ARC Centre of Excellence in Population Ageing Research

Working Paper 2021/09

Mortality Forecasting Using Stacked Regression Ensembles

Salvatory R. Kessy, Michael Sherris, Andrés M. Villegas and Jonathan Ziveyi

---

Mortality Forecasting Using Stacked Regression Ensembles

Salvatory R. Kessy*, Michael Sherris, Andrés M. Villegas, Jonathan Ziveyi

School of Risk and Actuarial Studies and ARC Centre of Excellence in Population Ageing Research (CEPAR)

University of New South Wales, Sydney

Abstract

We present a stacked regression ensemble method that optimally combines different mortality models to reduce the mean squared errors of mortality rate forecasts and mitigate model selection risk. Stacked regression uses a supervised machine learning algorithm to approximate the horizon-specific weights by minimizing the cross-validation criterion for each forecasting horizon. The horizon-specific weights facilitate the development of a mortality model combination customized to each horizon. Unlike other model combination methods, stacked regression simultaneously solves model selection and estimates model combinations to improve model forecasts. Our numerical illustrations based on 44 populations from the Human Mortality Database demonstrate that stacking mortality models increases predictive accuracy. Using one-year-ahead to 15-year-ahead out-of-sample mean squared errors, we find that stacked regression improves mortality forecast accuracy by 13% - 49% and 19% - 90% over the individual mortality models for males and females, respectively. Therefore, combining the mortality rate forecasts provides lower out-of-sample point forecast errors than selecting the single best individual mortality method. Stacked regression ensemble also achieves better predictive accuracy than other model combination methods, namely Simple Model Averaging, Bayesian Model Averaging, and Model Confidence Set. Our results support the stacked regression ensemble approach over individual mortality models and other model combination methods in forecasting mortality rates. We also provide a user-friendly open-source R package, `CoMoMo`, that combines multiple mortality rate forecasts using different model combination techniques.

Keywords. Stacked regression, ensemble learning, cross-validation, model uncertainty, model combination, age-period-cohort model, mortality forecasting.

1. Introduction

Life expectancy has been continuously increasing globally since the nineteenth century due to the decline in mortality rates. In particular, life expectancy at birth has risen approximately linearly by three months per year in the past four decades for several industrialized countries (Oeppen and Vaupel 2002). The decrease in mortality rates across ages and years is mainly due to enormous efforts dedicated to preventing and curing diseases through advances in public health, nutrition, and medical discoveries (Janssen 2018).

Longevity risk is a potential risk attached to the unexpected increasing life expectancy of policyholders. This risk obliges life annuity providers and defined-benefit pension funds to pay more benefits and pensions than expected. Longevity risk is also a risk to individuals who may outlive their retirement resources. The estimated potential size of the global longevity risk market for pension liabilities is between USD 60 trillion and USD 80 trillion (Blake et al. 2018).

Mortality models have consistently underestimated future longevity because they are often based on the assumption that the pace of mortality improvement would decline (Shaw 2007). However, in practice, life expectancy has been rising globally. Longevity risk is accentuated by prevailing low equity returns, low-interest rates, and volatility in financial markets resulting in the unavailability of suitable assets for matching liabilities that come with aging populations (Berdin and Gründl 2015). These concerns about longevity risk have led to a growing interest among practitioners and researchers in accurately modelling and forecasting mortality rates and capturing their corresponding uncertainty. The accurate forecasting of future mortality quantities is a challenging task as it depends on unknown future events such as medical advances and disease outbreaks like the COVID-19 pandemic. Even a small improvement in point and interval mortality forecast accuracy can provide considerable financial savings among financial entities. This study exploits the power of advanced data analytics methods in improving the forecast accuracy of mortality rate forecasts.

*Email: s.kessy@unsw.edu.au

Despite the diversity of existing mortality models, the "single best" mortality model is customarily used to forecast mortality rates (Booth and Tickle 2008). This single mortality model framework overlooks model choice uncertainty and can make decision-makers overconfident in the mortality rate forecasts from the single best mortality model. The presence of model uncertainty and the possible limitations of modelling the mortality rates using the single model approach are supported theoretically and empirically. Firstly, given the complexity of mortality data, there is no single stochastic mortality that performs consistently better than others, either in-sample or out-of-sample for all mortality data and forecasting horizons (Cairns et al. 2009; Rabbi and Mazzuco 2018; SriDaran et al. 2021). For example, mortality models that accurately predict short-term mortality rates tend to perform relatively poorly in longer-horizon forecasts (SriDaran et al. 2021). Secondly, there is no one accepted model selection criteria to evaluate mortality models. Several model selection criteria are based on different assumptions and may yield contradicting model choices (Kourentzes, Barrow, and Petropoulos 2019). Therefore, it is hard to select the model selection criteria to choose the single best mortality model (Atance, Debón, and Navarro 2020). Comparing different variants and extensions of the mortality models based on various assumptions does not always select the single best mortality model (Janssen 2013). In such situations, recommending a mortality model can be challenging. This underscores the benefits of model averaging, which empirically combines predictions from composite mortality models and reduces model choice uncertainty (Bates and Granger 1969).

A model combination approach is an alternative approach to model selection. It has been employed for more than 50 years in other forecasting fields to reduce model selection risk and improve forecast accuracy (Bates and Granger 1969; Genre et al. 2013). A few mortality forecasting studies have implemented the model averaging approach (Shang 2012; Kontis et al. 2017; Shang and Haberman 2018; Shang and Booth 2020; Barigou et al. 2021). Shang and Haberman (2018) combine multiple mortality models using the model confidence set approach proposed in Hansen, Lunde, and Nason (2011) to forecast national and sub-national Japanese mortality data. This technique selects a set of statistically superior mortality models from a predetermined set of models based on their in-sample forecast errors which are then combined using equal weights.

Kontis et al. (2017) use probabilistic Bayesian model averaging (BMA) to combine age-specific death rates from 21 mortality models applied to 35 industrialized countries. They choose different mortality models to capture different mortality rate dynamics such as trends, linearity, non-linearity, and cohort effects. These mortality models are then weighted depending on their in-sample predictive power. Kontis et al. (2017) find that, on average, the BMA approach has a smaller forecast error than the best single mortality model for different genders and countries. Barigou et al. (2021) also recently proposed a fully Bayesian approach for mortality forecasting using a validation set made of the most recent years. They show that BMA methods, based on an out-of-sample criterion, outperform in terms of prediction performance and robustness, the standard BMA which calculates the weights using a marginal likelihood approximation.

The model combination approaches proposed to date in the mortality projection literature have some limitations. Firstly, the model combination weights are often calculated using a validation set with the potential of overfitting (James et al. 2014). Therefore, overparameterized models are assigned higher weights, but they tend to perform poorly in longer forecasting horizons (Makridakis and Hibon 2000). Secondly, the out-of-sample forecast accuracy of BMA depends sensitively on the correct choice of the prior distribution of each model (Yao et al. 2018), and the true data-generating model should be in the list of models to be combined (Clarke 2004). Thirdly, the existing approaches use the same weights for all forecasting horizons, implicitly assuming that various mortality models have the same predictive power across all the forecasting horizons. However, the performance of different mortality models tends to vary with the forecasting horizon (SriDaran et al. 2021). This means that the forecasting accuracy of model combination approaches could be enhanced using horizon-specific weights (Shang and Booth 2020). Finally, while all the studies report that model averaging attained smaller forecast errors than the single mortality models, they do not examine whether the differences in point forecast accuracies between the individual mortality models and model combination methods are statistically significant.

Stacked regression ensembles, first proposed by Wolpert (1992), offer an alternative model combination approach that could overcome some of the shortcomings of traditional model combination methods. Stacked regression combines several diverse individual models into a single powerful prediction function via a secondary learning process known as meta-learning (Breiman 2004). Stacked regression can learn the complex patterns in the data, reduce model uncertainty, and enhance prediction accuracy (Wolpert 1992). When the true-data generating model is not amongst the list of models to be combined, Clarke (2004) empirically shows that stacked regression has a smaller out-of-sample point forecast error than BMA. This is because stacked regression is less dependent on the likelihood, and therefore it is more representative of the true data-generating model. Moreover, stacked regression assigns weights to the individual models based on the cross-validation technique, while BMA assigns weights to the individual models based

on the posterior probabilities (Yao et al. 2018). The combination weights from stacked regression reflect the ability of the individual model to generalize on new data, while BMA weights reflect the goodness-of-fit to the data. Owing to empirical and theoretical benefits of stacked regression, most winning teams in data science competitions have used such an approach (Sill et al. 2009; Puurula, Read, and Bifet 2014). The results of past data science competitions show that combined methods regularly outperform individual models in terms of out-of-sample predictive accuracy (Makridakis and Hibon 2000). Stacked regression ensemble methods have also been successfully applied with improved predictive accuracy on a wide range of problems such as credit risk assessment (Doumpos and Zopounidis 2007), forecasting global energy consumption (Khairalla et al. 2018), financial time series forecasting (Ma and Dai 2016), and prediction of infectious disease epidemics (Ray and Reich 2018; Reich et al. 2019).

This paper aims to develop and evaluate a stacked regression ensemble forecasting approach that combines forecasts from different stochastic mortality models. Despite the success of the stacked regression ensemble methods in other fields, to the best of our knowledge, they have not been explored previously in mortality modelling and forecasting literature. This paper seeks to fill this gap. In particular, this paper aims at answering the following research questions empirically:

- Does the stacked regression ensemble have lower out-of-sample point forecast errors than individual mortality models?
- Does the stacked regression ensemble have lower out-of-sample point forecast errors than other model combination techniques?
- Do the optimal combination weights estimated based on the forecasting horizons reduce the out-of-sample point forecast errors of model combination methods?
- Do individual mortality models and model combination methods have statistically different out-of-sample point forecast errors?

As implied by our research questions, this paper contributes to the existing mortality forecasting literature in the following three dimensions. Firstly, it introduces a new approach of learning model combination weights for each mortality model using the stacked regression ensemble method to improve model predictions. Secondly, it develops a mortality model combination approach that is dependent on the forecasting horizon. Finally, We develop a user-friendly open-source R package, `CoMoMo`, that combines multiple mortality rate forecasts using different model combination techniques.

The rest of the paper is organized as follows. In Section 2, we define the primary notation used to refer to mortality data. Section 3 briefly describes the properties of the individual mortality models, which underpin the different model combination approaches discussed in this paper. Section 4 discusses different metrics for evaluating the individual mortality models and cross-validating the mortality data. Section 5 discusses how to estimates the combination weights using different model combination methods. In Section 6, we introduce how to implement the stacked regression ensemble framework to combine multiple mortality models. In Section 7, we report on the application of different model combination approaches to 44 populations from the Human Mortality Database (University of California Berkeley and Max Planck Institute for Demographic Research 2020). Using one-year-ahead to 15-year-ahead out-of-sample mean squared errors, we find that stacked regression improves mortality forecast accuracy by 13% - 49% and 19% - 90% over the individual mortality models for males and females, respectively. Therefore, combining the mortality rate forecasts provides lower out-of-sample point forecast errors than selecting the single best individual mortality method. Finally, Section 8 concludes with the key findings of this paper and proposes possible future directions.

2. Notation

Let calendar year t run from time t to $t+1$ and let $\mathrm{D}(x,t)$ denote the number of deaths aged x at previous birthday during calendar year t. The actual observed number of deaths is denoted by $d(x,t)$ and the corresponding central exposed at risk by $E^c(x,t)$. The death and central exposure data are arranged in matrices $\mathbf{D} = (d(x,t))$ and $\mathbf{E}^c = E^c(x,t)$, respectively. Each matrix has dimension $n_a \times n_y$ with n_a ages, n_y years, and $n_b = n_a + n_y - 1$ cohorts. The force of mortality matrix can then be estimated with $\boldsymbol{\mu} = \left(\frac{d(x,t)}{E^c(x,t)}\right)$.

3. Single Mortality Models

We combine different individual mortality models that have been used in the literature on modelling and forecasting the mortality rates. Multiple mortality models capture various death rate dynamics, such as trends, linearity, non-linearity, curvature, mortality volatility, and cohort effects (Kontis et al. 2017). It is computationally expensive to combine a large number of mortality models in practice (Sharabiani and Mahani 2016). Therefore, we focus on the popular family of Generalized Age-Period-Cohort (GAPC) discrete-time mortality models formalized in Villegas, Kaishev, and Millossovich (2018). GAPC models decompose the force of mortality, $\mu(x,t)$ for age $x \in [x_1, x_{n_a}]$ at time $t \in [t_1, t_{n_y}]$ across the dimensions of age x, period t, and cohort $c \in \{t_1 - x_{n_a}, \ldots, t_{n_y} - x_1\}$. We assume that the observed number of deaths, $\mathrm{D}(x,t)$ are distributed according to a Poisson distribution so that $\mathrm{D}(x,t) \sim \text{Poisson}(E^c(x,t)\mu(x,t))$ (Brouhns, Denuit, and Vermunt 2002). Therefore, the GAPC models are represented as

$$\ln \mu(x,t) = \alpha(x) + \sum_{i=1}^{N} f^i(x)\kappa^i(t) + \gamma(c), \tag{1}$$

where $\alpha(x)$ is the age pattern of the log mortality rates averaged across years, N is the number of age-period terms describing the mortality trends using the time index $\kappa(t)$, $f^i(x)$ is an age-modulating function, and $\gamma(c)$ is the cohort factor. We consider various forms of GAPC mortality models summarized in Table 1. LC denotes the Lee-Carter model, CBD the Cairns-Blake-Dowd model, APC the age-period-cohort model, RH the Renshaw-Haberman model, mRH the modified Renshaw-Haberman model, M7 the quadratic CBD model, and PLAT the Plat model.

Table 1: Generalized age-period-cohort mortality models. Here, $\alpha(x)$ captures the general shape of the mortality by age, $\forall i = 1,2,3, \kappa^{(i)}(t)$ is the time index which specifies the mortality trend, γ_c is the cohort effects at the year of birth $c = t - x$, $\forall i = 0,1, \beta^{(i)}(x)$ measures the effect of $\kappa^{(1)}(t)$ or γ_c across ages, $\bar{x}$ is the average age in the sample range, $\widehat{\sigma}^2(x)$ is the average value of $(\bar{x} - x)^2$, and $(\bar{x} - x)^+ = \max(\bar{x} - x, 0)$.

