risks

MDPI

Article

Two-Population Mortality Forecasting: An Approach Based on Model Averaging

Luca De Mori [1,*], Pietro Millossovich [1,2], Rui Zhu [1] and Steven Haberman [1]

[1] Bayes Business School, City, University of London, London EC1Y 8TZ, UK; pietro.millossovich.1@city.ac.uk (P.M.); rui.zhu@city.ac.uk (R.Z.); s.haberman@city.ac.uk (S.H.)
[2] DEAMS, University of Trieste, 34127 Trieste, Italy
* Correspondence: luca.de-mori@bayes.city.ac.uk

Abstract: The analysis of residual life expectancy evolution at retirement age holds great importance for life insurers and pension schemes. Over the last 30 years, numerous models for forecasting mortality have been introduced, and those that allow us to predict the mortality of two or more related populations simultaneously are particularly important. Indeed, these models, in addition to improving the forecasting accuracy overall, enable evaluation of the basis risk in index-based longevity risk transfer deals. This paper implements and compares several model-averaging approaches in a two-population context. These approaches generate predictions for life expectancy and the Gini index by averaging the forecasts obtained using a set of two-population models. In order to evaluate the eventual gain of model-averaging approaches for mortality forecasting, we quantitatively compare their performance to that of the individual two-population models using a large sample of different countries and periods. The results show that, overall, model-averaging approaches are superior both in terms of mean absolute forecasting error and interval forecast accuracy.

Keywords: model averaging; mortality forecasting; two-population models; life expectancy; Gini index

Citation: De Mori, Luca, Pietro Millossovich, Rui Zhu, and Steven Haberman. 2024. Two-Population Mortality Forecasting: An Approach Based on Model Averaging. *Risks* 12: 60. https://doi.org/10.3390/risks12040060

Academic Editor: Han Li

Received: 13 February 2024
Revised: 15 March 2024
Accepted: 21 March 2024
Published: 27 March 2024

1. Introduction

In recent decades, as a consequence of life expectancy improvements and social and behavioural changes that have taken place in various countries, pension funds, annuities, and other insurance products that provide retirement income have become increasingly important. However, since these products are subjected to longevity risk, which refers to the systematic trend of mortality rates decreasing over time, it has become necessary to find effective models for forecasting mortality rates. Notably, several models have been developed to address this issue, including the Lee-Carter model (Lee and Carter 1992), its extension the Renshaw-Haberman model (Renshaw and Haberman 2006), the Cairns-Blake-Dowd model (Cairns et al. 2006) and its extensions—the M6, M7, and M8 models (Cairns et al. 2009)—and the Plat model (Plat 2009). In all these models, the mortality rates depend on two types of quantities: fixed parameters that represent the effect of age on the mortality and stochastic factors that represent the effect of cohort year and calendar year. All these quantities must be estimated using statistical techniques based on past data. Finally, the mortality rates are forecasted, extrapolating the stochastic factors on the more recent period. Originally, these models had been designed to forecast the mortality rates of single populations. Later, they were implemented in a multi-population framework by Li and Lee (2005). This class of models has become increasingly popular because it allows researchers to work simultaneously with different populations that are to some extent related (e.g., males and females in the same country or region), obtaining coherent forecasts; see Dowd et al. (2011), Li (2013), Yang et al. (2016), Enchev et al. (2017), and Shang et al. (2022). This means that these models are able to respect limitations and constraints that we set such as the biological ones like the sex gap between the life expectancy of females and

males. As an instance of practical application, we can refer to the longevity risk transfer products where it is necessary to quantify the basis risk that is the systematic difference between the population mortality and the pension fund mortality; see Villegas et al. (2017).

As researchers continue to make progress in developing single-population and multi-population mortality forecasting models, there has also been a growing interest in model-averaging approaches in recent years; see Shang et al. (2011), Shang (2012), and Benchimol et al. (2016). The idea behind them is that forecasts are obtained by averaging models' predictions using various weighting schemes. By adopting a model-averaging approach, we can avoid some potential drawbacks of using single or multi-population mortality forecasting models; see Hinne et al. (2020) and Benchimol et al. (2016). These models are affected by a sort of over-confidence in that they assume the selected model is the only one correct and will produce precise forecasts in any situation. This implies an all-or-nothing mentality. They could lead to large-scale forecasting errors (outliers) and incoherence: in the presence of new data (mortality rates of a different country, for instance), the selected model may no longer be optimal among those studied. Indeed, the model accuracy, on which optimal model selection depends, is heavily influenced by the dataset used in the selection process.

This paper aims at introducing and comparing model-averaging approaches in a two-population mortality forecasting context and quantitatively evaluating them, highlighting their differences and comparing them to specific two-population models. We used truncated life expectancy and the Gini index as metrics to capture the location and dispersion of the residual lifetime distribution. Our main conclusions are that a simple, equally weighted approach performs just as well as more sophisticated averaging approaches, and model-averaging approaches are overall superior in terms of mean absolute forecasting error and interval forecast accuracy to most common two-population models by considering a range of combinations of test and training periods and countries.

The remainder of this paper is organised as follows. In Section 2, we revisit some existing stochastic multi-population models in the literature with a particular focus on the two-population case. In Section 3, model-averaging approaches applied to two-population mortality forecasting are introduced from a theoretical point of view. In Section 4, we list the datasets that we use in our practical examples. In Section 5, the procedure implemented to obtain the quantitative results is described step-by-step. In Section 6, the results are presented and discussed. Finally, in Section 7, we summarise the most important findings of the previous sections and provide a future outlook.

2. Two-Population Mortality Models

Let us consider two populations, denoted by $p = m, f$, that represent males and females of one specific country, and assume that for both of them, the force of mortality is constant over each calendar year and age, so that the force of mortality coincides with the central death rate $m_{x,t}^{(p)}$. The number of deaths $D_{x,t}^{(p)}$ in population p, year t, and age x, conditionally on the central death rate $m_{x,t}^{(p)}$, is assumed to follow a Poisson distribution:

$$D_{x,t}^{(p)} \sim Po(N_{x,t}^{(p)} m_{x,t}^{(p)}), \tag{1}$$

where $N_{x,t}^{(p)}$ is the central exposure to risk. In order to forecast central death rates, several models have been proposed in the last three decades following the original idea of Lee and Carter (1992). In all these models, the natural logarithm of mortality rates $m_{x,t}^{(p)}$, or of probability of death $q_{x,t}^{(p)}$,[1] is expressed as a function of two different types of quantities: time-dependent stochastic factors and age-dependent parameters. A relevant selection of these models is considered in this paper in their two-population form and summarised in Table 1.[2] More precisely, in any given population, the mortality rates of both females ($m_{x,t}^{f}$) and males ($m_{x,t}^{m}$) are specified by one of the equations in 1–9 in Table 1. These models can be broadly classified as follows: models where the age is treated as categorical (1, 2, 8, and

9), as a quantitative variable (3–6) or hybrid (7); models which consider a cohort effect (2, 4, and 7) or not; models with a stochastic time factor common to both females and males populations (8 and 9) or models where the dependence between sexes only stems from the correlation in stochastic time factors; and models with one (1, 2, and 8), two (3, 4, 6, 7, and 9) or three stochastic time factors (5).

Table 1. Summary of multi-population models used. Here $\kappa_t^{(i,p)}$, $i = 1,2,3$, κ_t, and $\kappa_t^{(p)}$ are time-varying stochastic factors; $\gamma_{t-x}^{(p)}$ are cohort-related stochastic factors; $\beta_x^{(i,p)}$, $i = 1,2,3$, and $\beta_x^{(2)}$ are age-specific parameters; $\bar{x} = \frac{1}{m+1}\sum_{i=0}^{m} x_i$ is the mean age over the population age range; $\hat{\sigma}_x^2 = \frac{1}{m+1}\sum_{i=0}^{m}(x_i - \bar{x})^2$ is the age variance; and finally $x_c^{(p)}$ is an arbitrary fixed age.

Model	$\ln(m_{x,t}^{(p)})$
1. Lee–Carter model (LC)	$\beta_x^{(1,p)} + \beta_x^{(2,p)}\kappa_t^{(2,p)}$
2. Renshaw–Haberman model (RH)	$\beta_x^{(1,p)} + \beta_x^{(2,p)}\kappa_t^{(2,p)} + \gamma_{t-x}^{(p)}$
3. Cairns–Blake–Dowd model (CBD)	$\kappa_t^{(1,p)} + \kappa_t^{(2,p)}(x - \bar{x})$
4. CBD Model with a cohort effect (M6)	$\kappa_t^{(1,p)} + \kappa_t^{(2,p)}(x - \bar{x}) + \gamma_{t-x}^{(p)}$
5. CBD Model with quadratic and cohort effects (M7)	$\kappa_t^{(1,p)} + \kappa_t^{(2,p)}(x - \bar{x}) + \kappa_t^{(3,p)}((x - \bar{x})^2 - \hat{\sigma}_x^2) + \gamma_{t-x}^{(p)}$
6. CBD Model with an age-dependent cohort effect (M8)	$\kappa_t^{(1,p)} + \kappa_t^{(2,p)}(x - \bar{x}) + \gamma_{t-x}^{(p)}(x_c^{(p)} - x)$
7. Plat model (PLAT)	$\beta_x^{(1,p)} + \kappa_t^{(1,p)} + \kappa_t^{(2,p)}(x - \bar{x}) + \gamma_{t-x}^{(p)}$
8. Common Factor Model (CF)	$\beta_x^{(1,p)} + \beta_x^{(2)}\kappa_t$
9. Augmented Common Factor Model (ACF)	$\beta_x^{(1,p)} + \beta_x^{(2)}\kappa_t + \beta_x^{(2,p)}\kappa_t^{(2,p)}$

2.1. Model Estimation

The parameters of the models in Table 1 are usually estimated by maximizing the joint Poisson log-likelihood:

$$\ell = \sum_p \sum_x \sum_t \left\{ d_{x,t}^{(p)} \ln(N_{x,t}^{(p)} m_{x,t}^{(p)}) - N_{x,t}^{(p)} m_{x,t}^{(p)} - \ln(d_{x,t}^{(p)}!) \right\} \tag{2}$$

where $d_{x,t}^{(p)}$ are the observed deaths in population p, year t, and age x. The mortality rates $m_{x,t}^{(p)}$ for each model can be obtained from the corresponding equations in Table 1. The optimisation is performed using numerical algorithms. Note that for models 1–7, the log-likelihood for each population can be maximised separately.

