with the RH model while the APC model produces reasonable cohort parameter values in the range $[-0.5; 0.5]$. These high cohort parameter values for the RH model increase the forecast force of mortality and therefore decrease the life expectancy as depicted in Figure 10.

Overall, we find that the stacking and pseudo-BMA approaches provide more realistic forecasts in the presence of a Covid-type effect in the data. Both approaches predict an increase in life expectancy which is consistent with the historical trend and the 5% decrease in the number of deaths associated to the *compensation effect* of the pandemic. The fact that the BMA ap-

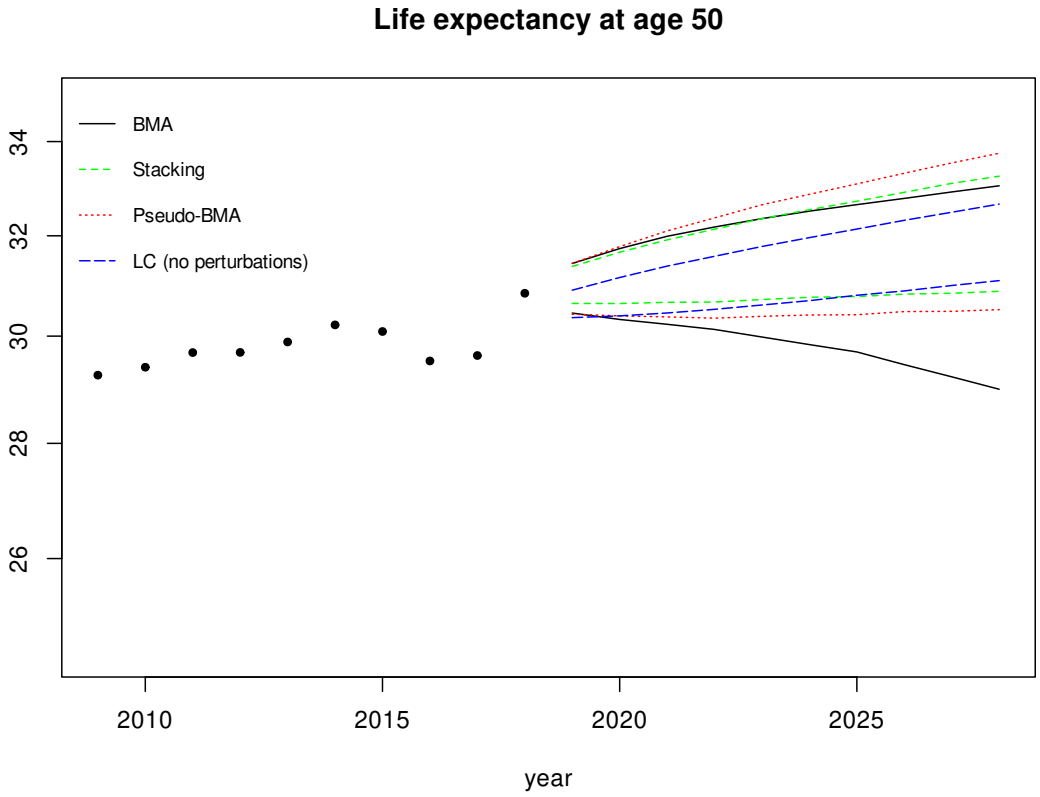


Figure 9: 95% prediction intervals for the life expectancy at age 50 (truncated at age 90) for the 10-year period 2019-2028 via the three model averaging approaches with perturbed data. For comparison, we also provide the 95% prediction intervals via the Lee-Carter (LC) model without perturbations.

proach selects the RH model despites the erratic behaviour of the cohort effect is problematic as it results in implausible and wide prediction intervals, see Figure 11.

7 Conclusion

In this work, we address the problem of stochastic mortality model averaging. We start by setting up an attractive Bayesian modeling framework because it allows us to consider several mortality models and to account for the uncertainty around the parameter estimates. Model averaging strategies are then applied to mitigate the risk of selecting the wrong model. The standard Bayesian model averaging, based on how well the model fits the training dataset is challenged by two other model averaging strategies, refered to as stacking and pseudo-BMA, that focus on the out-of-sample error.