Name	Model	Constraints
LC	$\ln \mu(x,t) = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)}$	$\sum_t \kappa^{(1)}(t) = 0, \sum_x \beta^{(1)}(x) = 1$
RH	$\ln \mu(x,t) = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(0)}\gamma_c$	$\sum_t \kappa^{(1)}(t) = \sum_x \gamma^{(1)}(c) = 0, \sum_x \beta^{(1)}(x) = \sum_x \beta^{(0)}(x) = 1$
mRH	$\ln \mu(x,t) = \alpha_x + \beta_x^{(1)}\kappa_t^{(1)} + \gamma_c$	$\sum_t \kappa^{(1)}(t) = \sum_x \gamma^{(1)}(c) = 0, \sum_x \beta^{(1)}(x) = 1$
APC	$\ln \mu(x,t) = \alpha_x + \kappa_t^{(1)} + \gamma_c$	$\sum_t \kappa^{(1)}(t) = \sum_x \gamma^{(1)}(c) = \sum_x c\gamma^{(1)}(c) = 0$
CBD	$\ln \mu(x,t) = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)}$	Not required
M7	$\ln \mu(x,t) = \kappa_t^{(1)} + (x - \bar{x})\kappa_t^{(2)} + \left((x - \bar{x})^2 - \hat{\sigma}_x^2\right)\kappa_t^{(3)} + \gamma_c$	$\sum_t \kappa^{(1)}(t) = \sum_t \kappa^{(2)}(t) = \sum_t \kappa^{(3)}(t) = \sum_x \gamma^{(1)}(c) = \sum_x c^2\gamma^{(1)}(c) = 0, \sum_x \beta^{(1)}(x) = 1$
PLAT	$\ln \mu(x,t) = \alpha_x + \kappa_t^{(1)} + (\bar{x} - x)\kappa_t^{(2)} + (\bar{x} - x)^+\kappa_t^{(3)} + \gamma_c$	$\sum_t \kappa^{(1)}(t) = \sum_t \kappa^{(2)}(t) = \sum_t \kappa^{(3)}(t) = \sum_x \gamma^{(1)}(c) = \sum_x c^2\gamma^{(1)}(c) = 0, \sum_x \beta^{(1)}(x) = 1$

GAPC models reflect different structures, assumptions, degrees of complexity, and extract different information from the mortality data. The LC model is perhaps the most widely used mortality model for predicting mortality rates (Lee and Carter 1992). Nevertheless, the LC model can not fully capture non-linear patterns in the mortality data. Therefore, Renshaw and Haberman (2006) extend LC model to RH by adding the cohort effect $\gamma^{(1)}(c)$ which captures the life experiences and life styles of people born at different time periods. When $\beta^{(0)}(x) = 1$, the RH model becomes mRH which is more stable than RH because of its simpler structure (Hunt and Villegas 2015) and we therefore implement mRH in this paper. The RH model does experience a much better fit than LC for mortality data with dominant cohort effects (Cairns et al. 2009). The APC model is derived from RH model by setting $\beta^{(1)}(x) = \beta^{(0)}(x) = 1$. The APC tends to produce more robust parameter estimates than RH with respect to the data changes (Cairns et al. 2009).

Cairns, Blake, and Dowd (2006) propose the CBD model which is an alternative parametric mortality method which models and forecasts the mortality rates using a two-factor model. In contrast to the non-parametric age structure in LC, RH, and APC, the CBD model treats age as a continuous variable that varies linearly with $\ln \mu(x,t)$. CBD has two time factors $\kappa^{(1)}(t)$ and $\kappa^{(2)}(t)$ that allows for substantial period effects than the LC, RH, and APC models. Cairns et al. (2009) generalize the CBD model to M7 to capture the curvature in the mortality rates by age using a quadratic age term. M7 captures the time variability more adequately than the LC, RH, APC, and CBD models using a three time factors, $\kappa^{(1)}(t)$, $\kappa^{(2)}(t)$, and $\kappa^{(3)}(t)$. It further includes the quadratic age term $(\bar{x} - x)^2 - \widehat{\sigma}^2(x)$ that capture the curvature along the age axis and $\gamma^{(1)}(c)$ to model the cohort effects. Plat (2009) combines the features of LC and CBD to model the mortality rates.

3.1. Fitting GAPC Models

The parameters of GAPC are estimated by assuming a Poisson distribution of the deaths $\mathrm{D}(x,t)$. The log-likelihood for GAPC models is presented in Villegas, Kaishev, and Millossovich (2018) as

$$\mathcal{L}(d(x,t)) = \sum_x \sum_t \omega(x,t) \left(d(x,t) \ln \widehat{d}(x,t) - \widehat{d}(x,t) - \ln d(x,t)! \right),$$

where the expected number of the deaths for the GAPC models are given as

$$\widehat{d}(x,t) = E(x,t) e^{\left(\alpha(x) + \sum_{i=1}^{N} f^i(x)\kappa^i(t) + \gamma(c)\right)},$$

and the weights $\omega(x,t)$ are given as

$$\omega(x,t) = \begin{cases} 0 & \text{if the } (x,t) \text{ data cell is excluded,} \\ 1 & \text{if the } (x,t) \text{ data cell is included.} \end{cases}$$

3.2. Mortality Projection with GAPC Models

The dynamics of the GAPC model are driven by the period indexes $\kappa^i(t), i = 1, \ldots, N$ and by the cohort index $\gamma(c)$. Therefore, to forecast the mortality rates, we need to model these indexes using the appropriate time series methods. Villegas, Kaishev, and Millossovich (2018) model the period indexes using a multivariate random walk with drift which can be expressed as

$$\boldsymbol{\kappa}(t) = \boldsymbol{\kappa}(t)(t-1) + \boldsymbol{\delta} + \boldsymbol{\epsilon}(t), \quad \boldsymbol{\epsilon}(t) \sim \mathbb{N}(0, \Sigma), \tag{2}$$

where $\boldsymbol{\kappa}(t) = [\kappa^1(t), \ldots, \kappa^N(t)]'$, Σ is the $N \times N$ variance-covariance matrix of the multivariate white noise $\boldsymbol{\epsilon}(t)$, and $\boldsymbol{\delta}$ is an N dimension vector of drift parameters. Villegas, Kaishev, and Millossovich (2018) model $\boldsymbol{\gamma}(c)$ using an autoregressive integrated moving average process ARIMA(p,q,d) with drift which is independent of the time index $\boldsymbol{\kappa}^i(t)$

$$\Delta^d \boldsymbol{\gamma}(c) = \delta(0) + \phi_1 \Delta^d \boldsymbol{\gamma}(c-1) + \ldots + \phi_p \Delta^d \boldsymbol{\gamma}(c-p) + \epsilon(c) + \phi_1 \epsilon(c-1) + \ldots + \phi_q \epsilon(c-q), \tag{3}$$

where $\epsilon(c)$ is a normal white noise process with variance σ_ϵ. In this paper, we implement an ARIMA$(1,1,0)$ because when $d = 1$, we avoid model identification issues (SriDaran et al. 2021). We set $p = 1$ because there is usually some autoregression among the cohorts.

The projected values of the period index $\widehat{\boldsymbol{\kappa}}(t_{n_y}+h) = [\widehat{\kappa}^1(t_{n_y}+h), \ldots, \widehat{\kappa}^N(t_{n_y}+h)]'$ and cohort index $\widehat{\boldsymbol{\gamma}}(t_{n_y}+h-x), h = 1, \ldots, H$ are extrapolated using Equations (2) and (3), respectively. Then, the mortality rate forecasts at horizon h are expressed as

$$\ln \widehat{\mu}(x, t+h) = \alpha(x) + \sum_{i=1}^{N} f^i(x) \widehat{\boldsymbol{\kappa}}^i(t_{n_y}+h) + \widehat{\boldsymbol{\gamma}}(t_{n_y}+h-x),$$

where t_{n_y} is the last year of the fitting period.

We use the `StMoMo` R package to fit and forecast the mortality rates using GAPC mortality models (Villegas, Kaishev, and Millossovich 2018).

4. Model Selection Criteria

It is challenging to predict mortality rates in the presence of multiple mortality models because there is no widely accepted best method to choose the best mortality model (Atance, Debón, and Navarro 2020). Multiple model selection methods are based on various assumptions and may yield contradicting model choices (Kourentzes, Barrow, and Petropoulos 2019). Therefore, the computation of model weights in different model combination approaches is underpinned by different model selection criteria. We group different model selection methods as to whether they focus on the in-sample goodness-of-fit of a model or the ability of the model to generalize on the unseen data.

4.1. In-Sample Model Evaluation

The in-sample goodness-of-fit criteria assess the performance of a model in relation to how well it explains the data. For example, the Akaike Information Criterion (AIC) selects the best model by balancing the quality of fit measured

using the likelihood function and model complexity measured using the number of parameters (Akaike 1974). The AIC of any mortality models is mathematically defined as

$$\text{AIC} = -2\mathcal{L} + 2\nu, \tag{4}$$

where $\mathcal{L}$ is the maximized value of the likelihood function and ν is number of parameters in a particular model. The AIC tends to select overparameterized models (Wagenmakers and Farrell 2004). As an alternative, Schwarz (1978) propose the Bayesian Information Criterion (BIC) that imposes a stronger penalty for the model complexity. The BIC of any mortality models is mathematically defined as

$$\text{BIC} = -2\mathcal{L} + \nu \ln(n), \tag{5}$$

where n is the number of observations. A comparison of AIC in Equation (4) and BIC in Equation (5) indicates that BIC penalty term is larger than the AIC penalty term when $\ln n > 2$. Therefore, BIC often selects simpler models than AIC. The models with the lower values of both criteria are generally preferred. Both criteria tend to choose models that adhere to the historical data but do not guarantee accurate out-of-sample forecasts, especially for longer horizons (Cairns et al. 2011; SriDaran et al. 2021). Furthermore, both criteria do not enable one to choose a mortality model tailored to a particular forecasting horizon.

4.2. Out-of-Sample Model Evaluation

An out-of-sample model evaluation of forecasting accuracy starts with splitting the mortality data into train and test periods. The final period in the training period is known as the forecasting origin, and the number of periods between the forecasting origin and time being forecast is the forecasting horizon. We can use either a single forecasting origin called fixed-origin evaluation or multiple forecasting origin such as a rolling window approach to perform the out-of-sample evaluation (Tashman 2000). The fixed-origin evaluation uses mortality data from t_1 to $t^*_{n_y}$ for ages $x \in [x_1, x_{n_a}]$ to estimate the model parameters, which are then used to forecast the mortality rates from $t^*_{n_y}+1$ to t_{n_y} (Kontis et al. 2017). We measure the forecasting accuracy of the model from $t^*_{n_y}+1$ to t_{n_y} for ages $x \in [x_1, x_{n_a}]$ using

$$\theta(\widehat{\mu}(x,t), \mu(x,t)) = \frac{1}{n_a n^*_y} \sum_{t=t^*_{n_y}+1}^{t_{n_y}} \sum_{x=x_1}^{x_{n_a}} \pi(\widehat{\mu}(x,t), \mu(x,t)), \tag{6}$$

where $\theta(\widehat{\mu}(x,t), \mu(x,t))$ is the forecasting error, $n^*_y = t_{n_y} - t^*_{n_y}$, and $\pi(\widehat{\mu}(x,t), \mu(x,t))$ is the respective loss function. The common loss functions are the quadratic loss, $(\ln \widehat{\mu}(x,t) - \ln \mu(x,t))^2$, which gives the mean squared error (MSE); the mean absolute loss, $|\ln \widehat{\mu}(x,t) - \ln \mu(x,t)|$, which yields the mean absolute error (MAE); and the bias loss, $\ln \widehat{\mu}(x,t) - \ln \mu(x,t)$, which yields the projection bias.

The fixed-origin evaluation generates only one forecast for each forecasting horizon, and therefore it needs a reasonably long test period to yield a forecasting track-record (Tashman 2000). The mortality forecasts produced from the fixed-origin evaluation may be influenced by occurrences unique to that origin. The fixed-origin evaluation also does not allow us to assess the forecasting accuracy of a mortality model at each forecasting horizon. We can overcome the shortcomings of the fixed-origin evaluation using a rolling window approach similar to Dowd et al. (2008). For instance, as depicted in Figure 1, we train the mortality models using all the in-sample data shown in blue and predict the unseen validation data $\mathcal{V}$ shown in red. Suppose that the validation data contains n_E number of years. We then roll forward one period of $n_E - h + 1$ times until all the data is finished. We extrapolate the period indices for horizon h using a multivariate random walk with drift given in Equation (2). The predictive power of a mortality model at horizon h is evaluated on $\mathcal{V}$ using

$$\theta^h(\widehat{\mu}(x,t), \mu(x,t)) = \frac{1}{n_a(n_E - h + 1)} \sum_{j=1}^{n_E-h+1} \sum_{(x,t)\in\mathcal{V}^j(h)} \pi(\widehat{\mu}(x,t), \mu(x,t)), \tag{7}$$

where $\mathcal{V}^j(h) = \{(x,t)\}_{t=n_T+j-1}$ for $j = 1, \ldots, n_E - h + 1$. We compute the evaluation metrics such as mean squared errors, mean absolute errors, and projection bias using their respective loss functions in Equation (7).