2.2. Stochastic Factor Assumptions

From Table 1, it can be seen that the models considered depend on a number of stochastic factors. More precisely, each model contains a combination of (one or more of) the following terms: population-specific time indices $\kappa_t^{(i,p)}$, common time index κ_t, and population-specific cohort effects $\gamma_t^{(p)}$. Inspired by Li et al. (2015), for the time indices $\kappa_t^{(i,p)}$ (models 1–7, 9) we consider a combination of a random walk with drift and first-order autoregression AR(1). The rationale of this choice is that there is a stable relation between the period indices of males and females.

- $\kappa_t^{(i,m)} = \mu^{(i,m)} + \kappa_{t-1}^{(i,m)} + Z_t^{(i,m)}, \quad i = 1,2,3$
- $\kappa_t^{(i,f)} = \kappa_t^{(i,m)} + \phi^{(i,f)}(\kappa_{t-1}^{(i,f)} - \kappa_{t-1}^{(i,m)}) + Z_t^{(i,f)}, \quad i = 1,2,3$

where $\mu^{(i,m)}$ are the drift parameters, $\phi^{(i,f)}$ are the autoregressive parameters, and $(Z_t^{(i,p)})_{p=m,f}$ are normal iid innovations.

For the time index κ_t (models 8–9), we consider a random walk with drift:

- $\kappa_t = \mu + \kappa_{t-1} + Z_t$,

where μ is the drift parameter, ϕ is the autoregressive parameter, and Z_t are normal iid innovations.

Finally, for the cohort terms $\gamma_{t-x}^{(p)}$ (models 2, 4–7), we consider a combination of ARIMA(1, 1, 0) (see Villegas et al. (2017) and Dowd et al. (2010)) and first-order autoregression AR(1) (see Li et al. (2015)). Again, the rationale is that there is a stable relation between the cohort effects of males and females.

- $\gamma_u^{(m)} = (1+\phi^{(m)})\gamma_{u-1}^{(m)} - \phi^{(m)}\gamma_{u-2}^{(m)} + Y_u^{(m)}$
- $\gamma_u^{(f)} = \gamma_u^{(m)} + \phi^{(f)}(\gamma_{u-1}^{(f)} - \gamma_{u-1}^{(m)}) + Y_u^{(f)}$

where $\phi^{(m)}$ and $\phi^{(f)}$ are the autoregressive parameters of the process, while $(Y_u^{(p)})_{p=m,f}$ are normal iid innovations.

In models 1–7, the dependence between female and male mortality is derived from the autoregressive components. In model 8, the dependence is given by the shared time index κ_t between female and male populations. In model 9, the dependence is derived by both the shared time index and the autoregressive component.

3. Model-Averaging Approaches

Suppose we have historical data on mortality for the period $[t_0, t_s]$, and we are interested in forecasting some mortality metric on the period $[t_{s+1}, t_n]$. The purpose of model-averaging approaches is to obtain forecasts of a given metric

$$U_t^{(p)} = f((m_{x,t}^{(p)})_{x=0,\ldots,\omega-1}) \tag{3}$$

that can be expressed as a function of mortality rates, where ω is the ultimate age, as an average of the forecasted metrics obtained using L different models. Notice that in this paper, we consider individual models listed in Table 1, and $L = 9$. As an example of the metric $U_t^{(p)}$, we can use the j-years survival probability ${}_jp_{x,t}^{(p)} = \exp\{-(m_{x,t}^{(p)} + m_{x+1,t}^{(p)} + \cdots + m_{x+j-1,t}^{(p)})\}$.

Let the metric of interest for the population p, in year t, and model l be

$$\hat{U}_t^{(p,l)} = f((\hat{m}_{x,t}^{(p,l)})_{x=0,\ldots,\omega-1}), \qquad l = 1, \ldots, L \tag{4}$$

where $\hat{m}_{x,t}^{(p,l)}$ is the forecasted mortality rate at age x, for the population p, in year t, obtained using model l. Following Fletcher (2018) and Shang (2012), the averaged metric is calculated as

$$\hat{U}_t^{(p,average)} = \lambda_1\hat{U}_t^{(p,1)} + \cdots + \lambda_L\hat{U}_t^{(p,L)} \tag{5}$$

where $\lambda_L, \ldots, \lambda_L$ are non-negative weights calculated based on the model-averaging approach considered dependently on the performance of the models in the validation period. In this paper, we consider the following four model-averaging approaches:

- Equal weights (EW):

$$\lambda_l^{EW} = \frac{1}{L}, \qquad l = 1, \ldots, L. \tag{6}$$

  This method is the most simple, as all models are assigned the same weight; see Shang (2012). There is no penalisation or reward depending on the performance in the validation period.
- Proportional weights (PW):

$$\lambda_l^{PR} = \frac{\frac{1}{g_l}}{\sum_{k=1}^{L}\frac{1}{g_k}}, \qquad l = 1, \ldots, L \tag{7}$$

  where $g_l = g(\hat{U}_t^{(p,l)}, U_t^{(p,l)}; p = m, f; t = t_s - h + 1, \ldots, t_s)$ is a strictly positive performance measure representing the performance of the model l in the validation period $[t_s - h + 1, t_s]$; see Shang (2012). In this way, models that have poor performance in the validation period are penalised with smaller weights.

- Weights based on the the softmax function (SM):

$$\lambda_l^{SM} = \frac{\exp\{-g_l\}}{\sum_{k=1}^{L}\exp\{-g_k\}}, \qquad l = 1, \ldots, L. \tag{8}$$

The concept is similar to the proportional weights model-averaging approach, but here we penalise less the models with poor performance in the validation period and reward less the models with good performance. See also Benchimol et al. (2016) for a similar formulation.

- Weights based on trimming (TR):

$$\lambda_l^{TR} = \begin{cases} \frac{1}{\hat{L}}, & \text{if } l \text{ is among the } \hat{L} \text{ best models in the validation period,} \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where the best models are determined in terms of the measure g_l; see Samuels and Sekkel (2017) and Shang (2012). With this method, we reward only the $\hat{L}$ models that have the best performance in the validation period, assigning the same weight ($\frac{1}{\hat{L}}$) for each one of them. In the following, we set $\hat{L}$ equal to 3.

In the remainder of this paper, for the definition of measure g used to evaluate the models' performance in the validation period, we adopt the mean absolute forecasting error (MAFE)

$$\text{MAFE}_l = \frac{\sum_{p=m,f}\sum_{t=t_s-h+1}^{t_s} |\hat{U}_t^{(p,l)} - U_t^{(p,l)}|}{P \cdot h} \tag{10}$$

where t_{s-h+1} and t_s are the first and last years of the validation period. Notice that alternative measures could also have been considered. For example, one that also takes into account the number of parameters in the model.

Finally, regarding the metric in (4), we choose the residual life expectancy at age 55 truncated at age 90, which represents a location metric of mortality rates; see Dickson et al. (2019),

$$\mathring{e}_{55:\overline{35|},t} = \sum_{j=1}^{35} {}_{j-1}p_{55,t}\left(1 - \frac{1}{2}q_{55+j-1,t}\right) \tag{11}$$

and the Gini index, calculated between 55 and 89, that represents the dispersion of mortality rates

$$G_{55:\overline{35|},t} = \frac{1}{2\mathring{e}_{55:\overline{35|},t}} \sum_{x=55}^{89}\sum_{y=55}^{89} {}_{x-55|1}q_{55,t}\ {}_{y-55|1}q_{55,t} \mid x - y \mid . \tag{12}$$

The Gini index is a metric that varies between the limits of 0 (perfect equality) and 1 (perfect inequality). For a length of life distribution, it is equal to zero if all individuals die at the same age and equal to one if all people die at age 0 and one individual dies at an infinitely old age. The choice of the Gini index as a metric representing the dispersion of mortality depends on the fact that, unlike other metrics such as interquartile range, variance, and standard deviation, it possesses all the desirable properties for an inequality index: population-size independence, mean or scale independence, and transfer principle; see Shkolnikov et al. (2003) for more details.