We recommend the use of the leave-future-out based model averaging approaches for the pur-

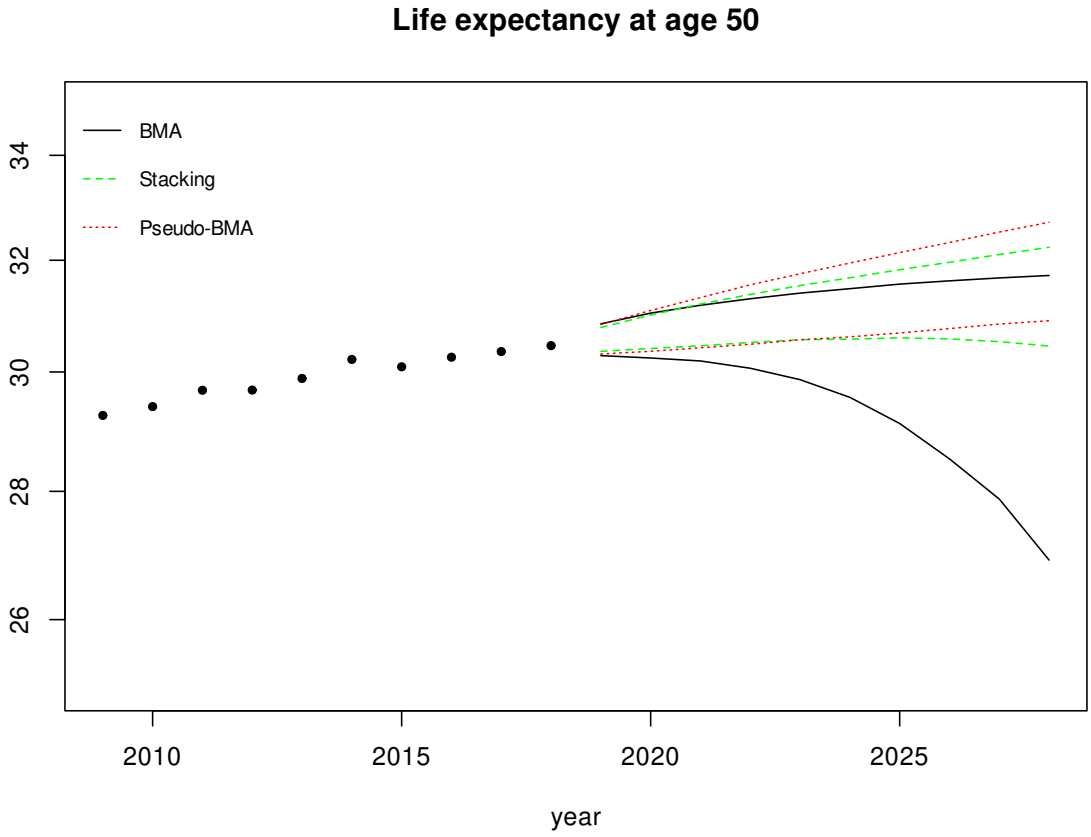


Figure 10: 95% prediction intervals for the life expectancy at age 50 (truncated at age 90) via the three model averaging approaches with no perturbations for the 10-year period 2019-2028 (French Males).