4.3. Resampling and Block Cross-Validation of Mortality Data

The out-of-sample evaluation approach described above uses a single validation period at the end of the data. To make a better use of the available data we can use cross-validation techniques to evaluate the models. Cross-validation is the

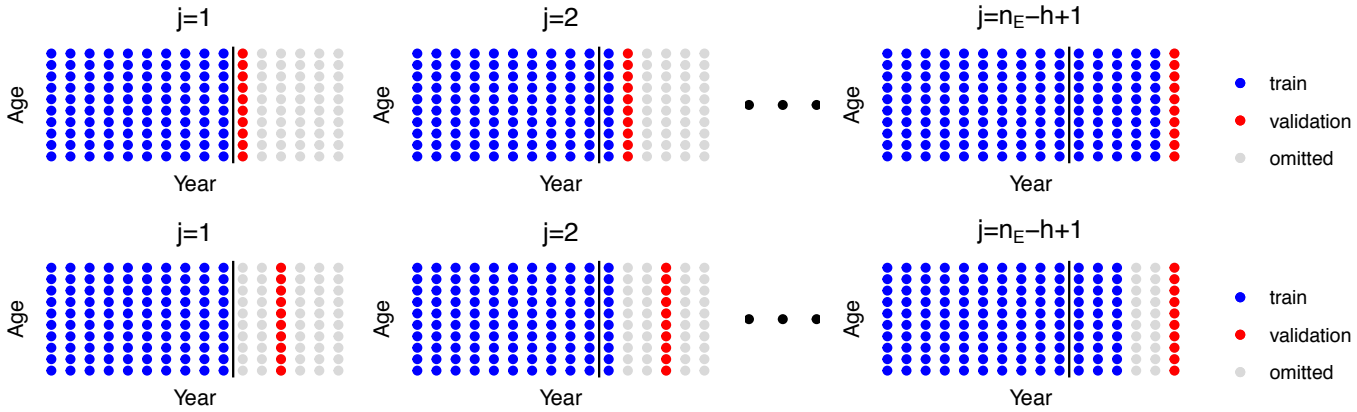

Figure 1: Iterations of rolling window out-of-sample evaluation for horizon one-year-ahead ($h = 1$) (top row) and three-years-ahead ($h = 3$) (bottom row).

resampling approach that randomly partitions the available data into k folds of roughly equal size. A model is trained k times, using $k - 1$ folds as a fitting set and the one remaining fold to evaluate the out-of-sample performance of the model using MSE or MAE. This is repeated until all k folds have been employed as out-of-sample. The hold-out performance is then averaged, yielding the cross-validation error. Cross-validation assumes the data are identically and independently distributed. This assumption does not hold on time series data such as mortality data. Therefore, we introduce block cross-validation, which preserves the dependency structure in time series data (Bergmeir, Costantini, and Benítez 2014).

Block cross-validation is an estimator used to assess and evaluate the accuracy of the forecasting methods over a given forecasting horizon for the time series data. It yields a lower forecast error than the standard out-of-sample scheme because it ensures full data use (Bergmeir, Costantini, and Benítez 2014). SriDaran et al. (2021) implement the block cross-validation method for mortality data using the following steps. Firstly, they split the mortality data into the in-sample data for model training and validation set for model performance testing. Secondly, the in-sample data is iteratively further divided into training data to fit the individual mortality models, known as base learners, and test data to evaluate their forecasting performances. The testing data can take different widths to represent varying forecasting horizons. For example, when forecasting one-year-ahead mortality rates, test data should be defined as a one-year block, as illustrated in Figure 2 (top row). While when forecasting three-years-ahead mortality rates, test data should be defined as a three-year block as illustrated in Figure 2 (bottom row).

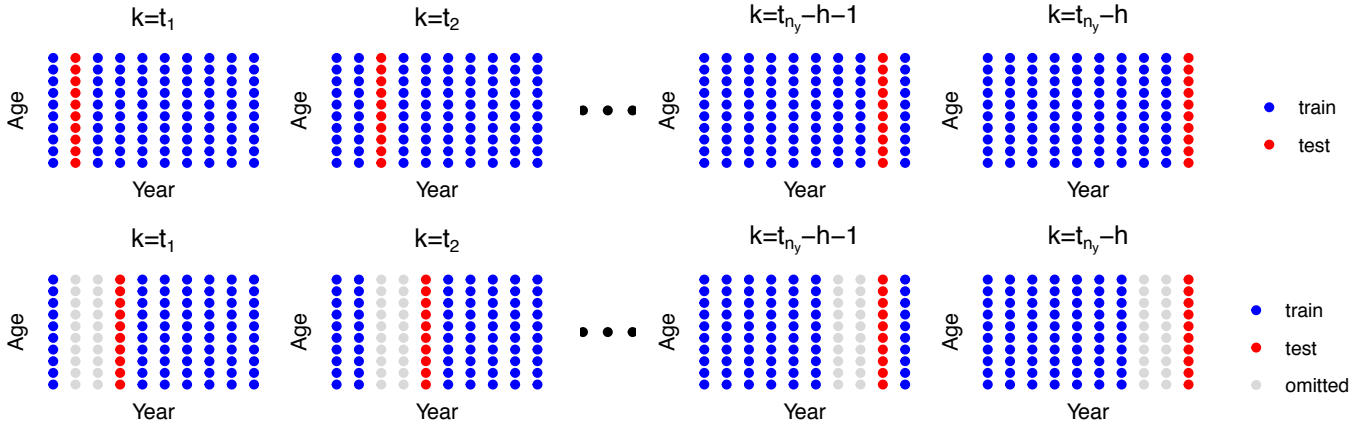

Figure 2: Iterations of cross validation for horizon one-year-ahead ($h = 1$) (top row) and three-years-ahead ($h = 3$) (bottom row).

The data is iteratively used to train and test the base learners for the time periods $t \in [t_1, t_2, \ldots, t_{n_y}]$. In particular, $\forall k = t_1, \ldots, t_{n_y} - h$, the training data shown in blue is defined as $\mathcal{D}^k_{\text{train}}(h) = \{(x, t)\}_{t \notin [k+1, k+h]}$ while the testing data shown in red is defined as $\mathcal{D}^k_{\text{test}}(h) = \{(x, t)\}_{t=k+h}$. The data in the left and right of the test data set (red) are used for training the base learners (Bergmeir, Costantini, and Benítez 2014).

As shown in Figure 2 (bottom row), using the training data, we can estimate the coefficients corresponding to every age group and cohort. However, we can not fully approximate the period index $\kappa^i(t)$ for non-training years with no observations. Figure 3 illustrates the time-series process of imputing the missing values. We can impute the missing values shown in grey by fitting a random walk with drift using all observable period indices shown in blue on both sides of the missing region as

$$\kappa^{(i)}(t) = \kappa^{(i)}(t-1) + \delta^{(i)} + \epsilon^{(i)}(t), \tag{8}$$

where $\epsilon^{(i)}(t) \sim \mathbb{N}(0, {\sigma_k}^{2^{(i)}})$ and $\delta^{(i)}$ is a drift parameter. Once we approximate the drift term, we can then predict the missing values using a forward fill approach (SriDaran et al. 2021)

$$\hat{\kappa}^{(i)}(t) = \hat{\kappa}^{(i)}(t-1) + \hat{\delta}^{(i)}.$$

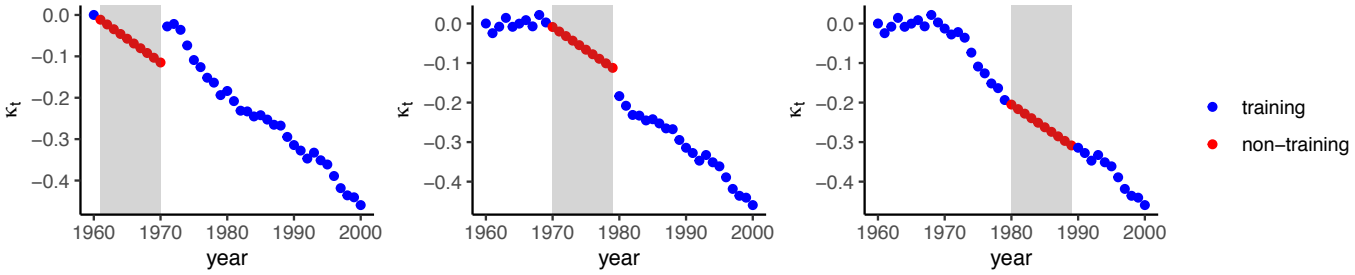

Figure 3: Illustration of time-series imputation approach. The blue points correspond to κ_t estimates that can be calculated directly from the training data and the grey area to the non-training years that can not be estimated. The red points indicating the values imputed using a forward fill procedure.

The cross-validation errors at horizon h are then estimated for the test data using

$$\theta^h(\widehat{\mu}(x,t), \mu(x,t)) = \frac{1}{n_a(t_{n_y} - t_1 - h + 1)} \sum_{k=t_1}^{t_{n_y}-h} \sum_{(x,t)\in\mathcal{D}^k_{\text{test}}(h)} \pi(\widehat{\mu}(x,t), \mu(x,t)). \tag{9}$$

We estimate the cross-validation mean squared error (CVMSE(h)) and cross-validation projection bias (CVPB(h)) by substituting their respective loss functions in Equation (9).

Finally, the best model at forecast horizon h is selected as the model with the least $\theta^h(\widehat{\mu}(x,t), \mu(x,t))$. However, cross-validation offers more information than merely identifying which model is the best. It permits the estimation of combinations of models, which can yield more precise out-of-sample forecasts. Therefore, instead of picking the model with the least $\theta^h(\widehat{\mu}(x,t), \mu(x,t))$, we use the cross-validated predictions to develop a stacked regression ensemble as discussed in Section 6.

5. Model Combination

Model selection and combination are two competing forecasting approaches. The model selection approach produces forecasts using the single best-selected model, assuming that the chosen single best model is close to the underlying data-generating process (Kourentzes, Barrow, and Petropoulos 2019). However, due to limited data and complexity of the data structures, it is often hard to identify the true data-generating process using real data (Makridakis, Spiliotis, and Assimakopoulos 2019). Even when a particular best-selected model appears to be close to a true data-generating process, using the single best-selected model is unstable and wastes information in the rejected models (Ahlburg 1995; Yao et al. 2018). Alternatively, combining forecasts from multiple models is a naturally reasonable substitute for using a single forecasting technique. It empirically combines predictions from composite models that reduce model choice uncertainty and improve point forecast accuracy (Bates and Granger 1969).

Let the h-year-ahead mortality rate forecasts from M mortality models $L_1, \ldots, L_M$ be $\widehat{\mu}_1(x, t_{n_y}+h), \ldots, \widehat{\mu}_M(x, t_{n_y}+h)$ for age $x \in [x_1, x_{n_a}]$ at time $t_{n_y}+h$. Then, in general then the combined mortality rate forecasts $\ln\left(\widehat{\mu}(x, t_{n_y}+h)\right)_{\text{comb}}$ are given as

$$\ln\left(\widehat{\mu}(x,t_{n_y}+h)\right)_{\text{comb}} = \sum_{m=1}^{M} w_m(h)\ln\widehat{\mu}_m(x,t_{n_y}+h), \tag{10}$$

where h is the forecasting horizon and $w_m(h)$ is the horizon specific weight assigned to a particular forecasting method such that $(w_1(h),\ldots,w_M(h)) \in \mathbb{R}^M$ and $\sum_{m=1}^{M} w_m(h) = 1$ which makes the model combination a weighted average.

In what follows, we consider four different model combination approaches: Simple Model Averaging (Genre et al. 2013), Bayesian Model Averaging (Bates and Granger 1969), Model Confidence Set (Hansen, Lunde, and Nason 2011), and Stacked Regression Ensemble (Wolpert 1992). These methods vary from each other depending on how they use the historical data to choose the combination weights $w_m(h)$ or the individual forecasting methods in Equation (10).

5.1. Simple Model Averaging

In a simple model averaging (SMA) approach, the mortality rate forecasts from multiple mortality models are assigned equal weights in Equation (10). That is, $w_m(h) = \frac{1}{M}$, with the final mortality rate forecasts given by

$$\ln\left(\widehat{\mu}(x,t_{n_y}+h)\right)_{\text{SMA}} = \sum_{m=1}^{M}\frac{1}{M}\ln\widehat{\mu}_m(x,t_{n_y}+h).$$

The SMA approach assigns weights to individual mortality models without considering their historical performances. Empirically, sometimes a simple model averaging outperforms sophisticated ways of assigning weights to the individual forecasting methods (Genre et al. 2013; Graefe et al. 2014; Makridakis, Spiliotis, and Assimakopoulos 2019). This is referred to as a "model combination puzzle" (Stock and Watson 2004). Therefore, we consider the SMA as a benchmark to other implemented model combinations in this paper.

5.2. Bayesian Model Averaging

The Bayesian model averaging (BMA) estimates the weights in Equation (10) using the posterior model probabilities. Let $L_1,\ldots,L_M$ be the set of M mortality models and suppose that at least one of these models is the true data-generating mortality model. Let a vector $\boldsymbol{\Phi} = \phi_m,\ldots,\phi_M$ be parameters corresponding to each of the mortality models, and ξ be the combined mortality forecasts. Then, the posterior distribution given the mortality data $\mathcal{D}$, is given by

$$\mathbb{P}(\xi|\mathcal{D}) = \sum_{m=1}^{M}\mathbb{P}(\xi|L_m,\mathcal{D})\mathbb{P}(L_m|\mathcal{D}) = \sum_{m=1}^{M} w_m\mathbb{P}(\xi|L_m,\mathcal{D}), \tag{11}$$

where $w_m = \frac{\mathbb{P}(\mathcal{D}|L_m)}{\sum_{m=1}^{M}\mathbb{P}(\mathcal{D}|L_m)\mathbb{P}(L_m)}$, $\mathbb{P}(\mathcal{D}|L_m) = \int \mathbb{P}(\mathcal{D}|\phi_m,L_m)\mathbb{P}(\phi_m|L_m)\,d\phi_m$, $\mathbb{P}(\phi_m|L_m)$ is the prior density of ϕ_m under L_m, $\mathbb{P}(\mathcal{D}|\phi_m,L_m)$ is the likelihood function, and $\mathbb{P}(L_m)$ is the prior probability that L_m is a true data-generating model. Therefore, the Bayesian prediction is the weighted average of individual mortality model predictions, with weights proportional to the posterior probability of each mortality model. The Bayesian model averaging assigns weight to each mortality model depending on its number of parameters and how well it fits the mortality data measured by Akaike or Bayesian Information Akaike Criterion. A full Bayesian model averaging in Equation (11) is computationally hard to compute, so traditionally, the weights are approximated using much simpler Akaike or Bayesian Information approximation (Bates and Granger 1969).