4. Data

In the following, we implement the multi-population models and the model-averaging approaches with historical mortality data from ten pairs of populations, namely the female and male populations of Australia, Canada, France, England and Wales, Italy, Japan, Netherlands, Spain, Sweden, and the US. The choice of these countries depends on the fact that they are developed with large populations, and their data are complete and easily

obtainable in the Human Mortality Database (HMD). For these countries, we considered the age range between 55 and 89 for the mortality rates. The reason for this choice derives from the fact that most of the deaths are concentrated after the age of 55 years and that after the age of 90 years, we have fewer data, especially for the older cohorts, so this could lead to biased estimations of the models. Furthermore, we note that, at ages over 90, there is evidence of age misstatements that may lead to biased estimates of mortality indices. Moreover, this age range is the most relevant from an actuarial point of view; see Cairns et al. (2006). Finally, concerning the periods considered, we have two cases: in the first one, we have a 30-year rolling training period from 1950–1979 to 1975–2004 and a 15-year rolling test period from 1980–1994 to 2005–2019 (last year for which the data are available at the time of writing this paper); in the second one, we have a variable length training period from 1966–2004 to 1985–2004 and a 15-year fixed test period 2005–2019.

5. Implementation

Step-By-Step Procedure

For a generic country, training period $[t_0, t_s]$, and test period $[t_s + 1, t_n]$ (see Figure 1), we follow the steps listed below:

1. First stage
   - 1.1 We fit the two-population models (LC, RH, CBD, PLAT, M6, M7, M8, CF, and ACF) on the period $[t_0, t_s - 10]$ (training period 1) using the StMoMo package; see Villegas et al. (2018). Notice that this implies h, i.e., the length of the validation period, is set equal to 10.
   - 1.2 We simulate mortality rates for the period $[t_s - 9, t_s]$ (validation period) using the models fitted in 1.1.
   - 1.3 We calculate the corresponding truncated life expectancy and Gini index for each model as functions of the mortality rates obtained in 1.2 using Formulas (11) and (12).
   - 1.4 We repeat steps 1.2 and 1.3 1000 times, and we obtain the forecasted truncated life expectancy and Gini index as the average of these for each model.
   - 1.5 We calculate the MAFE as the difference between forecasted truncated life expectancy and Gini index calculated in 1.4 and the historical ones (Formula (10)) for each model.
   - 1.6 We calculate the weights of each model-averaging approach based on the MAFEs calculated in 1.5 using Formulas (6)–(9).
2. Second stage
   - 2.1 We repeat step 1.1 using the period $[t_0, t_s]$ (training period 2) instead of $[t_0, t_s - 10]$.
   - 2.2 We repeat steps 1.2 and 1.3 using the period $[t_s + 1, t_n]$ (test period) instead of $[t_s - 9, t_s]$.
   - 2.3 We repeat step 2.2 10,000 times[3] and we obtain the forecasted truncated average life expectancy and Gini index as the average of these for each model.
   - 2.4 For each model-averaging approach, we carry out 1 simulation from a multinomial distribution with parameters equal to 10,000, 9, and the vector of the weights obtained in 1.6. The result of this simulation will be a vector with 9 elements, which sum to 10,000, that represent the number of truncated life expectancy and Gini index trajectories that are considered in the model-averaging approach from the 9 two-population models.
   - 2.5 Using the results of the simulation at point 2.4 as parameters, we resample by bootstrapping from the truncated life expectancy and Gini index trajectories obtained in step 2.3, and we average them using Formula (5) obtaining the forecasted truncated life expectancy and Gini index for all the model-averaging approaches. Similarly, we take the 5th and 95th percentile to build the 90% confidence forecasting intervals for the two metrics. See Figure 2 for an exam-

ple of forecasted life expectancy and Gini index using the model-averaging approach with equal weights.

2.6 We calculate the MAFE as the difference between the forecasted truncated life expectancy and the Gini index calculated in 2.3 and 2.5 with the observed ones. Similarly, we compare the confidence forecasting intervals of the two metrics with the observed values in order to obtain the interval forecast accuracy that represents the proportion of times in which the observed truncated life expectancy and Gini index fall within the respective prediction intervals.

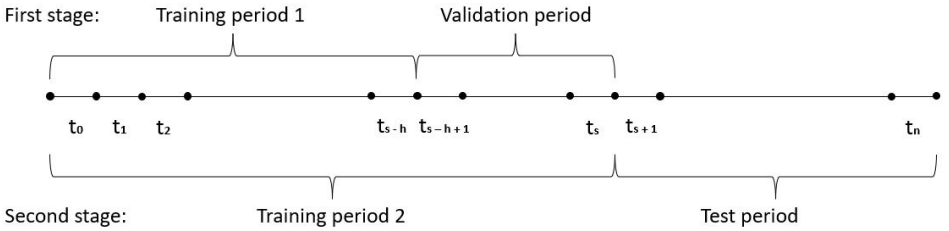

Figure 1. An illustration of the training, validation, and test periods to train and evaluate the models.

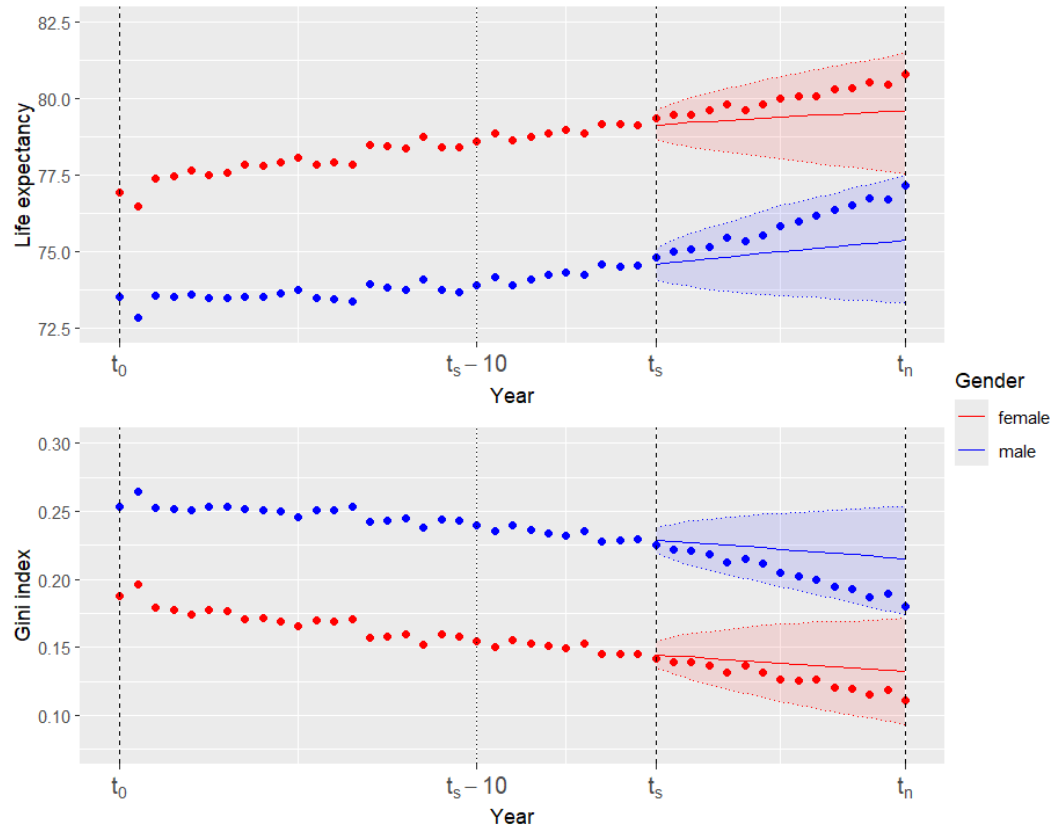

Figure 2. Example of forecasted truncated life expectancy and Gini index, the respective 90% prediction intervals, and the observed values of the two metrics (points). Training period: 1950–1979. Test period: 1980–1994. Country: England and Wales. Model-averaging approach: equal weights.

6. Results

Following Shang (2012), in order to evaluate the performance of each model, we must consider the goodness of both point and interval forecasts. For the first one, we consider the mean absolute forecasting error (MAFE; see Formula (10)), while for the second one, the interval forecast accuracy, here defined as the proportion of cases in which the observed life expectancy or Gini index falls within the 90% confidence forecasting interval, is considered. In Figure 3, we can find boxplots summarising the mean absolute forecasting errors for life

expectancy and the Gini index obtained by all the models and model-averaging approaches previously mentioned in the rolling test period case. Figure 4 shows, for the rolling test period case, which is the best model, i.e., the one with the lowest MAFE, by period, country, and metric. Figures 5 and 6 have the same content as Figures 3 and 4 respectively, but for the fixed test period case. Finally, Tables 2–5 and 6–9 report the interval forecast accuracy of the models by period, country, and metric, respectively, for rolling and fixed test period cases.