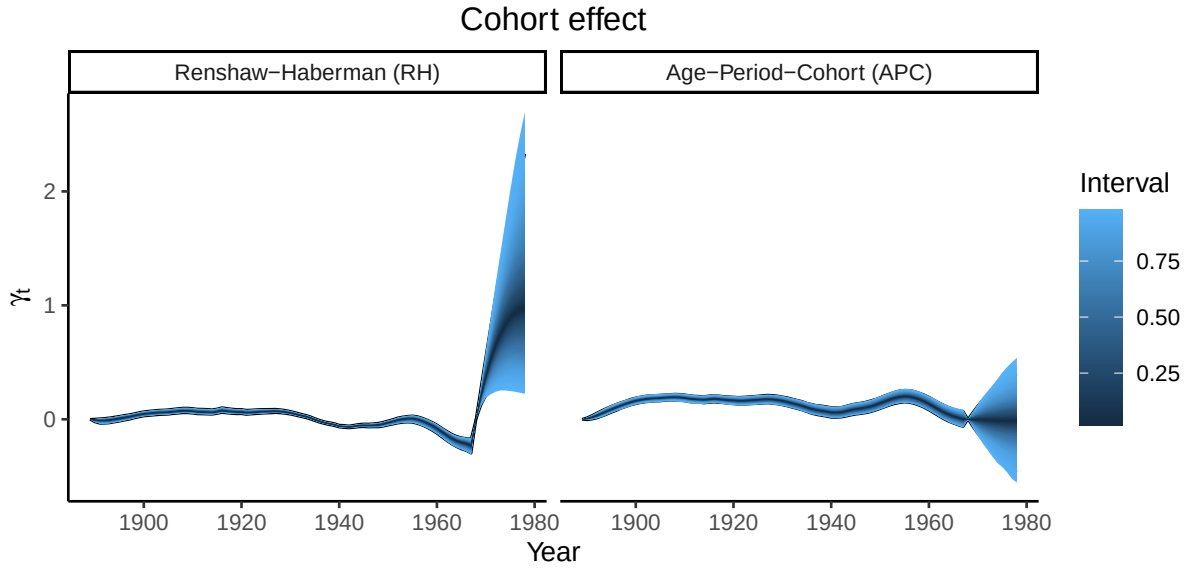


Figure 11: 95% prediction intervals for the cohort parameter γ_t in the RH and APC models. Parameters in the RH model present a clear erratic behaviour while the ones in the APC model are more stable.

Table 6: Median Life Expectancy at age 50 (truncated at age 90) for French Male with and without Covid-type effect.

	BMA		Stacking		Pseudo-BMA	
Perturbations	Without	With	Without	With	Without	With
2019	30.56	30.95	30.57	31.01	30.57	30.94
2020	30.64	31.05	30.70	31.15	30.72	31.09
2021	30.69	31.13	30.83	31.30	30.87	31.26
2022	30.70	31.19	30.95	31.43	31.02	31.41
2023	30.68	31.23	31.06	31.57	31.16	31.56
2024	30.64	31.25	31.15	31.69	31.30	31.71
2025	30.57	31.27	31.24	31.81	31.43	31.85
2026	30.49	31.28	31.32	31.93	31.57	31.98
2027	30.41	31.29	31.40	32.04	31.70	32.12
2028	30.32	31.29	31.47	32.16	31.83	32.26

pose of forecasting mortality trends. Our study draws on extensive simulation study and applications to real-world mortality data sets (with and without COVID-like disruption).

This work could be extended in many interesting ways. First, the validation technique could be adapted to the case where the mortality patterns exhibit a change of regime. In fact, as discussed with the COVID-type impact, the model averaging approach should assign more weights to models that are not only good at representing the past but also at forecasting the future. Here, we should introduce some potential regime switching techniques into the considered models in order to tackle such a problem. However, this interesting problem is beyond the scope of the current paper and will be investigated in a future work. Finally, given the ability of the averaging techniques to accommodate classic and most used models, an R package implementing the three model averaging approaches is available for download to researchers as well as practitioners at https://kabarigou.github.io/StanMoMo.

Acknowledgment

This work was supported by the Joint Research Initiative on "Mortality Modeling and Surveillance" funded by AXA Research Fund. S. Loisel and Y. Salhi also acknowledge support from the BNP Paribas Cardif Chair "New Insurees, Next Actuaries" (NINA) and the Milliman research initiative "Actuariat Durable". Y. Salhi benefited from the support of the CY Initiative of Excellence (grant "Investissements d'Avenir" ANR-16-IDEX-0008), Project "EcoDep" PSI-AAP2020–0000000013.

References

Antonio, K., Bardoutsos, A., and Ouburg, W. (2015). Bayesian poisson log-bilinear models for mortality projections with multiple populations. *European Actuarial Journal*, 5(2):245–281.

Benchimol, A. G., Diazaraque, J. M. M., Lozano, I. A., and Alonso-González, P. J. (2018). Mortality projection using bayesian model averaging. In *Mathematical and Statistical Methods for Actuarial Sciences and Finance*, pages 111–115. Springer.