Define a set of M mortality models with the comparable Akaike Information Criterion (AIC) such that $\text{AIC}(M) = \{\text{AIC}_m\}$ for $m = 1,\ldots,M$. Given non-informative priors and equal model prior probabilities, the BMA approximates the normalized weights using the AIC criterion (Shang 2012) as

$$w_m^{\text{BMA}}(h) \approx \frac{e^{-0.5\text{AIC}_m}}{\sum_{m=1}^{M} e^{-0.5\text{AIC}_m}}, \quad \forall m = 1,2,\ldots,M, \tag{12}$$

where AIC_m is the raw Akaike Information Criterion of a particular model m given in Equation (4). However, the raw AIC are not interpretable as they contain arbitrary constants and are much affected by the sample size (Piłatowska 2009). Therefore, in our implementation we follow Shang and Booth (2020) and transform the raw AIC to the difference in AIC of each model with reference to AIC of the best model. That is, $\Delta_m = \text{AIC}_m - \min(\text{AIC}(M))$, so that the normalized Bayesian weights in Equation (12) becomes

$$w_m^{\text{BMA}}(h) \approx \frac{e^{-0.5\Delta_m}}{\sum_{m=1}^{M} e^{-0.5\Delta_m}}, \quad \forall m = 1, 2, \ldots, M. \tag{13}$$

The weights $w_m^{\text{BMA}}(h)$ approximates the probability that model m is the best model given the data and the set of candidate models, that is, $w_m^{\text{BMA}}(h) \approx \mathcal{L}(L_m|\mathcal{D})$. Empirically, the forecasting methods with $\Delta_m < 4$ are considered plausible in terms of Kullback-Leibler information (Piłatowska 2009).

When a particular mortality model is superior to other models when measured by Akaike or Bayesian Information Criteria, the posterior probability tends to be close to one (Wagenmakers and Farrell 2004). The combined results will be indistinguishable from those of the best-fitting individual mortality model. Therefore, for each mortality model to contribute information to the ensemble, we consider an alternative Bayesian model averaging approach used in Kontis et al. (2017), where they compute the Bayesian weights, $w_m^{\text{bias}}(h)$ using normalized exponentiated projection bias given by

$$w_m^{\text{bias}}(h) \approx \frac{e^{-0.5|\text{Projection Bias}_m|}}{\sum_{m=1}^{M} e^{-0.5|\text{Projection Bias}_m|}}, \quad \forall m = 1, 2, \ldots, M. \tag{14}$$

We calculate the projection bias in Equation (14) using Equation (6) when the loss function is $\ln \widehat{\mu}(x,t) - \ln \mu(x,t)$.

The weights in Equation (14) depend on a single validation set with the potential of overfitting. Moreover, Equation (14) assumes the same weights for all forecasting horizons. Thus, for our numerical experiments in Section 7, we also propose two modified versions of computing the model weights using a cross-validation approach which makes full use of the data and produces horizon-dependent weights. In a similar spirit to Kontis et al. (2017), we replace projection bias in Equation (14) with the cross-validation mean squared error for horizon h, (CVMSE(h)) and the cross-validation projection bias for horizon h, (CVPB(h)) in Equation (9). The weights based on the CVMSE and CVPB at horizon h are given as

$$w_m^{\text{CVMSE}}(h) \approx \frac{e^{-0.5\text{CVMSE}_m(h)}}{\sum_{m=1}^{M} e^{-0.5\text{CVMSE}_m(h)}} \quad \forall m = 1, 2, \ldots, M,$$
$$w_m^{\text{CVPB}}(h) \approx \frac{e^{-0.5\text{CVPB}_m(h)}}{\sum_{m=1}^{M} e^{-0.5\text{CVPB}_m(h)}}, \quad \forall m = 1, 2, \ldots, M.$$

5.3. Model Confidence Set

The model confidence set (MCS) approach chooses the subset of superior mortality models to combine in Equation (10) using equal weights. Superior models are identified by assessing if they have an equal predictive ability at a given confidence interval (Hansen, Lunde, and Nason 2011). An equal predictive ability test of the models can be conducted using any arbitrary loss functions, such as the square or absolute loss function explained in Section 4. Formally, suppose $\mathcal{M}^*$ is a subset of the set of original models denoted by $\mathcal{M}^0 = \{L_1, \ldots, L_M\}$. According to Hansen, Lunde, and Nason (2011), we define the square loss function for evaluating any model m at time t for ages $x \in [x_1, x_{n_a}]$ as $\varphi_{L_m,x,t} = \left(\ln \mu(x,t) - \ln \widehat{\mu}_m(x,t)\right)^2$. Therefore, the loss differential between any two models L_a and L_b for a finite time $t \in [t_1, \ldots, t_{n_y}]$ and ages $x \in [x_1, x_{n_a}]$ is defined as

$$\zeta_{ab,x,t} = \varphi_{L_a,x,t} - \varphi_{L_b,x,t}, \forall\ a, b = 1, 2, \ldots, M.$$

Let $\eta_{ab} = \mathbb{E}(\zeta_{ab,x,t})$ be finite and not time dependent (Hansen, Lunde, and Nason 2011). Now, the role of the model confidence set is to choose a set of superior models $\widehat{\mathcal{M}}_{1-\alpha} \equiv \{L_a \in \mathcal{M}^* : \eta_{ab} \leq 0\ \forall\ L_b \in\ \mathcal{M}^*\}$ at the test level α through a sequence of significance tests. The equal predictive ability test hypothesis is constructed as follows

$$\begin{aligned} H_{0,\mathcal{M}^*} &: \eta_{ab} = 0, \forall a, b = 1, 2, \ldots, M, \\ H_{1,\mathcal{M}^*} &: \eta_{ab} \neq 0 \text{ for some } a, b = 1, 2, \ldots, M. \end{aligned} \tag{15}$$

We use η_{ab} to formulate the hypothesis test as follows:

$$t_{ab} = \frac{\bar{\zeta}_{ab}}{\sqrt{\widehat{\text{Var}}(\bar{\zeta}_{ab})}},$$

where $\bar{\zeta}_{ab} = \frac{1}{n_a n_y^{'}} \sum_{t=t_1}^{t_{n_y}} \sum_{x=x_1}^{x_{n_a}} \zeta_{ab,x,t}$, here, $n_y^{'} = t_{n_y} - t_1 + 1$ and $\bar{\zeta}_{ab}$ measures the relative sample loss between L_a and L_b. The quantity $\widehat{\text{Var}}(\bar{\zeta}_{ab})$ is estimated using the bootstrap method (Hansen, Lunde, and Nason 2011). Model confidence set works sequentially by eliminating the worst model at each stage until the null hypothesis given in Equation (15) is accepted at a given level of confidence. The elimination rule for the worst model is coherent with the test statistic and is defined as $\arg\max_{L_a \in \mathcal{M}^*} \{\sup_{L_b \in \mathcal{M}^*} t_{ab}\}$.

We calculate the combination weight of each model m in the subset $\widehat{\mathcal{M}}_{1-\alpha}$ using

$$w_m(h) = \frac{\mathbb{I}(m \in \widehat{\mathcal{M}}_{1-\alpha})}{|\widehat{\mathcal{M}}_{1-\alpha}|},$$

where $|\widehat{\mathcal{M}}_{1-\alpha}|$ is the number of selected superior models and the indicator function $\mathbb{I}(\cdot)$ has the value of one if the model m is included in $\widehat{\mathcal{M}}_{1-\alpha}$ and zero otherwise.

Shang and Haberman (2018) use a single validation set approach to select the same set of superior mortality models for all the forecasting horizons. We modify this procedure by choosing the superior mortality models via cross-validation. We calculate the loss differential at each forecasting horizon h as $\zeta_{ab,t}^h = \varphi_{L_a,x,t}^h - \varphi_{L_b,x,t}^h, \forall\, a, b = 1, 2, \ldots, M$ and we select the superior models at each forecasting horizon. We compute $\varphi_{L_m,x,t}^h$ using a quadratic loss, $(\ln \mu(x, t+h) - \ln \widehat{\mu}_m(x, t+h))^2$ at horizon h.

6. Stacked Regression Ensemble

The stacked regression ensemble combines point forecasts from multiple base learners using weights that optimize a cross-validation criterion (Wolpert 1992). The base learners can, for example, be a family of generalized age-period-cohort discrete-time mortality models. The stacked regression ensemble often proceeds in two steps in generating the final predictions. The first step consists of multiple base models which separately generate cross-validated predictions from the training data. The predictions from various models and the observed response variable constitute the metadata. Secondly, a meta-learner is trained on the metadata to learn the optimal weights for combining multiple base learners while minimizing the cross-validation criterion. Figure 4 schematizes the implementation of the stacked regression ensemble framework when forecasting three-year ahead mortality rates.

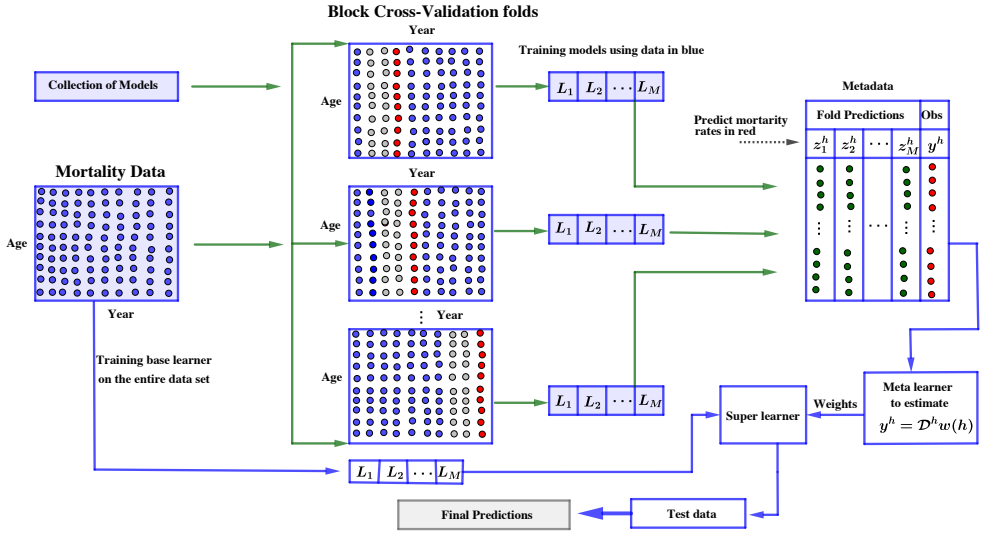

Figure 4: Stacked regression ensemble framework when forecasting three-year ahead mortality rates. The framework can be generalized for predicting mortality rates in any forecast horizon by varying the width of the testing data in red.

Like the standard model combination approaches, the stacked regression ensemble combines the mortality rate forecasts as a linear combination of individual mortality model forecasts. However, in the stacked regression ensemble, the combination weights are viewed as coefficients of a linear regression problem in which the observed mortality rates are treated as the dependent variable and the point forecasts from the individual mortality models are treated

as the independent variables. That is

$$\underbrace{\ln \mu(x, t_{n_y} + h)}_{\text{Dependent variables}} = \sum_{m=1}^{M} \underbrace{w_m(h)}_{\text{coefficients}} \underbrace{\ln \widehat{\mu}_m(x, t_{n_y} + h)}_{\text{covariates}}, \tag{16}$$

where $\ln \mu(x, t + h)$ are observed mortality rates and $\ln \widehat{\mu}_m(x, t + h)$ are the mortality rate forecasts from individual models generated via cross-validation. For each forecast horizon h, the set of combining weights $w_m(h)$ $\forall m = 1, \ldots, M$ are learned using any supervised machine learning algorithm that optimizes the respective loss function (Gunes, Wolfinger, and Tan 2017). The squared loss function is usually preferred (Wolpert 1992). The optimization is constrained such that these weights sum to one.

6.1. Implementation of Stacked Regression Ensemble

We combine the individual mortality rate forecasts linearly using the stacked regression ensemble via block cross-validation as follows:

1. Split the mortality data into the in-sample data to train the model and validation set for model performance testing. We use two-thirds of the data to train the model and the remaining data for model testing.

2. Fit each base learner $L_1, \ldots, L_M$ on the training data $\mathcal{D}^k_{\text{train}} = \{(x, t)\}_{t \notin [k+1, k+h]}, \forall k = t_1, \ldots, t_{n_y} - h$ depending on the forecasting horizon and ages $x \in [x_1, x_{n_a}]$. For instance, we train each mortality base learners on the training data set shown in blue in Figure 2 to forecast one-year-ahead ($h = 1$) and three-years-ahead ($h = 3$) mortality rates, respectively.

3. For each mortality base learner $L_1, \ldots, L_M$, predict the mortality rates $\ln \widehat{\mu}(x, t + h)$ using the test data set $\mathcal{D}^k_{\text{test}} = \{(x, t)\}_{t=k+h}, \forall k = t_1, \ldots, t_{n_y} - h$ for different forecasting horizons and ages $x \in [x_1, x_{n_a}]$. For instance, we predict the mortality rates shown in red in Figure 2 for $h = 1$ and $h = 3$, respectively. Let $h = 1, \ldots, H$ be the forecasting horizon for estimating the weights $\mathbf{w}(h)$. Also, let the predictions of a base learner m, $m = 1, \ldots, M$ for the forecasting horizon h, $h = 1, \ldots, H$ on the entire test data be an $n_a \times n^*_y(h)$ dimensional matrix Z^h_m with $n^*_y(h) = n_y - h$. Let $z^h_m = \mathbf{vec}(Z^h_m)$ be the vectors of predictions from a base learner m on the test set. The operator **vec** stacks the columns of Z^h_m in column order on top of each other. Therefore, for any base learner m, $z^h_m = [\ln \widehat{\mu}_m(x_1, t_1 + h), \ldots, \ln \widehat{\mu}_m(x_{n_a}, t_1 + h), \ldots, \ln \widehat{\mu}_m(x_1, t_{n_y}), \ldots, \ln \widehat{\mu}_m(x_{n_a}, t_{n_y})]'$. For each forecasting horizon h, the matrix $\mathcal{D}^h$ of M columns (base learners) and $n_a \times n^*_y(h)$ rows is given by

$$\mathcal{D}^h = [z^h_1, \ldots, z^h_M].$$

   The matrix $\mathcal{D}^h$ along with the observed mortality rates $\ln \mu(x, t)$ forms the metadata. With these metadata we formulate the following linear regression

$$y^h = \mathcal{D}^h \mathbf{w}(h), \tag{17}$$

   where $\mathbf{w}(h) = [w_1(h), \ldots, w_M(h)]'$ and y^h is the corresponding observed mortality rates with $n_a \times n^*_y(h)$ rows given by $y^h = [\ln \mu(x_1, t_1 + h), \ldots, \ln \mu(x_{n_a}, t_1 + h), \ldots, \ln \mu(x_1, t_{n_y}), \ldots, \ln \mu(x_{n_a}, t_{n_y})]'$. The parameter, $w(h)$, of the linear regression Equation (17) can be estimated using any supervised machine learning algorithm or meta-learner.