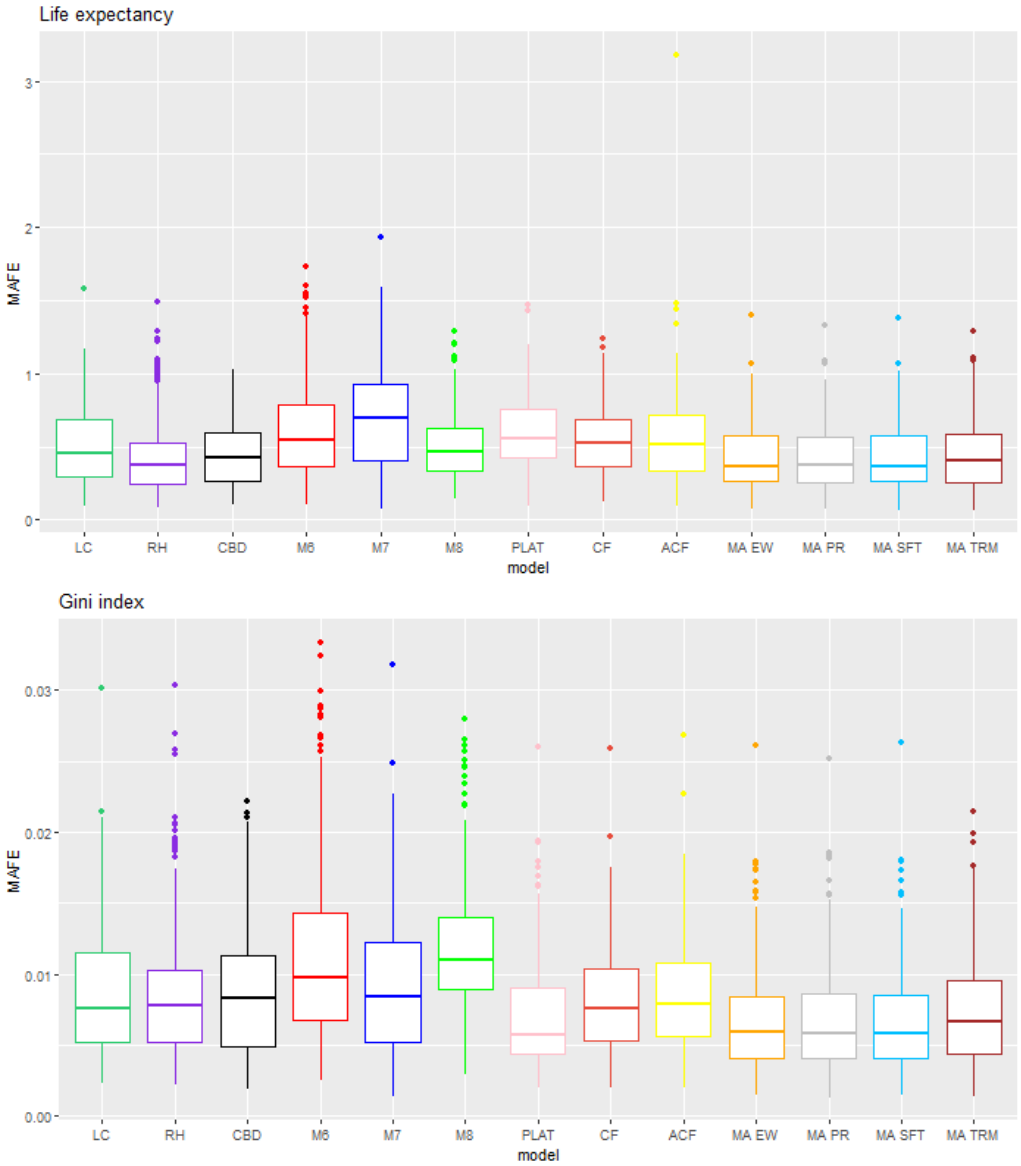

Figure 3. Summary of the MAFEs by model. Results for individual models and corresponding weighted average. Rolling test period case.

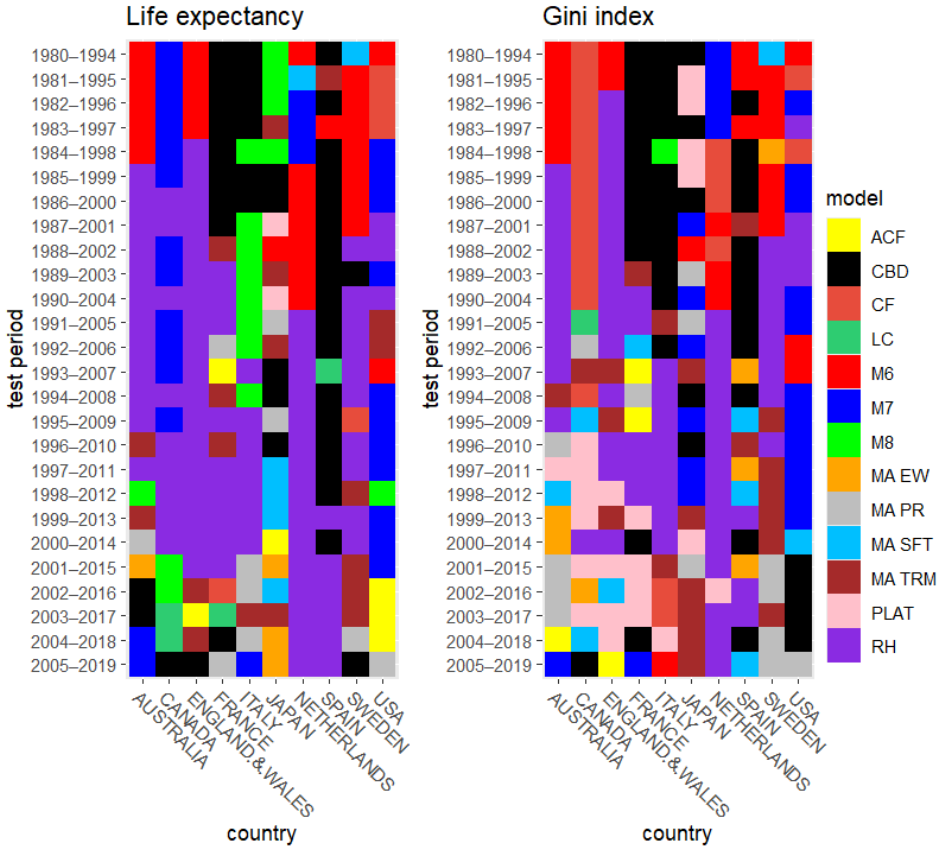

Figure 4. Model with the lowest MAFE by period and country. Results for individual models and model-averaging approaches. Rolling test period case.