Bürkner, P.-C., Gabry, J., and Vehtari, A. (2020). Approximate leave-future-out cross-validation for bayesian time series models. *Journal of Statistical Computation and Simulation*, 0(0):1–25.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, pages 603–618.

Cade, B. S. (2015). Model averaging and muddled multimodel inferences. *Ecology*, 96(9):2370–2382.

Cairns, A. J., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., and Khalaf-Allah, M. (2011a). Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, 48(3):355–367.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from england and wales and the united states. *North American Actuarial Journal*, 13(1):1–35.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., and Khalaf-Allah, M. (2011b). Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin: The Journal of the IAA*, 41(1):29–59.

Cairns, A. J., Blake, D. P., Kessler, A., and Kessler, M. (2020). The impact of covid-19 on future higher-age mortality. *Available at SSRN 3606988*.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Currie, I. D. (2016). On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal*, 2016(4):356–383.

Czado, C., Delwarde, A., and Denuit, M. (2005). Bayesian poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics*, 36(3):260–284.

Delwarde, A., Denuit, M., and 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.

Dickson, D. C., Hardy, M., Hardy, M. R., and Waters, H. R. (2013). *Actuarial mathematics for life contingent risks*. Cambridge University Press.

Fernandez, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for bayesian model averaging. *Journal of Econometrics*, 100(2):381–427.

Gao, G., Meng, S., and Shi, Y. (2019). Stochastic payments per claim incurred. *North American Actuarial Journal*, 23(1):11–26.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534.

Gronau, Q., Singmann, H., and Wagenmakers, E.-J. (2020). bridgesampling: An r package for estimating normalizing constants. *Journal of Statistical Software, Articles*, 92(10):1–29.

Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., Leslie, D. S., Forster, J. J., Wagenmakers, E.-J., and Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of mathematical psychology*, 81:80–97.

Haberman, S. and Renshaw, A. (2011). A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48(1):35–55.

Hilton, J., Dodd, E., Forster, J. J., and Smith, P. W. (2019). Projecting uk mortality by using bayesian generalized additive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(1):29–49.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623.

Hunt, A. and Blake, D. (2020). On the structure and classification of mortality models. *North American Actuarial Journal*, 0(0):1–20.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kogure, A. and Kurachi, Y. (2010). A bayesian approach to pricing longevity risk based on risk-neutral predictive distributions. *Insurance: Mathematics and Economics*, 46(1):162–172.

Koop, G. and Korobilis, D. (2012). Forecasting inflation using dynamic model averaging. *International Economic Review*, 53(3):867–886.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671.

Lovász, E. (2011). Analysis of finnish and swedish mortality data with stochastic mortality models. *European Actuarial Journal*, 1(2):259–289.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860.

Neal, R. M. (2011). Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2.

Overstall, A. M. and Forster, J. J. (2010). Default bayesian model determination methods for generalised linear mixed models. *Computational Statistics & Data Analysis*, 54(12):3269–3288.

Pedroza, C. (2006). A bayesian forecasting model: predicting us male mortality. *Biostatistics*, 7(4):530–550.

Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393–404.

Renshaw, A. E. and Haberman, S. (2006). A cohort-based extension to the lee–carter model for mortality reduction factors. *Insurance: Mathematics and economics*, 38(3):556–570.

Shang, H. L. (2012). Point and interval forecasts of age-specific life expectancies: A model averaging approach. *Demographic Research*, 27:593–644.

Shang, H. L. and Haberman, S. (2018). Model confidence sets and forecast combination: an application to age-specific mortality. *Genus*, 74(1):19.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., Bürkner, P.-C., et al. (2020). Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc. *Bayesian Analysis*.

Venter, G. and Şahın, Ş. (2018). Parsimonious parameterization of age-period-cohort models by bayesian shrinkage. *ASTIN Bulletin: The Journal of the IAA*, 48(1):89–110.

Wong, J. S., Forster, J. J., and Smith, P. W. (2018). Bayesian mortality forecasting with overdispersion. *Insurance: Mathematics and Economics*, 83:206–221.

Yao, Y., Vehtari, A., Simpson, D., Gelman, A., et al. (2018). Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1007.