4. Estimate the optimal weights $\mathbf{w}(h)$ in Equation (17) using a meta-learner. The optimal weights are estimated by regressing the dependent variable y^h on to the cross-validation predictions of base models $\mathcal{D}^h$. Intuitively, the weighting coefficients indicate the forecasting strength of each base learner at a particular horizon.

5. Predict the unseen data by fitting the base learners $L_1, \ldots, L_M$ on the in-sample data, and use them to predict mortality rates on the validation set. Finally, combine these predictions using Equation (10).

6.2. Base Learners Integration

Any supervised machine learning algorithm can be used as a meta-learner for optimally estimating the weights in Equation (17) to combine multiple mortality models (Gunes, Wolfinger, and Tan 2017). Standard least squared regression is the simplest meta-learner we could first consider (Khairalla et al. 2018). Thus, using the metadata $\mathcal{D}^h = [z^h_1, \ldots, z^h_M]$, the weighting coefficients $\widehat{w_m}^*(h)$ in Equation (10) are determined as the minimizers of

$$\widehat{\mathbf{w}}^*(h) = \underset{\mathbf{w}(h)}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i^h - \sum_{m=1}^{M} w_m(h) z_{im}^h \right)^2, \tag{18}$$

where $N = n_a \times n_y^*(h)$. Standard least squared regression tends to perform poorly because of the variability of weighting coefficients (James et al. 2014).

Breiman (2004) applied non-negative least squared regression to estimate the weighting coefficients by optimizing Equation (18) such that it restricts all the weighting coefficients to be positive $\widehat{w_m}^*(h) > 0, \ \forall m = 1, \ldots, M$. The non-negativity condition ensures that the predictive accuracy of the stacked regression ensemble in Equation (10) is better than selecting the single best individual mortality model (Breiman 2004). Also, the positive weights increase the ensemble interpretability (Gunes, Wolfinger, and Tan 2017). Linear and non-negative least squared regressions are preferred as meta-learners when the base learners are less correlated because these methods cannot penalize highly correlated base learners (James et al. 2014).

Typically, regularization methods such as lasso, ridge, and elastic net regressions are used to reduce overfitting and increase the predictive accuracy of the stacked regression ensemble (Breiman 2004). In lasso regression, we estimate the weights as

$$\widehat{\mathbf{w}}^*(h) = \underset{\mathbf{w}(h)}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i^h - \sum_{m=1}^{M} w_m(h) z_{im}^h \right)^2 + \lambda \sum_{m=1}^{M} |w_m(h)|, \tag{19}$$

where $\widehat{\mathbf{w}}^*(h) \in \mathbb{R}^M$ is a vector of optimal weighting coefficients (James et al. 2014). Equation (19) has a tuning parameter λ that controls the degree of shrinkage applied to the weights of the resulting stacked regression ensemble. If the penalty parameter λ is chosen correctly, the total generalization error will decrease, and the resulting ensemble will be more stable (Gunes, Wolfinger, and Tan 2017). This penalty parameter is optimally determined via cross-validation (Gunes, Wolfinger, and Tan 2017). Lasso regression can force some weighting coefficients to be identically zero if the constraint λ is tight enough, which results in a simple and interpretable ensemble. Therefore, lasso regression can be used as a pruning tool in which only some mortality models will contribute to the final mortality rate forecasts. In the presence of groups of correlated base learners, lasso regression indifferently selects only one base learner from each group (Ahrens, Hansen, and Schaffer 2019).

Ridge regression can be used when all the base learners are relevant in the ensemble presented in Equation (10). Ridge shrinks all of the weighting coefficients towards zero via shrinkage parameter λ, however, none of them will be set exactly to zero (James et al. 2014). Ridge regression approximates the weighting coefficients by minimizing the quantity

$$\widehat{\mathbf{w}}^*(h) = \underset{\mathbf{w}(h)}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i^h - \sum_{m=1}^{M} w_m(h) z_{im}^h \right)^2 + \lambda \sum_{m=1}^{M} w_m^2(h).$$

Ridge regression tends to give small and well-distributed weights in Equation (10).

The elastic net regression can be used to keep or drop the correlated base learners jointly (Ahrens, Hansen, and Schaffer 2019). When there are many accurate individual mortality models to combine, their mortality rate forecasts will be correlated, and the elastic-net linear regression will be able to perform groupwise selection. The elastic net regression combines the ℓ_1 and ℓ_2 properties of lasso and ridge regressions, respectively into

$$\widehat{\mathbf{w}}^*(h) = \underset{\mathbf{w}(h)}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i^h - \sum_{m=1}^{M} w_m(h) z_{im}^h \right)^2 + \lambda_1 \sum_{m=1}^{M} w_m^2(h) + \lambda_2 \sum_{m=1}^{M} |w_m(h)|.$$

The elastic net produces a sparse ensemble depending on the choices of λ_1 and λ_2 (James et al. 2014).

7. Empirical Results

For our empirical results, we use mortality data from the Human Mortality Database provided by the University of California Berkeley and Max Planck Institute for Demographic Research (2020). The years of data vary by country, but for consistency, we consider the period of 1960 to 2015. There are 22 countries with reliable and quality data for this period, for both males and females, which provides 44 empirical datasets for testing the performance of stacked regression ensembles. To illustrate the application of stacked regression ensemble, we first consider in Subsection 7.1

the commonly referenced mortality data of England and Wales for both males and females. Then, in Subsection 7.2, we present the results of applying the stacked regression ensemble to 44 populations.

In all cases, we use mortality data from 1960 to 1990 to train six individual mortality models via cross-validation. Additionally, we consider the restricted age range 50 to 89, the range of greatest interest to pensions and annuities providers (Currie 2016). The fitted individual mortality models are used to predict one to 15-year-ahead mortality rates. That is because pension funds and life insurers are principally interested in accurate mortality rate forecasts in longer forecasting horizons. However, it is challenging to achieve accurate mortality rate forecasts in longer forecasting horizons because of the limited historical mortality data, and mortality models lose their forecasting strength as the horizon lengthens. The mortality rate forecasts from these individual mortality models are combined using different model combination methods.

7.1. Application of Stacked Regression Ensemble to England and Wales

7.1.1. Mortality Model Selection Dilemma

Table 2 presents the performance of different GAPC mortality models using AIC, BIC, and CVMSE for England and Wales males. Both AIC and CVMSE concurrently choose RH as the best mortality model while BIC chooses M7 for males. The CVMSE chooses APC as the second-best model while AIC and BIC select PLAT and RH, respectively. For the third-best model, CVMSE chooses different models depending on the forecasting horizons while AIC and BIC select M7 and PLAT, respectively. Based on all the selection criteria, LC and CBD are the worst-performing mortality models because of their non-inclusion of cohort effects dominant in England and Wales mortality data (Villegas, Kaishev, and Millossovich 2018). These results align with the findings in Cairns et al. (2009) and Atance, Debón, and Navarro (2020) that different model selection criteria can lead to different mortality model choices.

Table 2: Values of AIC, BIC, and CVMSE (with their respective rankings shown in parentheses) for the different GAPC mortality models fitted to the England and Wales males population for ages $50-89$ and the period $1960-1990$. The values of CVMSE at different forecasting horizons namely one, five, ten, and 15 year-ahead are presented.

Criterion	LC	RH	APC	CBD	M7	PLAT
AIC	19076.90 (5)	15133.12 (1)	16005.34 (4)	37906.72 (6)	15216.00 (3)	15189.67 (2)
BIC	19635.30 (5)	16039.86 (2)	16712.30 (4)	38224.34 (6)	16035.66 (1)	16045.19 (3)
CVMSE ($h=1$)	0.001750 (5)	0.001150 (1)	0.001300 (2)	0.006100 (6)	0.001310 (4)	0.001300 (3)
CVMSE ($h=5$)	0.003120 (5)	0.001270 (1)	0.001780 (2)	0.007410 (6)	0.002350 (3)	0.002400 (4)
CVMSE ($h=10$)	0.005670 (5)	0.001190 (1)	0.002490 (2)	0.009600 (6)	0.004110 (3)	0.004120 (4)
CVMSE ($h=15$)	0.008320 (5)	0.002050 (1)	0.003680 (2)	0.012500 (6)	0.005950 (4)	0.005580 (3)
Δ_m(AIC)	3943.78	0	872.22	22773.6	82.88	56.55

Note: Values of CVMSEs for all the forecasting horizons are presented in the top panel of Figure 5.
AIC Difference: [1] $\Delta_m(\text{AIC}) = \text{AIC}_m - \min(\text{AIC}_m)$.

The cross-validation criterion allows the influence of the forecasting horizon in selecting the best mortality model. Conversely, both AIC and BIC are not customized to a specific forecasting horizon. Hence, they are less effective in choosing mortality models for longer forecasting horizons. Table 2 further shows that the values of different model selection criteria differ minimally from each other for some mortality models. Therefore, it is hard to know how much statistical importance we should attach to a difference in the values of the selection criteria between the single best model and the next best model (Wagenmakers and Farrell 2004). For instance, RH and M7 have BIC values of 16039.86 and 16035.66, respectively for males. Traditionally, we would choose M7 over RH based on a small difference of 4.2 on their BIC values, but it is questionable to select a single model given such small differences among the models (Wagenmakers and Farrell 2004). Moreover, we compare only a few mortality models in Table 2 which does not guarantee that the single best-selected mortality model will represent the underlying true mortality data-generating process.

Different model selection criteria can choose different single best mortality models with different out-of-sample performances. For example, in Table 2, BIC chooses M7 as the best mortality model for England and Wales males. However, as depicted in Figure 5 (Lower left panel), this model consistently outperforms other mortality models only on one-to-five-ahead mortality rate forecasts and performs poorly in the medium and long horizons[1]. Furthermore,

[1]In this study, the short-term horizon corresponds to a period of one-to-five years, medium-term horizon corresponds to a period of

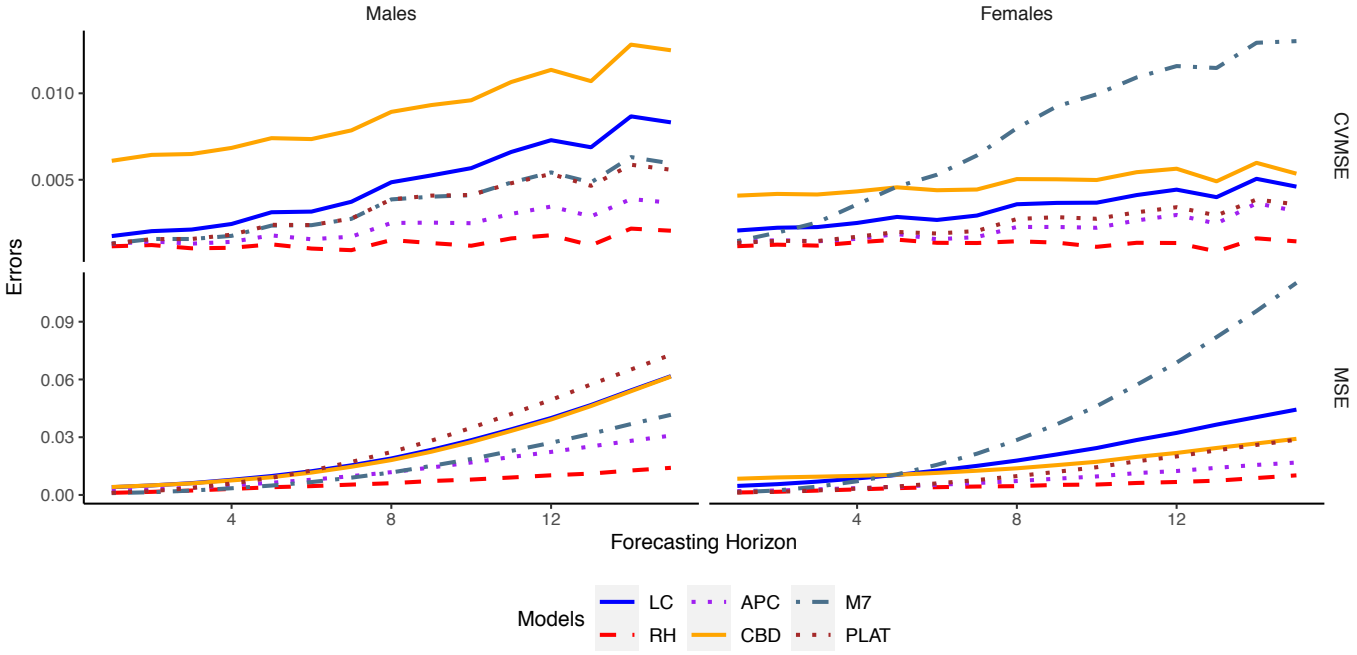

Figure 5: Top panel shows the CVMSEs of the one-step-ahead to 15-step-ahead mortality rate forecasts using different mortality methods and forecast horizons for England and Wales male mortality data (Top left panel) and females (Top right panel). The CVMSEs are estimated using mortality data from 1960 to 1990 for ages 50 to 89 via cross-validation. The lower panel shows the MSEs of one-step-ahead to 15-step-ahead mortality rate forecasts using different mortality methods and forecast horizons for England and Wales male mortality data (Lower left panel) and females (Lower right panel). The MSEs are estimated using mortality data from 1991 to 2015 for ages 50 to 89.

the RH model, which is the best mortality model by both the AIC and CVMSE criteria, is outperformed by the M7 model at least in short-term male mortality rate forecasts. This presents subjectivity to the mortality forecasters on which model selection criteria to employ.

The forecasting model selection paradigm assumes that the chosen single best mortality model is close to the underlying true mortality model. However, it is difficult to identify one mortality forecasting method that performs consistently well over all the forecasting horizons and for both males and females, making the choice of the single best forecasting mortality model a challenging task. Prudent mortality forecasters should acknowledge model risk and take advantage of combining various mortality models with different forecasting abilities instead of selecting the single best forecasting mortality model (Cairns et al. 2011).