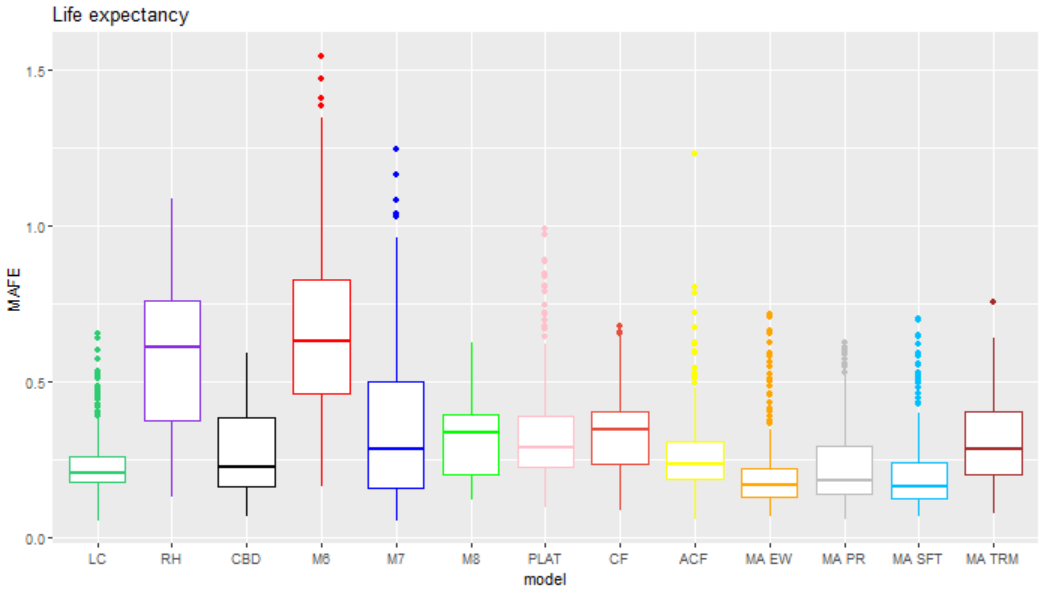

Figure 5. *Cont.*

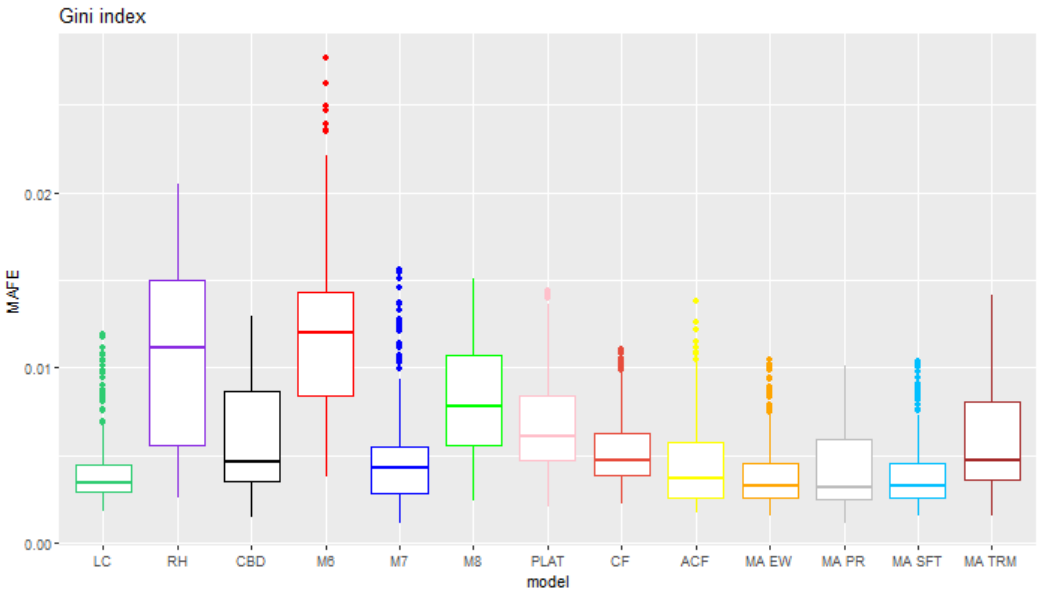

Figure 5. Summary of the MAFEs by model. Results for individual models and corresponding weighted average. Fixed test period case.

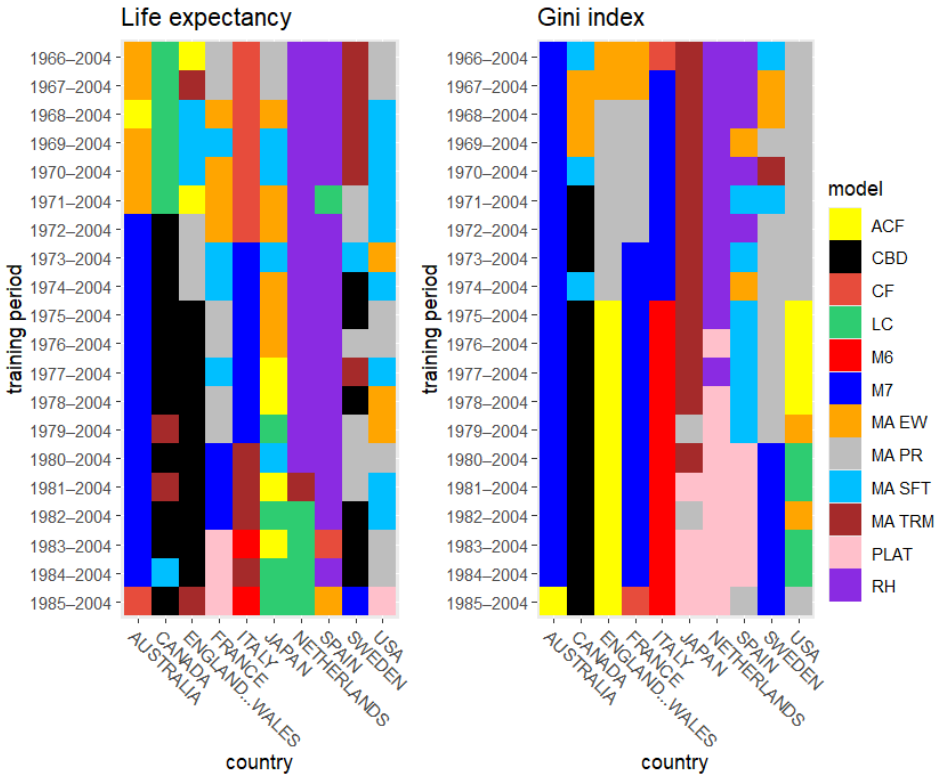

Figure 6. Model with the lowest MAFE by period and country. Results for individual models and model-averaging approaches. Fixed test period case.

Table 2. Interval forecast accuracy by period. Proportion of cases in which the observed life expectancy falls in the forecasting interval. Dark shades of green and red signify higher interval forecast accuracy, whereas lighter shades denote lower interval forecast accuracy. Rolling test period case.

Training Period	Test Period	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
1950–1979	1980–1994	83%	84%	95%	84%	77%	95%	71%	82%	91%	100%	98%	100%	96%
1951–1980	1981–1995	82%	89%	95%	83%	81%	91%	71%	81%	86%	99%	98%	99%	96%
1952–1981	1982–1996	83%	87%	90%	79%	81%	88%	72%	81%	86%	99%	98%	99%	97%
1953–1982	1983–1997	81%	84%	91%	83%	79%	86%	81%	84%	84%	99%	98%	98%	96%
1954–1983	1984–1998	75%	81%	87%	75%	76%	82%	77%	77%	79%	98%	97%	97%	95%
1955–1984	1985–1999	86%	87%	92%	83%	84%	92%	87%	88%	86%	99%	98%	99%	97%
1956–1985	1986–2000	80%	82%	83%	77%	80%	78%	86%	75%	76%	98%	98%	98%	95%
1957–1986	1987–2001	81%	82%	85%	80%	80%	84%	88%	77%	76%	98%	97%	98%	96%
1958–1987	1988–2002	82%	83%	87%	77%	83%	82%	92%	82%	80%	98%	97%	98%	95%
1959–1988	1989–2003	79%	80%	76%	74%	81%	74%	90%	75%	75%	96%	95%	96%	92%
1960–1989	1990–2004	78%	84%	85%	76%	80%	80%	95%	77%	74%	97%	97%	97%	91%
1961–1990	1991–2005	86%	88%	79%	75%	89%	82%	89%	81%	81%	97%	97%	97%	96%
1962–1991	1992–2006	85%	89%	81%	75%	85%	82%	87%	79%	78%	96%	97%	97%	96%
1963–1992	1993–2007	86%	85%	79%	75%	83%	81%	84%	76%	82%	98%	97%	97%	90%
1964–1993	1994–2008	76%	79%	61%	59%	83%	63%	86%	66%	73%	91%	92%	92%	86%
1965–1994	1995–2009	84%	86%	70%	70%	82%	75%	87%	71%	81%	99%	99%	99%	95%
1966–1995	1996–2010	82%	85%	60%	60%	84%	65%	90%	70%	76%	97%	97%	97%	87%
1967–1996	1997–2011	81%	82%	61%	60%	81%	66%	90%	67%	76%	97%	98%	98%	96%
1968–1997	1998–2012	85%	81%	60%	62%	79%	72%	89%	70%	79%	98%	98%	98%	93%
1969–1998	1999–2013	82%	79%	53%	59%	74%	66%	88%	68%	76%	96%	95%	96%	89%
1970–1999	2000–2014	74%	75%	33%	41%	73%	54%	86%	61%	73%	94%	95%	95%	88%
1971–2000	2001–2015	88%	83%	45%	59%	77%	68%	72%	65%	80%	96%	94%	96%	91%
1972–2001	2002–2016	90%	86%	49%	70%	79%	68%	62%	64%	86%	98%	98%	98%	94%
1973–2002	2003–2017	87%	82%	35%	63%	74%	67%	70%	62%	85%	95%	94%	95%	78%
1974–2003	2004–2018	75%	82%	25%	52%	68%	56%	61%	59%	79%	92%	90%	92%	82%
1975–2004	2005–2019	90%	77%	51%	83%	77%	83%	41%	68%	87%	99%	99%	98%	95%
Average		82%	83%	70%	71%	80%	76%	81%	73%	80%	97%	97%	97%	93%

Table 3. Interval forecast accuracy by country. Proportion of cases in which the observed life expectancy falls in the forecasting interval. Dark shades of green and red signify higher interval forecast accuracy, whereas lighter shades denote lower interval forecast accuracy. Rolling test period case.

	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
AUSTRALIA	90%	94%	82%	95%	96%	84%	91%	92%	91%	99%	99%	99%	97%
CANADA	61%	62%	49%	78%	59%	54%	58%	62%	52%	99%	96%	99%	78%
ENGLAND AND WALES	74%	75%	78%	69%	71%	70%	89%	63%	75%	99%	99%	99%	99%
FRANCE	97%	96%	34%	63%	89%	80%	88%	87%	97%	100%	99%	100%	97%
ITALY	84%	88%	70%	59%	88%	76%	83%	88%	85%	97%	97%	98%	95%
JAPAN	99%	87%	87%	62%	75%	96%	80%	50%	99%	100%	100%	100%	96%
NETHERLANDS	56%	61%	77%	71%	57%	76%	80%	67%	58%	86%	87%	87%	86%
SPAIN	98%	98%	85%	70%	95%	94%	95%	88%	99%	100%	100%	100%	95%
SWEDEN	73%	80%	62%	42%	78%	65%	71%	69%	66%	91%	88%	90%	84%
USA	91%	92%	71%	96%	89%	67%	71%	67%	80%	100%	100%	100%	98%
Average	82%	83%	70%	71%	80%	76%	81%	73%	80%	97%	97%	97%	93%

Table 4. Interval forecast accuracy by period. Proportion of cases in which the observed Gini index falls in the forecasting interval. Dark shades of green and red signify higher interval forecast accuracy, whereas lighter shades denote lower interval forecast accuracy. Rolling test period case.