7.1.2. Mortality Model Combination Using Stacked Regression Ensemble

As shown in the lower panel of Figure 5, multiple mortality forecasting techniques generate forecasts with varying prediction accuracies. Forecasts from a particular mortality model, for example, RH provides additional information, especially in longer horizons than forecasts from other forecasting techniques such as M7. This reveals that multiple mortality models satisfactorily learn different parts of the mortality data at different forecasting horizons. Their combinations complement each other, leveraging their advantages and avoiding their drawbacks.

7.1.2.1. Comparisons of the Point Forecast Accuracies

Before combining multiple GAPC mortality models, we test whether their mortality rate forecasts at different forecasting horizons are statistically different from each other. The distribution of MSEs for different mortality models at different forecasting horizons is not normally distributed. That allows us to employ the Friedman test, a non-parametric test that does not depend on the distributional assumptions of the data (see Appendix A) (Friedman 1937). The $p-$values are 1.518×10^{-11} and 2.137×10^{-12} for England and Wales males and females, respectively. The $p-$values are both less than a customary 5% level of significance. Thus, we reject the null hypothesis that there is no difference in the forecasting accuracy among the compared individual mortality models. Shang (2015) also reports

six-to-10 years, and long-term horizon as a period of 11-to-15 years.

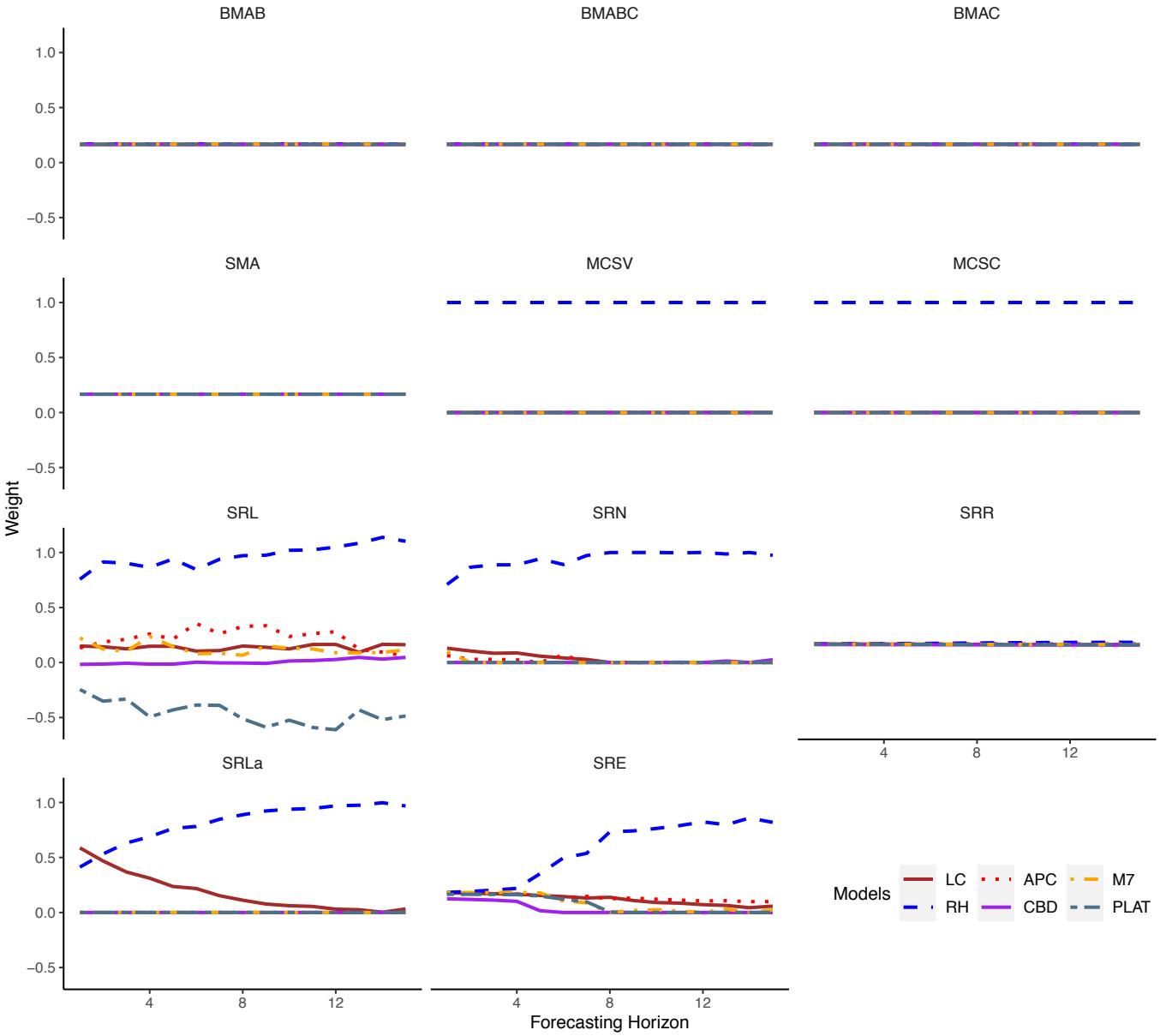

Figure 6: Component models weights for males: Horizon-specific optimal combining weights learned using different model combinations for England and Wales males mortality data from 1960 to 1990 and ages 50 to 89.

that the forecasting accuracy of different individual mortality models is statistically different from each other. This indicates that the GAPC models produce different mortality forecasts for both males and females, and hence they can be combined using multiple model combination methods.

7.1.2.2. Combination Weights for Mortality Models

We estimate the weights of the individual mortality models using multiple model combination approaches described in Section 5 and 6. Namely, we apply Simple Model Averaging (SMA), Bayesian Model Averaging using a projection bias estimated from a single withheld data (BMAB), Bayesian Model Averaging using a projection bias estimated from a cross-validation approach (BMABC), and Bayesian Model Averaging using cross-validation mean squared error (BMAC). We also use a model confidence set based on single validation and cross-validation approaches, which we refer to as MCSV and MCSC, respectively. Finally, we apply stacked regression ensemble using lasso regression (SRLa), ridge regression (SRR), elastic net regression (SRE), non-negative least square regression (SRN), and linear regression (SRL) as meta-learners. We did not implement the Bayesian Model Averaging which uses modified Akaike Information Criteria in Equation (13) because as shown in Table 2, most of the individual models have $\Delta_m(\text{AIC}) > 4$

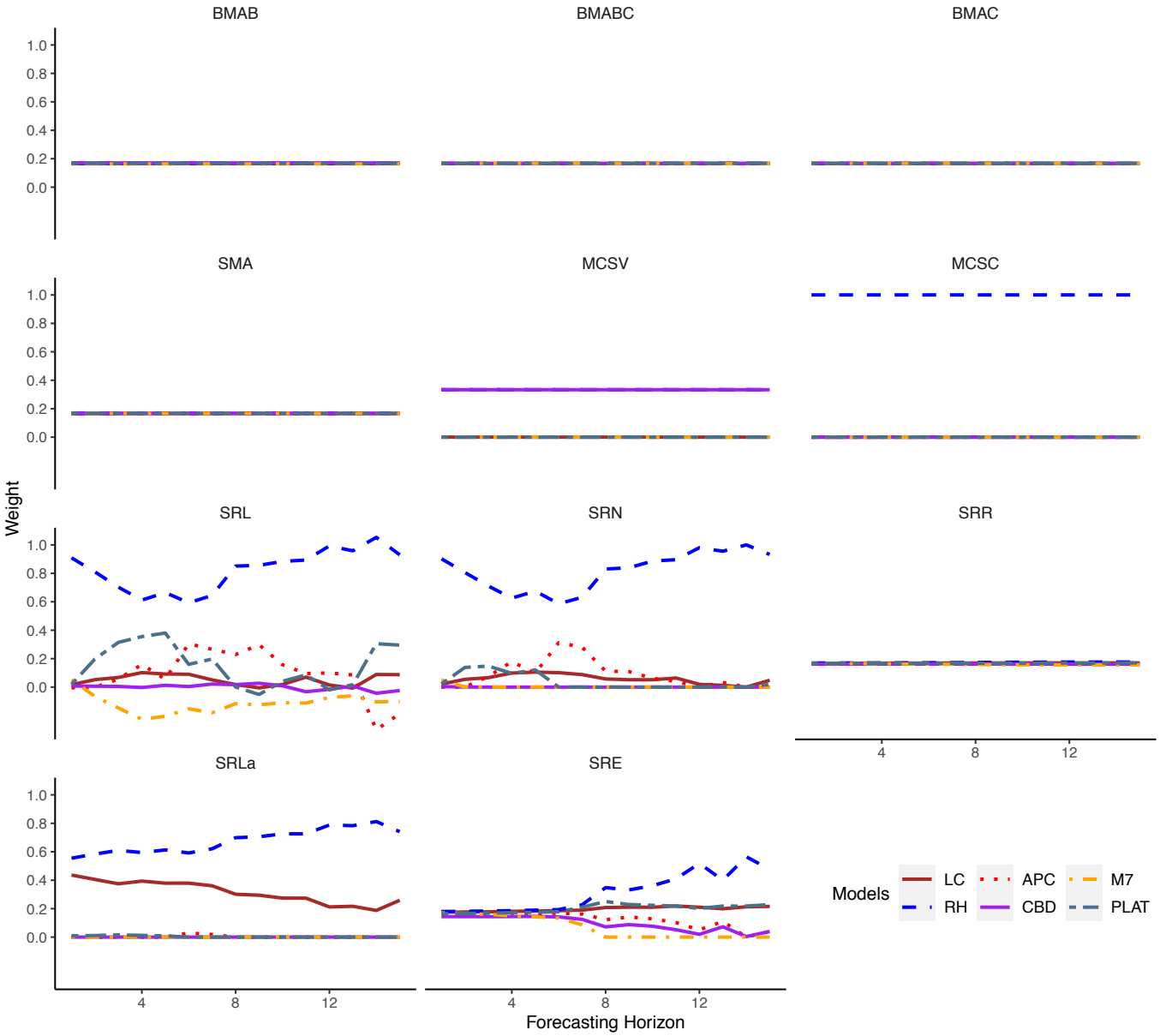

Figure 7: Component models weights for females: Horizon-specific optimal combining weights learned using different model combinations for England and Wales females mortality data from 1960 to 1990 and ages 50 to 89.

(Piłatowska 2009). This implies that all models are assigned zero weights except for the mortality model with the smallest AIC and hence the combination will include only one individual mortality model (see Barigou et al. (2021)). In all cases, we normalize the weights of the six mortality models to sum to one.

Figures 6 and 7 depict the corresponding weights for males and females, respectively. The weights estimated using BMAB, BMABC, BMAC, MCSC, and MCSV vary less among the individual mortality models and over the forecasting horizons. The BMAB, BMABC, and BMAC assign weights to individual mortality models which are close to the weighs from SMA. Shang and Booth (2020) report similar findings that the combination weights estimated using BMA approaches do not vary much over the forecasting horizons. This is because the weights estimated using these approaches do not reflect the ability of the model to generalize on the new unknown future mortality data. Additionally, all BMA approaches do not perform model selection and assign small and well-distributed weights to all the six mortality models. For the model confidence set approaches, all the selected superior models are combined using equal weights. At a 90% confidence level, MCSC chooses {RH} for females and males over all the forecasting horizons, respectively while MCSV chooses {RH} and {RH, APC, CBD} for females and males, respectively.

The stacked regression ensemble learns horizon-specific optimal weights for combining individual mortality models at

different forecasting horizons using different meta-learners. The mortality models with stronger predictive strength for a given horizon receive higher weights. This can be seen from the top panel of Figure 5, which indicates that models with the lower CVMSE get higher proportional weights. For instance, RH generalizes well for the new unknown future mortality data and hence gets higher weights that increase with the forecasting horizons. Other models such as LC, APC, CBD, M7, and PLAT receive corresponding smaller weights that decrease with the forecasting horizons across all the meta-learners.

The size of the weight assigned to a particular model using the stacked regression ensemble also depends on how the model differs from other mortality models. Both lasso and elastic net linear regressions give zero weights to some of the models with similar predictive ability. The less significant mortality models are more shrunk and hence contribute less to the combined mortality rate forecasts. For instance, from Figure 5 (Top left panel), the set of mortality models $\{\text{APC}, \text{RH}\}$, $\{\text{M7}, \text{PLAT}\}$, and $\{\text{LC}, \text{CBD}\}$ generate closely related CVMSEs, respectively for males. In such situations where groups of individual mortality models have similar predictive ability, lasso indifferently selects only one mortality model from each group and the models with the complex structures[2] such as M7 and PLAT are highly penalized due to their poor out-of-sample performance. Therefore, as shown in Figure 6, LC and RH are selected from each group, and complex models, M7 and PLAT are zero-weighted as a result of penalization. In contrast, in the same situation, elastic net regression does a group-wise selection of the individual mortality model. For instance, Figure 6 shows that the mortality models from each group get a smaller fraction of weights.