Training Period	Test Period	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
1950–1979	1980–1994	92%	81%	97%	96%	74%	95%	79%	90%	91%	100%	100%	100%	98%
1951–1980	1981–1995	89%	88%	97%	96%	74%	95%	79%	86%	86%	99%	99%	99%	99%
1952–1981	1982–1996	90%	82%	92%	93%	73%	94%	79%	85%	86%	99%	99%	99%	97%
1953–1982	1983–1997	89%	77%	97%	98%	71%	93%	86%	88%	87%	99%	99%	100%	98%
1954–1983	1984–1998	81%	76%	87%	86%	68%	96%	83%	80%	76%	98%	97%	97%	95%
1955–1984	1985–1999	92%	83%	94%	96%	73%	100%	89%	90%	89%	100%	100%	100%	99%
1956–1985	1986–2000	85%	78%	85%	95%	72%	97%	87%	78%	78%	98%	97%	98%	93%
1957–1986	1987–2001	86%	79%	85%	95%	69%	98%	90%	81%	78%	99%	99%	99%	97%
1958–1987	1988–2002	86%	79%	91%	95%	66%	98%	92%	86%	81%	99%	98%	98%	96%
1959–1988	1989–2003	84%	79%	82%	91%	72%	95%	90%	77%	74%	96%	96%	96%	94%
1960–1989	1990–2004	84%	75%	87%	93%	61%	97%	94%	79%	76%	99%	99%	99%	94%
1961–1990	1991–2005	90%	84%	85%	93%	68%	95%	88%	75%	84%	98%	98%	98%	96%
1962–1991	1992–2006	87%	83%	88%	93%	63%	93%	84%	75%	81%	98%	98%	98%	95%
1963–1992	1993–2007	84%	81%	85%	90%	58%	89%	82%	74%	76%	95%	94%	95%	91%
1964–1993	1994–2008	78%	82%	69%	81%	63%	91%	87%	68%	67%	95%	95%	96%	90%
1965–1994	1995–2009	90%	83%	79%	96%	62%	93%	83%	74%	78%	99%	99%	99%	96%
1966–1995	1996–2010	82%	83%	69%	87%	64%	93%	89%	75%	68%	100%	99%	100%	94%
1967–1996	1997–2011	85%	81%	71%	86%	63%	92%	88%	71%	69%	99%	98%	99%	95%
1968–1997	1998–2012	86%	81%	70%	86%	59%	90%	84%	74%	77%	99%	98%	99%	96%
1969–1998	1999–2013	84%	78%	64%	81%	59%	91%	86%	69%	72%	97%	97%	97%	95%
1970–1999	2000–2014	74%	81%	42%	67%	63%	88%	87%	59%	58%	97%	96%	97%	90%
1971–2000	2001–2015	86%	82%	58%	87%	66%	86%	72%	66%	73%	97%	96%	97%	91%
1972–2001	2002–2016	89%	80%	60%	91%	67%	85%	58%	67%	83%	100%	100%	100%	91%
1973–2002	2003–2017	82%	79%	49%	85%	68%	89%	70%	60%	76%	95%	95%	95%	85%
1974–2003	2004–2018	73%	86%	38%	69%	63%	86%	64%	55%	67%	95%	95%	95%	83%
1975–2004	2005–2019	89%	72%	61%	95%	64%	72%	40%	73%	84%	100%	100%	100%	93%
Average		85%	81%	76%	89%	66%	92%	81%	75%	77%	98%	98%	98%	94%

Table 5. Interval forecast accuracy by country. Proportion of cases in which the observed Gini index falls in the forecasting interval. Dark shades of green and red signify higher interval forecast accuracy, whereas lighter shades denote lower interval forecast accuracy. Rolling test period case.

	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
AUSTRALIA	93%	97%	83%	99%	95%	89%	90%	90%	85%	100%	100%	100%	97%
CANADA	69%	51%	61%	78%	38%	80%	55%	67%	57%	100%	98%	100%	82%
ENGLAND AND WALES	83%	75%	78%	84%	84%	94%	88%	64%	67%	100%	100%	100%	96%
FRANCE	94%	89%	42%	92%	68%	97%	86%	86%	92%	100%	100%	100%	97%
ITALY	88%	95%	84%	91%	78%	95%	82%	87%	79%	99%	99%	99%	98%
JAPAN	93%	83%	89%	96%	56%	98%	84%	46%	88%	100%	100%	100%	100%
NETHERLANDS	59%	56%	80%	83%	47%	91%	82%	75%	66%	91%	90%	91%	85%
SPAIN	98%	98%	88%	92%	71%	97%	95%	91%	97%	100%	100%	100%	100%
SWEDEN	81%	82%	78%	79%	59%	89%	80%	75%	66%	93%	92%	93%	89%
USA	93%	82%	79%	98%	67%	89%	70%	72%	78%	99%	99%	99%	95%
Average	85%	81%	76%	89%	66%	92%	81%	75%	77%	98%	98%	98%	94%

Table 6. Interval forecast accuracy by period. Proportion of cases in which the observed life expectancy falls in the forecasting interval. Dark shades of green and red signify higher interval forecast accuracy, whereas lighter shades denote lower interval forecast accuracy. Fixed test period case.

Training Period	Test Period	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
1966–2004	**2005–2019**	91%	84%	48%	62%	83%	78%	62%	69%	84%	98%	99%	97%	93%
1967–2004	**2005–2019**	91%	84%	49%	63%	82%	77%	59%	70%	85%	97%	99%	98%	93%
1968–2004	**2005–2019**	91%	82%	52%	67%	80%	79%	55%	71%	88%	96%	97%	97%	92%
1969–2004	**2005–2019**	90%	81%	52%	72%	79%	80%	54%	71%	88%	97%	99%	99%	93%
1970–2004	**2005–2019**	91%	81%	50%	70%	79%	79%	50%	70%	88%	98%	99%	98%	94%
1971–2004	**2005–2019**	91%	80%	49%	75%	79%	79%	47%	69%	88%	99%	99%	99%	94%
1972–2004	**2005–2019**	91%	80%	47%	75%	78%	78%	45%	66%	87%	99%	99%	99%	93%
1973–2004	**2005–2019**	91%	78%	50%	78%	77%	81%	43%	68%	88%	98%	99%	98%	89%
1974–2004	**2005–2019**	91%	79%	51%	81%	78%	82%	44%	68%	88%	99%	98%	98%	90%
1975–2004	**2005–2019**	90%	77%	50%	83%	78%	84%	40%	69%	87%	97%	99%	99%	94%
1976–2004	**2005–2019**	90%	78%	52%	86%	78%	85%	39%	70%	88%	99%	98%	98%	95%
1977–2004	**2005–2019**	91%	79%	49%	85%	77%	84%	40%	72%	89%	96%	98%	97%	94%
1978–2004	**2005–2019**	90%	79%	51%	87%	78%	85%	37%	75%	90%	97%	98%	97%	96%
1979–2004	**2005–2019**	90%	81%	49%	89%	78%	84%	37%	76%	90%	97%	98%	98%	94%
1980–2004	**2005–2019**	89%	79%	52%	89%	77%	85%	35%	73%	88%	98%	98%	97%	99%
1981–2004	**2005–2019**	89%	79%	53%	91%	77%	85%	35%	74%	88%	98%	98%	98%	100%
1982–2004	**2005–2019**	89%	80%	56%	90%	78%	84%	36%	75%	88%	98%	98%	98%	98%
1983–2004	**2005–2019**	86%	80%	57%	90%	77%	82%	33%	76%	86%	99%	99%	99%	95%
1984–2004	**2005–2019**	86%	82%	57%	93%	78%	86%	34%	76%	87%	99%	99%	99%	97%
1985–2004	**2005–2019**	82%	78%	62%	93%	80%	85%	34%	75%	85%	99%	99%	99%	94%
Average		90%	80%	52%	81%	79%	82%	43%	72%	87%	98%	98%	98%	94%

Table 7. Interval forecast accuracy by country. Proportion of cases in which the observed life expectancy falls in the forecasting interval. Dark shades of green and red signify higher interval forecast accuracy, whereas lighter shades denote lower interval forecast accuracy. Fixed test period case.

	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
Australia	96%	100%	71%	100%	100%	98%	37%	95%	99%	100%	100%	100%	97%
Canada	85%	87%	28%	94%	84%	52%	30%	32%	65%	99%	98%	99%	84%
England and Wales	99%	100%	66%	100%	98%	79%	32%	64%	96%	100%	100%	100%	100%
France	90%	54%	18%	89%	64%	97%	33%	98%	92%	100%	100%	100%	99%
Italy	91%	73%	87%	96%	69%	93%	24%	97%	95%	100%	100%	100%	83%
Japan	99%	48%	58%	73%	50%	100%	53%	60%	100%	100%	100%	100%	100%
Netherlands	55%	62%	47%	45%	58%	50%	80%	48%	51%	80%	87%	82%	97%
Spain	100%	90%	39%	65%	77%	93%	99%	84%	100%	100%	100%	100%	92%
Sweden	100%	100%	52%	81%	100%	89%	21%	93%	99%	100%	100%	100%	100%
USA	82%	88%	48%	66%	87%	71%	21%	46%	79%	100%	100%	100%	92%
Average	90%	80%	52%	81%	79%	82%	43%	72%	87%	98%	98%	98%	94%

Table 8. Interval forecast accuracy by period. Proportion of cases in which the observed Gini index falls in the forecasting interval. Dark shades of green and red signify higher interval forecast accuracy, whereas lighter shades denote lower interval forecast accuracy. Fixed test period case.