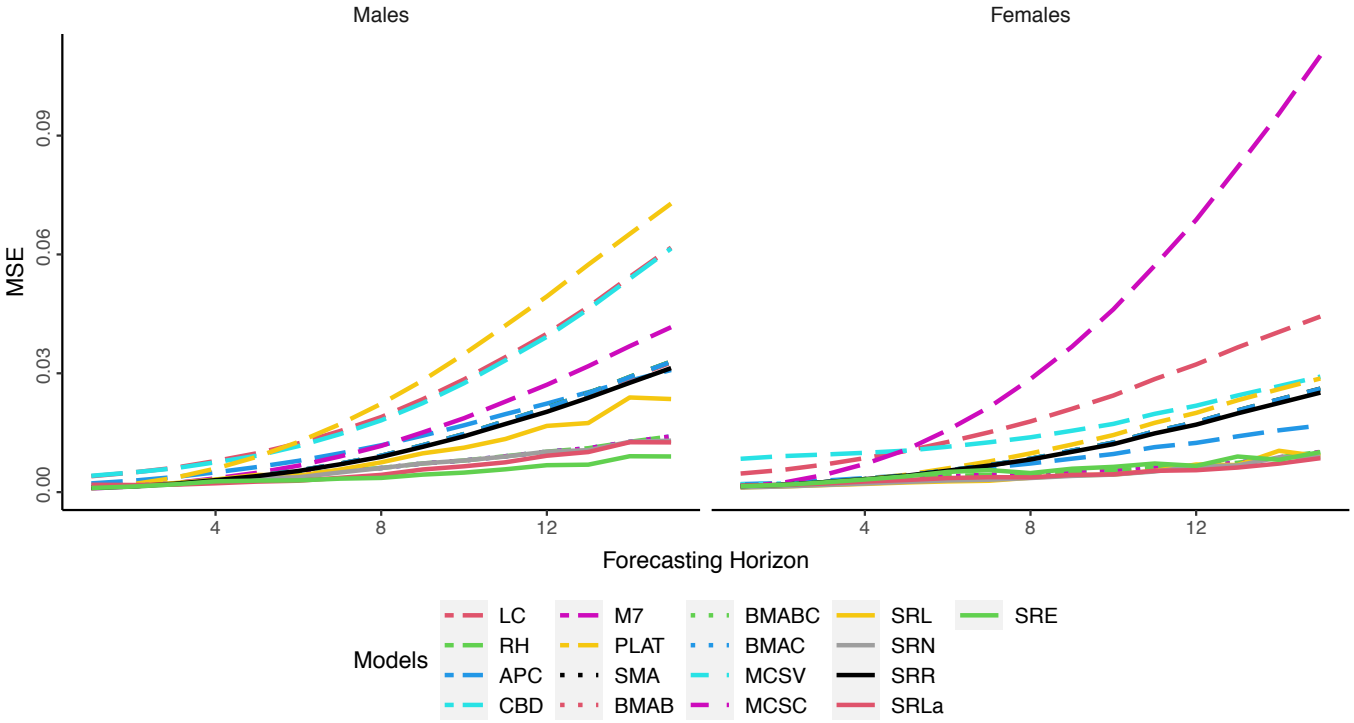

Figure 8: Out-of-sample performance of different models: MSEs of the one-step-ahead to 15-step-ahead mortality rate forecasts using different mortality methods and forecast horizons for England and Wales male mortality data (Left panel) and females (Right panel). The MSEs are estimated using the mortality data from 1991 to 2015 for ages 50 to 89. The values of MSEs for the mortality rate forecasts of males and females generally increase linearly with increasing forecast horizon.

Linear and non-negative least squares regressions estimate weights which are close to each other except that linear regression allows negative weights to some of the poorly performing individual mortality models such as M7 for females as shown in Figure 5 (Top right panel). SRL assigns negative weights to PLAT because the information in PLAT is better captured by M7. In contrast, non-negative least squares regression assigns zero weights to the poorly performing individual mortality models. Ridge regression assigns a small proportion of weights that are well distributed to all the six mortality models. The weights estimated from ridge regression are more consistent than linear regression over the forecasting horizons and among the models because ridge regression shrinks all the weights. The

[2]These are the mortality models with many numbers of parameters. M7 and PLAT have $3n_y + n_b$ and $n_a + 3n_y + n_b$ number of parameters, respectively.

Table 3: Percentage improvement over the best model combination for males: MSEs ($\times 10^3$) of male mortality rates for one-year-ahead to 15-year-ahead forecasts by model and forecast horizon. The SRE approaches produce gains in point forecast accuracy of 33% to 84% over the six individual mortality models for males. The underlined and bolded values correspond to the smallest MSE within the forecast horizon.

Horizon	LC	RH	APC	CBD	M7	PLAT	SMA	BMAB	BMABC	BMAC	MCSV	MCSC	SRL	SRN	SRR	SRLa	SRE
1	4.02	1.1	2.22	4.13	**0.95**	1.13	1.11	1.1	1.1	1.11	1.1	1.1	1.02	0.97	1.1	1.87	1.06
2	4.95	1.6	2.88	4.97	**1.41**	2.01	1.49	1.49	1.49	1.49	1.6	1.6	1.64	1.41	1.48	1.8	1.41
3	6.13	2.29	3.81	6	2.26	3.6	2.12	2.11	2.11	2.12	2.29	2.29	2.27	1.96	2.1	**1.89**	1.96
4	7.86	3.13	4.99	7.54	3.48	5.95	3.05	3.03	3.05	3.05	3.13	3.13	3.2	2.59	3.02	**2.24**	2.73
5	9.82	4.01	6.34	9.2	4.88	8.92	4.1	4.08	4.1	4.1	4.01	4.01	4.36	3.52	4.05	**2.66**	2.92
6	12.32	4.59	7.91	11.63	6.69	12.67	5.45	5.41	5.44	5.44	4.59	4.59	4.16	3.75	5.35	**2.89**	2.97
7	15.4	5.38	9.78	14.63	9	17.14	7.18	7.14	7.18	7.18	5.38	5.38	5.7	4.97	7.02	3.61	**3.41**
8	18.96	6.09	11.84	18.13	11.67	22.3	9.19	9.12	9.18	9.18	6.09	6.09	7.52	6.08	8.92	4.36	**3.6**
9	23.44	7.24	14.26	22.46	15.07	28.18	11.84	11.75	11.84	11.83	7.24	7.24	9.79	7.24	11.43	5.72	**4.43**
10	28.46	7.99	16.88	27.51	18.67	34.86	14.66	14.55	14.67	14.65	7.99	7.99	11.22	7.99	14.1	6.52	**4.94**
11	34.08	9.07	19.7	33.28	22.86	42.06	17.96	17.83	17.97	17.94	9.07	9.07	13.42	8.99	17.21	7.53	**5.77**
12	39.97	10.22	22.35	39.21	27.1	49.4	21.27	21.11	21.29	21.25	10.22	10.22	16.72	10.22	20.28	9.19	**6.8**
13	46.74	11.09	25.23	46.2	31.79	57.43	24.98	24.8	25.01	24.96	11.09	11.09	17.44	10.6	23.8	10.12	**6.93**
14	54.33	12.71	28.12	53.88	36.84	65.17	29.08	28.86	29.12	29.05	12.71	12.71	23.9	12.71	27.6	12.6	**9.06**
15	61.72	14.09	30.83	61.49	41.64	72.84	32.94	32.69	32.98	32.9	14.09	14.09	23.5	13	31.32	12.56	**9**
Mean	**24.55**	**6.71**	**13.81**	**24.02**	**15.62**	**28.24**	**12.43**	**12.34**	**12.44**	**12.42**	**6.71**	**6.71**	**9.72**	**6.4**	**11.92**	**5.71**	**4.47**
% Gain	**81.81**	**33.42**	**67.66**	**81.41**	**71.42**	**84.19**	**64.07**	**63.81**	**64.1**	**64.04**	**33.42**	**33.42**	**54.08**	**30.23**	**62.54**	**21.74**	**0**
Average Rank	**16.2**	**6.17**	**13.13**	**15.47**	**11.73**	**16.07**	**10.8**	**8.27**	**10.33**	**9.67**	**6.17**	**6.17**	**7.73**	**3.03**	**7.2**	**3.27**	**1.6**

Note:
% Gain $= \left(1 - \frac{\text{Mean of the Best combined Method}}{\text{Mean of the Typical Method}}\right) \times 100\%$.

complex mortality models such as M7 and PLAT receive the least and decreasing weights over forecasting horizons, reflecting the fact that these models tend not to generalize well in longer forecasting horizons. On the contrary, simple models such as LC, RH, and APC generalize well in longer horizons and hence receive a considerable proportion of the weights as the forecasting horizon lengthens.

7.1.2.3. Final Mortality Rate Forecasts

After estimating the optimal weights as illustrated in Figures 6 and 7, we can produce the final mortality rate forecasts using Equation (10). For instance, using SRE, the super-learner mortality model for forecasting one-year-ahead males mortality rate forecasts is given by combining the individual mortality models LC, RH, APC, CBD, M7, and PLAT using their corresponding weights $0.18, 0.18, 0.18, 0.14, 0.17$, and 0.18. Similarly, the 15-year-ahead males mortality rate forecasts is obtained by combining LC, RH, CBD, APC, and PLAT with their corresponding weights $0.22, 0.48, 0.04, 0.04$, and 0.23. Thus, the super-learner mortality models for forecasting one-year-ahead and 15-year-ahead males mortality rates are respectively given by

$$\ln\left(\widehat{\mu}(x, t_{n_y+1})\right)_{\text{SRE}} = 0.18 \times \widehat{\text{LC}} + 0.18 \times \widehat{\text{RH}} + 0.17 \times \widehat{\text{M7}} + 0.18 \times \widehat{\text{PLAT}} + 0.18 \times \widehat{\text{APC}} + 0.14 \times \widehat{\text{CBD}}, \tag{20}$$

$$\ln\left(\widehat{\mu}(x, t_{n_y+15})\right)_{\text{SRE}} = 0.22 \times \widehat{\text{LC}} + 0.48 \times \widehat{\text{RH}} + 0.23 \times \widehat{\text{PLAT}} + 0.04 \times \widehat{\text{APC}} + 0.04 \times \widehat{\text{CBD}}, \tag{21}$$

where $\widehat{\text{LC}}$, $\widehat{\text{RH}}$, $\widehat{\text{APC}}$, $\widehat{\text{CBD}}$, $\widehat{\text{M7}}$, and $\widehat{\text{PLAT}}$ represent the log-mortality rate forecasts from each of the individual models.

Noticeably, the weights of each of the individual mortality models that form the one-year-ahed and 15-year-ahead super-learner mortality model in Equations (20) and (21) differ. In one-year-ahead mortality rate forecasts, complex mortality models, namely M7 and PLAT, contribute 35% of the weights in the ensemble while the simple and more accurate models get the remaining weights. However, as the forecasting horizon increases to 15-year-ahead, complex mortality models do not generalize well to the new unknown future mortality data and lose their forecasting ability, and hence their contribution decreases to 23%.

7.1.2.4. Performance of Stacked Regression Ensemble

Figure 8 presents the performance of different methods measured in MSEs over different forecasting horizons. The stacked regression ensembles, namely SRN, SRLa, and SRE, consistently outperform individual mortality models in predicting the mortality rates over all the forecasting horizons using England and Wales mortality data. This improved forecast accuracy reflects the fact that the stacked regression ensemble optimally combines the features of

Table 4: Percentage improvement over the best model combination for females: MSEs (×1000) of female mortality rates for one-year-ahead to 15-year-ahead forecasts by model and forecast horizon. The SRN approaches produce gains in point forecast accuracy of 15% to 89% over the six individual mortality models for females. The underlined and bolded values correspond to the smallest MSE within the forecast horizon.

Horizon	LC	RH	APC	CBD	M7	PLAT	SMA	BMAB	BMABC	BMAC	MCSV	MCSC	SRL	SRN	SRR	SRLa	SRE
1	4.7	1.26	2.03	8.42	1.39	1.39	1.55	1.56	1.55	1.55	1.99	1.26	**1.21**	1.21	1.54	1.65	1.5
2	5.59	1.64	2.28	9.09	2.33	1.62	1.87	1.87	1.87	1.87	2.16	1.64	**1.42**	1.42	1.86	1.81	1.81
3	6.93	2.25	2.91	9.49	4.31	2.34	2.51	2.51	2.51	2.51	2.46	2.25	**1.78**	1.8	2.5	2.15	2.44
4	8.59	2.88	3.54	9.93	7.12	3.26	3.3	3.29	3.3	3.3	2.73	2.88	**2.08**	2.18	3.27	2.68	3.19
5	10.4	3.52	4.27	10.48	10.71	4.39	4.2	4.18	4.2	4.19	2.98	3.52	**2.44**	2.59	4.16	3.07	4
6	12.66	4.03	5.26	11.46	15.66	6.01	5.47	5.44	5.47	5.47	3.35	4.03	**2.71**	2.94	5.41	3.55	5.1
7	15.11	4.38	6.19	12.58	21.43	7.74	6.81	6.77	6.81	6.81	3.63	4.38	**2.88**	3.14	6.7	3.78	5.57
8	17.86	4.57	7.23	13.84	28.55	9.79	8.43	8.37	8.42	8.42	3.94	4.57	3.65	**3.57**	8.24	3.73	4.82
9	21.04	5.22	8.45	15.5	36.71	12.01	10.46	10.38	10.45	10.45	4.65	5.22	4.25	**4.09**	10.17	4.33	5.92
10	24.38	5.41	9.62	17.23	46.14	14.41	12.51	12.41	12.49	12.5	5.08	5.41	4.44	**4.43**	12.12	4.49	6.38
11	28.54	6.22	11.37	19.77	57.19	17.5	15.34	15.22	15.31	15.32	6.14	6.22	5.36	**5.22**	14.83	5.39	7.23
12	32.23	6.74	12.44	21.82	68.74	20.01	17.67	17.52	17.61	17.64	6.62	6.74	6.98	6.47	17.04	**5.54**	6.64
13	36.53	7.41	14.09	24.4	82	23.19	20.63	20.45	20.56	20.59	7.51	7.41	6.84	6.75	19.88	**6.21**	8.99
14	40.44	8.79	15.61	26.76	95.54	25.99	23.42	23.22	23.35	23.38	8.45	8.79	10.44	8.79	22.5	**7.21**	8.16
15	44.33	10.18	16.83	29.16	110.19	28.69	26.14	25.91	26.03	26.09	9.15	10.18	9.04	8.94	25.13	**8.57**	10.12
Mean	**20.62**	**4.97**	**8.14**	**15.99**	**39.2**	**11.89**	**10.69**	**10.61**	**10.66**	**10.67**	**4.72**	**4.97**	**4.37**	**4.24**	**10.36**	**4.28**	**5.46**
% Gain	**79.45**	**14.67**	**47.95**	**73.51**	**89.19**	**64.36**	**60.36**	**60.05**	**60.26**	**60.3**	**10.28**	**14.67**	**2.99**	**0**	**59.09**	**0.94**	**22.38**
Average Rank	**15.93**	**5.2**	**10**	**15.6**	**15.87**	**11.73**	**12.6**	**10.2**	**11**	**11.4**	**5.33**	**5.2**	**2.4**	**1.87**	**8.8**	**3.53**	**6.33**

Note:
% Gain $= \left(1 - \frac{\text{Mean of the Best combined Method}}{\text{Mean of the Typical Method}}\right) \times 100\%$.

different mortality models to form a super-learner mortality model which captures the features of the mortality data more adequately than any of the single mortality models.