Training Period	Test Period	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
1966–2004	**2005–2019**	91%	80%	60%	88%	64%	85%	62%	73%	87%	99%	100%	99%	93%
1967–2004	**2005–2019**	91%	78%	62%	91%	67%	85%	61%	73%	87%	99%	100%	99%	93%
1968–2004	**2005–2019**	91%	75%	63%	91%	59%	80%	56%	73%	87%	99%	100%	99%	94%
1969–2004	**2005–2019**	91%	74%	62%	92%	62%	80%	56%	73%	87%	99%	100%	99%	91%
1970–2004	**2005–2019**	91%	76%	61%	92%	62%	81%	51%	72%	87%	99%	99%	99%	91%
1971–2004	**2005–2019**	90%	74%	60%	93%	62%	82%	50%	73%	87%	99%	100%	99%	92%
1972–2004	**2005–2019**	91%	76%	57%	93%	60%	77%	47%	72%	88%	100%	100%	100%	93%
1973–2004	**2005–2019**	90%	71%	61%	93%	61%	75%	43%	72%	85%	99%	99%	99%	94%
1974–2004	**2005–2019**	90%	72%	60%	94%	60%	73%	44%	73%	85%	100%	100%	100%	96%
1975–2004	**2005–2019**	89%	73%	60%	95%	66%	73%	41%	73%	84%	100%	100%	100%	93%
1976–2004	**2005–2019**	89%	75%	60%	94%	62%	68%	40%	74%	84%	100%	100%	100%	94%

Table 8. *Cont.*

Training Period	Test Period	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
1977–2004	2005–2019	88%	73%	59%	92%	64%	72%	38%	73%	83%	99%	99%	99%	91%
1978–2004	2005–2019	88%	72%	61%	91%	67%	70%	36%	75%	82%	99%	99%	99%	93%
1979–2004	2005–2019	88%	75%	56%	91%	70%	73%	37%	74%	83%	99%	99%	99%	92%
1980–2004	2005–2019	88%	72%	60%	89%	67%	69%	35%	74%	83%	99%	99%	99%	94%
1981–2004	2005–2019	88%	74%	61%	89%	68%	71%	35%	74%	82%	99%	99%	99%	93%
1982–2004	2005–2019	88%	75%	60%	90%	71%	71%	36%	74%	83%	99%	100%	99%	95%
1983–2004	2005–2019	84%	71%	62%	85%	71%	66%	35%	71%	78%	98%	97%	98%	88%
1984–2004	2005–2019	85%	73%	62%	90%	73%	68%	38%	72%	80%	99%	99%	99%	93%
1985–2004	2005–2019	81%	71%	66%	88%	68%	62%	37%	70%	77%	98%	98%	98%	86%
Average		89%	74%	61%	91%	65%	74%	44%	73%	84%	99%	99%	99%	93%

Table 9. Interval forecast accuracy by country. Proportion of cases in which the observed Gini index falls in the forecasting interval. Dark shades of green and red signify higher interval forecast accuracy, whereas lighter shades denote lower interval forecast accuracy. Fixed test period case.

	LC	CBD	M6	M7	M8	PLAT	RH	CF	ACF	MA E.W	MA PR.W	MA S.M	MA TR
AUSTRALIA	99%	100%	77%	100%	84%	32%	37%	96%	97%	100%	100%	100%	75%
CANADA	64%	95%	48%	82%	91%	69%	29%	46%	42%	100%	100%	100%	81%
ENGLAND AND WALES	99%	100%	58%	100%	97%	79%	33%	70%	96%	100%	100%	100%	100%
FRANCE	93%	23%	27%	97%	49%	87%	31%	97%	91%	100%	100%	100%	96%
ITALY	84%	54%	96%	93%	50%	43%	22%	85%	81%	97%	98%	98%	89%
JAPAN	99%	37%	84%	84%	27%	96%	67%	46%	98%	99%	100%	99%	99%
NETHERLANDS	52%	72%	53%	64%	74%	100%	81%	50%	45%	96%	97%	96%	86%
SPAIN	98%	87%	50%	96%	50%	92%	95%	89%	96%	99%	100%	100%	99%
SWEDEN	100%	94%	62%	98%	58%	74%	21%	93%	99%	100%	100%	100%	100%
USA	100%	79%	53%	96%	74%	69%	25%	57%	96%	100%	100%	100%	100%
Average	89%	74%	61%	91%	65%	74%	44%	73%	84%	99%	99%	99%	93%

6.1. Rolling Test Period

Observing Figure 3, we can generally say that the best results, in terms of MAFE, are given by the model-averaging approaches based on equal and proportional weights and on the softmax function. They all present a median MAFE lower than 0.4 and 0.006, respectively, for life expectancy and the Gini index. The RH model is the third-best model (after model-averaging approaches with equal weights and with weights based on the softmax function) for life expectancy (but with weak performance for the Gini index), while the PLAT model is the best for the Gini index (but has weak performance for life expectancy). The model-averaging approach based on trimming has good overall performance as well. Indeed, it is overcome in terms of median MAFE only by the RH model for life expectancy and the PLAT model for the Gini index. Among the other models, good results are given by the CBD model regarding life expectancy and by the CF model for the Gini index. The worst results here are found in the M7 model for life expectancy and the M6 and M8 models for the Gini index. Other models, such as LC and ACF, do not show remarkable results.

Figure 4 shows the model with the lowest MAFE by country and period. The RH model has the highest number of best performances for both life expectancy and the Gini index (34% and 26%),[4] followed by the CBD model (17% and 17%). It is interesting to notice how here the M7 model is the third-best model for life expectancy, despite it being the worst one in terms of median MAFE, and the M6 model is the third-best for the Gini index, despite being the second-worst one in terms of median MAFE. Focusing on the model-averaging approaches, it can be noted that they are seldom the best ones. Among them, the trimming averaging approach has the highest proportion of winning cases (7% and 8%). These results, even if they appear to contradict those in Figure 3, are easily explained. Indeed, model-averaging approaches have solid performance (ranking in the top five positions in terms of lowest MAFE) across all countries, periods, and metrics. On the contrary, individual models that are the best for specific combinations of countries and periods, such as M7 for life expectancy and M6 for the Gini index, are overall among the worst, as shown in Figure 3. Consequently, the advantage of model averaging is striking, in particular when several countries must be considered and models must be updated on different training periods.

Tables 2–5 show the interval forecast accuracy results. As happened for the median MAFE analysis, the model-averaging approaches based on equal and proportional weights and on the softmax function are the best ones, with an interval forecast accuracy of 97% for

the life expectancy and 98% for the Gini index. They are followed by the model-averaging approach based on trimming (93% and 94%). Among the other models, we have that the RH model has an interval forecast accuracy of 81% for the life expectancy (lower than the LC and CBD models), and the PLAT model has an interval forecast accuracy of 92% for the Gini index (the highest outside the model-averaging approaches, followed by M7 and LC models).

In the following, we report other findings from the tables and figures mentioned above, focusing particularly on the analysis by country, period, and metric considered. In Figure 4, we notice how the best model changes over time due to changes in mortality trends, such as the slowing down in mortality improvements observed in several developed countries since 2010; see Djeundje et al. (2022). Similar behaviour can also be noticed in Tables 2 and 4 for the forecast interval accuracy. In this regard, the evolution of the RH model's performance is enlightening. Indeed, it goes from being the best model for life expectancy in most periods considered to becoming the second-worst in the last test period (see Figure 5). This fact strengthens the motivation to choose a model-averaging approach that considers all or several models over a single model.

In Figure 4 and Tables 3 and 5, we observe how most individual models have a good performance in terms of interval forecast accuracy and number of cases with the lowest MAFE, for at least one country. However, there are also countries where these models show poor performance, with low interval forecast accuracy and no cases in which they are the best. This is a consequence of the fact that each country has a specific mortality trend, which is fitted better by certain models than others. On the other hand, model-averaging approaches have more robust results with good performance in all the countries.

In conclusion, observing Tables 2–5 and Figures 3 and 4, we see how the performance of the models varies substantially based on whether we consider life expectancy or the Gini index, while for the model-averaging approaches there are no large-scale variations. This means that the latter are more robust even regarding the choice of the metric considered.

6.2. Fixed Test Period

Observing Figure 5, we see how, similarly to what happened with the rolling test period case, the model-averaging approaches based on equal and proportional weights and on the softmax function have the best results in terms of MAFE. They all present a median MAFE lower than 0.2 and 0.004 for life expectancy and the Gini index, respectively. Furthermore, notable performance in terms of median MAFE is obtained with the LC model for both the Gini index and life expectancy, while here the RH model is among the worst for life expectancy, and the Plat model is now overcome by other models regarding the Gini index. The model-averaging approach based on trimming has a higher MAFE, and it is outperformed by other models such as CBD and ACF for life expectancy and M7 and ACF for the Gini index. Indeed, the trimming based averaging approach heavily relies on the RH (for life expectancy) and PLAT (for Gini index) models, which have poor performance in the test period considered.

Figure 6 shows the model with the lowest MAFE by period and country. The RH model still has the highest number of best performances for life expectancy (16%),[5] while the M7 model has the highest number for the Gini index (23%). All the model-averaging approaches here have good performance; indeed, they globally account for 40% for life expectancy and 37% for the Gini index, with the model-averaging approach with proportional weights showing the best results in both cases (12% and 17%). Other remarkable results are obtained with the CBD and M7 models for life expectancy, and the PLAT model regarding the Gini index. As in the rolling period case, the apparent discrepancy between between some results in Figure 5 and those in Figure 6 is explained by the superior robustness of the averaging approaches over the individual models.

Tables 6–9 show the proportion of cases in which the observed metrics fall in the forecasting intervals. As was the case for the rolling test period case, the model-averaging approaches based on equal and proportional weights and on the softmax function have the

highest interval forecast accuracy, with 98% for life expectancy and 99% for the Gini index. They are followed by the model-averaging approach based on trimming (94% and 93%). Among the other models, we find that the LC (90%) and ACF (87%) models show good results for life expectancy, while the M7 (91%) and LC (89%) models show good results for the Gini index. The RH model here is the worst model for life expectancy forecast accuracy as well as the Gini index (43% and 44%), while the PLAT model performance decreases for the Gini index (from 92% to 74%). This is an additional demonstration that it is not always optimal to rely on models that performed well previously.

To conclude, in Tables 6–9 and Figures 5 and 6, we notice that, compared with the rolling test period case, there is less variability in the models' performance by training period as the test period remains fixed. A similar remark can be made with respect to the results by life expectancy and the Gini index. Instead, regarding the results by country, we observe how the performance variability remains, with many individual models performing well in some countries and badly in others. In contrast, the model-averaging approaches, except for the one based on trimming that is slightly weaker, show good results in all the countries considered.

7. Conclusions

In this paper, we compared the forecast performance of existing two-population models with those of four different model-averaging approaches. We considered ten countries and 46 different combinations of training and test periods, using truncated life expectancy and the Gini index as metrics. Our results show that model-averaging approaches outperformed the individual models, achieving superior results both in terms of MAFE (difference between the forecasted central truncated life expectancy and Gini index and the observed ones) and interval forecast accuracy (proportion of cases in which the observed values fall within the prediction interval). Among the model-averaging approaches, the best results are given by the ones with equal weights, proportional weights, and weights based on the the softmax function, while the one based on trimming shows a poorer performance, although it is still better than many individual models. The fact that the model-averaging approach based on trimming produces a higher MAFE on average is likely to be a consequence of changes in mortality trends: the models that perform well in the validation period do not necessarily perform equally well in the test period due to the speed up or slow down in the mortality improvements over the years. Finally, a further advantage of model-averaging approaches is that, while the performance of individual models is heavily affected by the choice of the metric (life expectancy or the Gini index), country, and period, the performance of the model-averaging approaches is shown to be more robust concerning this choice.

In terms of extensions of this research, a straightforward development could be considering multi-population models that simultaneously consider three or more populations rather than just two-population models. Furthermore, alternative measures of mean absolute forecasting error and interval forecast accuracy could be considered for comparing the goodness of forecasts of model-averaging approaches and traditional models.

Author Contributions: Conceptualization, P.M.; methodology, L.D.M., S.H. and P.M.; software, L.D.M. and P.M.; formal analysis, L.D.M.; investigation, L.D.M., S.H., P.M. and R.Z.; data curation, L.D.M.; writing—original draft preparation, L.D.M.; writing—review and editing, S.H., P.M. and R.Z.; supervision, S.H., P.M. and R.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

[1] Probabilities of death q_x can be calculated from the corresponding mortality rates m_x by using the relation $q_x = m_x/(1 + \frac{1}{2}m_x)$, and vice versa, $m_x = q_x/(1 - \frac{1}{2}q_x)$.

[2] For consistency, we use the Poisson distribution assumption coupled with the log-link function and mortality rates for models such as M6, M7, and M8, which usually are presented under a binomial assumption coupled with the logit-link function and probabilities of death.

[3] We carry out more simulations than in the first stage since here we consider the interval forecast accuracy in addition to the MAFE.

[4] These percentages have been calculated as the ratio of the number of cases in which each model is the best over the total number of cases considered (260).

[5] These percentages have been calculated as the ratio of the number of cases in which each model is the best over the total number of cases considered (200).

References

Benchimol, Andrés Gustavo, Pablo J. Alonso, Juan Miguel Marín Díazaraque, and Irene Albarrán Lozano. 2016. Model Uncertainty Approach in Mortality Projection with Model Assembling Methodologies. Available online: https://e-archivo.uc3m.es/rest/api/core/bitstreams/cd474c81-da11-4ee0-92f2-04be244fc37e/content (accessed on 8 March 2024).

Cairns, Andrew J. G., David Blake, and Kevin Dowd. 2006. A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance 73*: 687–718.

Cairns, Andrew J. G., David Blake, Kevin Dowd, Guy D. Coughlan, David Epstein, Alen Ong, and Igor Balevich. 2009. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal* 13: 1–35. [CrossRef]

Dickson, David C. M., Mary R. Hardy, and Howard R. Waters. 2019. *Actuarial Mathematics for Life Contingent Risks*. Cambridge, MA: Cambridge University Press.

Djeundje, Viani B., Steven Haberman, Madhavi Bajekal, and Joseph Lu. 2022. The slowdown in mortality improvement rates 2011–2017: A multi-country analysis. *European Actuarial Journal* 12: 839–78. [CrossRef]

Dowd, Kevin, Andrew J. G. Cairns, David Blake, Guy D. Coughlan, David Epstein, and Marwa Khalaf-Allah. 2010. Evaluating the goodness of fit of stochastic mortality models. *Insurance: Mathematics and Economics* 47: 255–65. [CrossRef]

Dowd, Kevin, Andrew J. G. Cairns, David Blake, Guy D. Coughlan, and Marwa Khalaf-Allah. 2011. A gravity model of mortality rates for two related populations. *North American Actuarial Journal* 15: 334–56. [CrossRef]

Enchev, Vasil, Torsten Kleinow, and Andrew J. G. Cairns. 2017. Multi-population mortality models: fitting, forecasting and comparisons. *Scandinavian Actuarial Journal* 2017: 319–42. [CrossRef]

Fletcher, David. 2018. *Model Averaging*. Berlin/Heidelberg: Springer.

Hinne, Max, Quentin F. Gronau, Don van den Bergh, and Eric-Jan Wagenmakers. 2020. A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science* 3: 200–15. [CrossRef]

Lee, Ronald D., and Lawrence R. Carter. 1992. Modeling and forecasting US mortality. *Journal of the American Statistical Association* 87: 659–71.

Li, Jackie. 2013. A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. *Population Studies* 67: 111–26. [CrossRef] [PubMed]

Li, Johnny Siu-Hang, Rui Zhou, and Mary Hardy. 2015. A step-by-step guide to building two-population stochastic mortality models. *Insurance: Mathematics and Economics* 63: 121–34. [CrossRef]

Li, Nan, and Ronald Lee. 2005. Coherent mortality forecasts for a group of populations: An extension of the Lee–Carter method. *Demography* 42: 575–94. [CrossRef]

Plat, Richard. 2009. On stochastic mortality modeling. *Insurance: Mathematics and Economics* 45: 393–404.

Renshaw, Arthur E., and Steven Haberman. 2006. A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics* 38: 556–70. [CrossRef]

Samuels, Jon D., and Rodrigo M. Sekkel. 2017. Model confidence sets and forecast combination. *International Journal of Forecasting* 33: 48–60. [CrossRef]

Shang, Han Lin. 2012. Point and interval forecasts of age-specific life expectancies: A model averaging approach. *Demographic Research* 27: 593–644. [CrossRef]

Shang, Han Lin, Heather Booth, and Rob J. Hyndman. 2011. Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. *Demographic Research* 25: 173–214. [CrossRef]

Shang, Han Lin, Steven Haberman, and Ruofan Xu. 2022. Multi-population modelling and forecasting life-table death counts. *Insurance: Mathematics and Economics* 106: 239–53. [CrossRef]

Shkolnikov, Vladimir M., Evgueni E. Andreev, and Alexander Z. Begun. 2003. Gini coefficient as a life table function: Computation from discrete data, decomposition of differences and empirical examples. *Demographic Research* 8: 305–58. [CrossRef]

Villegas, Andrés, Pietro Millossovich, and Vladimir Kaishev. 2018. StMoMo: Stochastic mortality modeling in R. *Journal of Statistical Software* 84: 1–38. [CrossRef]

Villegas, Andrés M., Steven Haberman, Vladimir Kaishev, and Pietro Millossovich. 2017. A comparative study of two-population models for the assessment of basis risk in longevity hedges. *ASTIN Bulletin: The Journal of the IAA* 47: 631–79. [CrossRef]
Yang, Bowen, Jackie Li, and Uditha Balasooriya. 2016. Cohort extensions of the Poisson common factor model for modelling both genders jointly. *Scandinavian Actuarial Journal* 2016: 93–112. [CrossRef]