Figure 8 also shows that the stacked regression ensembles, namely SRN, SRLa, and SRE, consistently outperform all other model combination methods over all the forecasting horizons using England and Wales mortality data. Notably, the competitive performance of the stacked regression ensemble is evident in medium and long-term forecasting horizons. The stacked regression ensemble assigns weights that reflect the out-of-sample performance of the mortality models and incorporate future mortality data uncertainty into the weights. The stacked regression ensemble also selects and combines individual mortality models simultaneously. This inherent feature in the stacked regression ensemble assigns higher weights to the individual GAPC mortality models, which generalize well to the new unknown future mortality data. In other model combinations, the model selection and model combination are treated as separate processes. Therefore, mortality models that do not generalize well on unseen data can be combined with good models, reducing the performance of a model combination. Finally, shrinkage techniques allow only accurate and diverse models to form the super-learner mortality model, improving forecasting accuracy.

Figure 8 further shows that SMA, BMABC, BMAB, BMAC, and SRR produce consistently close mortality rate forecasts for both genders, reflecting the fact that these methods assign approximately the same proportion of weights to all the individual mortality models over the forecasting horizons. All forms of Bayesian model combinations perform poorly compared to the stacked regression ensembles because they assign weights to the individual mortality models that reflect their goodness of fit rather than their ability to generalize the new unknown future mortality data. This poor performance of the Bayesian model averaging is consistent with previous research (Clarke 2004; Shang 2012; Shang and Booth 2020). Similar to Kontis et al. (2017), for all model combination methods, the improvement in forecasting performance becomes evident as the forecasting horizon increases. This is because the effects of mortality model misspecification are reduced through averaging compared to the case of the individual models, where the effects are accumulated over the forecasting horizons.

Tables 3 and 4 present the percentage improvement of out-of-sample mortality forecast accuracy based on one-year-ahead to 15-year-ahead MSEs. The percentage improvement measures the magnitude to which the MSEs of the best model combination is smaller than the MSEs of a typical mortality model across the forecasting horizons. The smallest mean values of MSEs are 0.00447 (SRE) and 0.00424 (SRN) for males and females, respectively. The SRE and SRN indicate gains in prediction accuracy over the six individual mortality models of 33% to 84% and 15% to 89% for males and females, respectively. The SRE and SRN achieved a percentage improvement of 33.42% and 14.67% over the RH model, which is the best mortality model selected based on both AIC and CVMSE criteria for males and females, respectively. By contrast, the model averaging methods based on the Bayesian approach perform poorly for both males and females for similar reasons given before. Both MCSV and MCSC achieved relative better performance than the simple model averaging for both males and females because they choose and combine only the

superior mortality models[3] using equal weights.

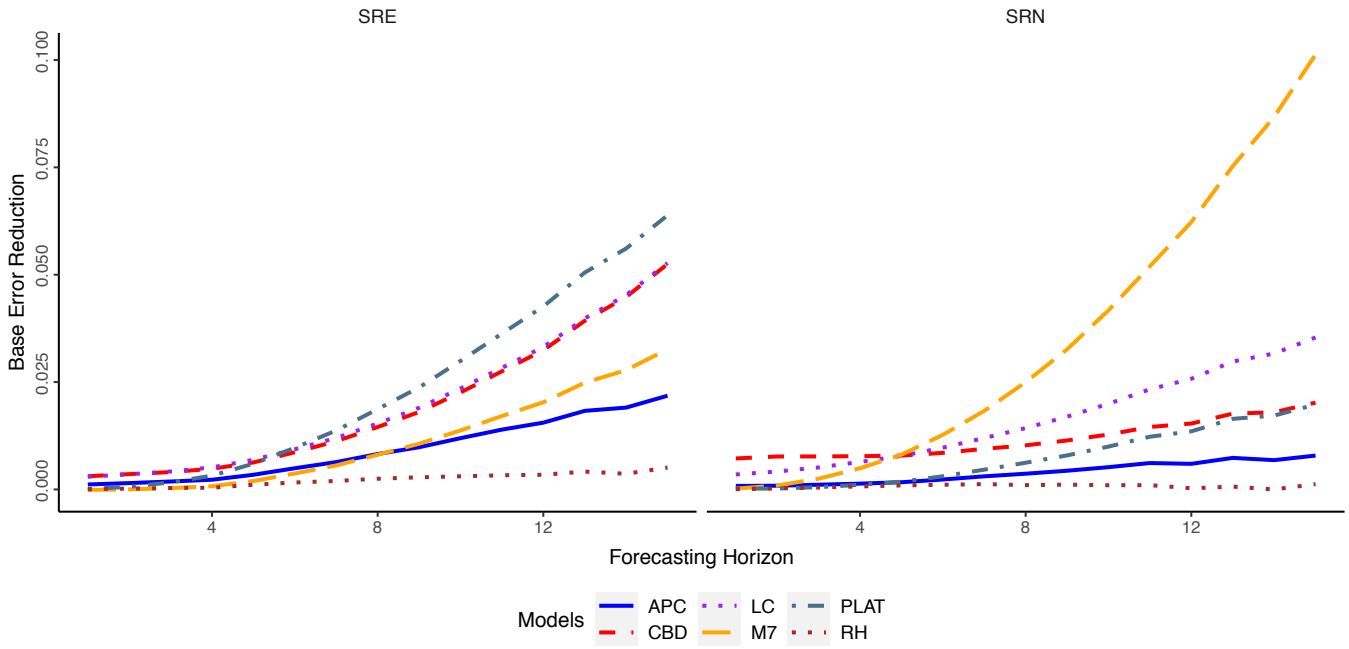

Figure 9: Base error reduction (BER) for the top-ranked model combination methods, namely SRE and SRN for males and females, respectively. The BER is computed based on different individual mortality models for England and Wales male and female mortality data.

The model combination approach achieve better predictive accuracy than the individual mortality models, in the process reducing model selection risk. The range of the MSEs of multiple model combination methods is 0.00797 and 0.00645 for males and females, respectively. By contrast, the corresponding range for the individual mortality model is 0.0215 and 0.0342 for males and females, respectively. This signifies that even when the model combination is not over all the best among the individual models, we can use it to forecast the mortality rates because it reduces the model selection risk. The strength of the stacked regression ensemble is further confirmed by being ranked at the top of both individual and combined mortality models in terms of the average ranking of the MSEs over the forecasting horizons. The top three mortality models in terms of the MSEs are all the stacked regression ensembles, namely SRE, SRN, and SRLa; and SRN, SRLa, SRL for males and females, respectively. The same models are also confirmed in terms of the average ranking of the models.

To further illustrate the forecasting strength of the stacked regression ensemble (SR), we plot in Figure 9 the base error reduction (BER) which is the difference between the MSEs of base learners and stacked regression ensemble given by

$$\text{BER} = \text{MSE}_{\text{Base Learner}} - \text{MSE}_{\text{SR}}.$$

The base error reduction gives a better indicator of the improvement in the stacked regression ensemble performance over the base learners (Aldave and Dussault 2014). A positive difference shows that performance accuracy is in favor of the stacked regression ensemble and vice versa. The bigger the BER, the better the stacked regression ensemble, and it represents a bigger gain in prediction accuracy of the stacked regression ensemble than individual mortality models. As depicted in Figure 9, both SRE and SRN have positive BER even for the single best mortality model selected by AIC, BIC, or CVMSE for males and females, respectively. Also, as shown in Figure 9, the forecasting accuracy gains of the stacked regression ensemble becomes more apparent as the forecasting horizon increases. For example, the predictive accuracy at a 15-year horizon is 13 and eight times worse than when predicting at a one-year horizon using RH and SRE, respectively for males. That is because the uncertainity[4] among the individual mortality models increases with the forecasting horizon. This implies that mortality rate forecasts from different mortality models diverge as the uncertainty rises. Additionally, the mortality rate features change over time, and in longer forecasting horizons, they can be better captured through combining multiple mortality models than using an individual mortality

[3]MCSC and MCSV select RH as the only superior model for males and hence it is not combined with other models. Therefore, MSEs of RH are similar to both MCSC and MCSV. For females, MCSV selects RH as the only superior model.

[4]Uncertainity is the difference between the highest and lowest mortality rate forecasts at any given forecasting horizon.

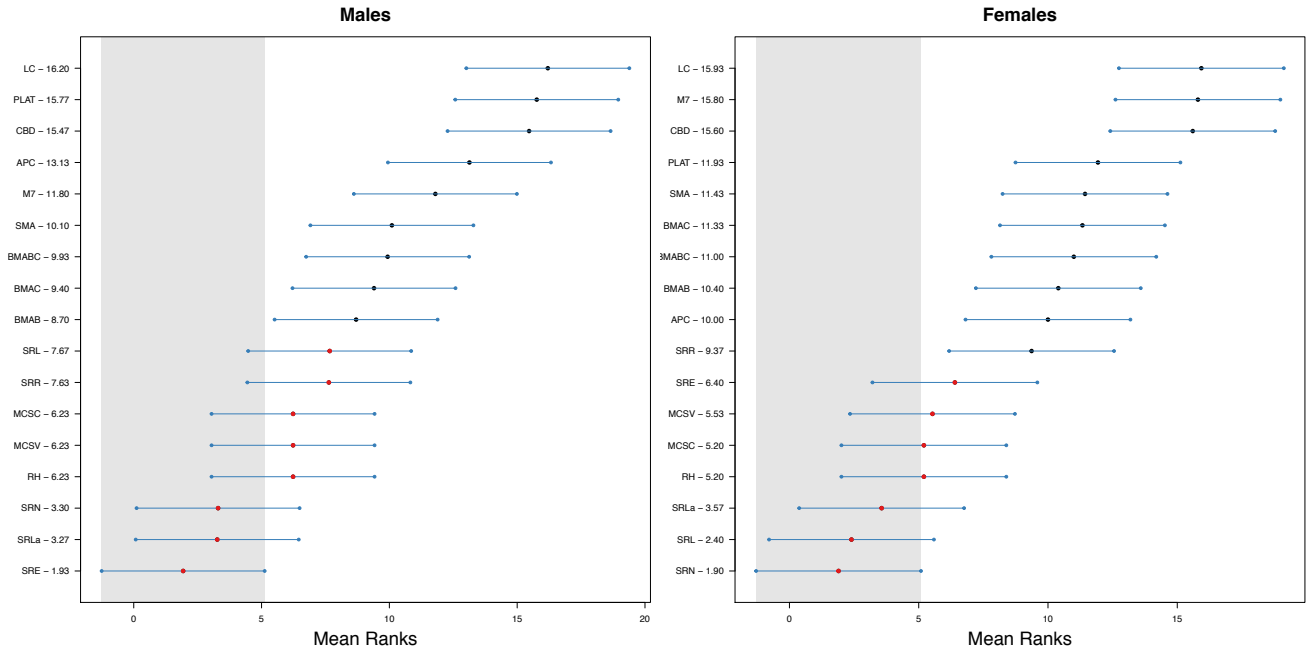

Figure 10: Multiple comparison with the best style plot for the mortality models considered for England and Wales males (Left panel) and females (Right panel). The average ranks are computed according to MSEs across the forecasting horizons at a customary 5% level of significance. Any mortality model with mean rank (plotted with •) outside the grey bounds indicating significant differences. The number along the $y-$ axis represents the average ranking of mortality model over the forecasting horizons. The smaller numbers represent models with better predictive accuracy.

model (Kontis et al. 2017). Therefore, this confirms that stacked regression ensemble should be preferred relative to individual mortality models in predicting mortality rates in longer forecasting horizons.

7.1.2.5. Statistical Test of Forecasting Accuracies

Finally, we perform a statistical test on the MSEs across the forecasting horizons to check whether the mortality rate forecasts from both the individual and combined mortality models are statistically different from each other. The Friedman test gives $p-$values of 1.889×10^{-32} and 1.652×10^{-34} for males and females, respectively. As these $p-$values are less than the customary 5%, we confidently reject the null hypothesis that the forecasting accuracy of models is the same. We can then employ the Nemeyi test to identify explicitly models which are statistically different from each other (see Appendix A). Figure 10 displays the average ranks of different mortality models. If the confidence interval[5] of two mortality models does not overlap, then their ranked forecasting performances are statistically different, and vice versa. A mortality model is statistically different from other models if its mean rank (plotted with •) is outside the grey bounds (Kourentzes and Petropoulos 2017). For males and females mortality rates forecasting, the stacked regression ensembles, namely SRE and SRN outperform all other models, respectively. However, their mortality rate forecasts do not differ significantly from SRN and SRLa; and SRL and SRLa for males and females, respectively. These results align with the results reported in Tables 3 and 4.

7.2. Application of Stacked Regression Ensemble to Other Countries

We now assess whether the stacked regression ensemble performs competitively relative to the existing individual and combined mortality methods using mortality data from other countries. We summarize the results for 44 populations from the Human Mortality Database with reliable mortality data from 1960 to 2015. To measure the forecast accuracy for all the countries, we follow Kontis et al. (2017) and we average the MSEs at each forecasting horizon across all the countries. Tables 5 and 6 show one-step-ahead to 15-step-ahead MSEs of mortality rate forecasts averaged at each forecasting horizon for all countries for males and females, respectively. Among the investigated mortality models,

[5]The confidence interval for each mortality model is $\bar{R}_m \pm \text{CD}$, where $\bar{R}_m$ is the mean rank of each model and CD is the critical difference (see Appendix A).

Table 5: Percentage improvement over the best model combination for males: MSEs ($\times 10^3$) of male mortality rates averaged across all the countries for one-year-ahead to 15-year-ahead forecasts by model and forecast horizon. The SRE approaches produce gains in point forecast accuracy of 13% to 49% over the six individual mortality models for males. The underlined and bolded values correspond to the smallest MSE within the forecast horizon.

Horizon	LC	RH	APC	CBD	M7	PLAT	SMA	BMAB	BMABC	BMAC	MCSV	MCSC	SRL	SRN	SRR	SRLa	SRE
1	4.82	2.98	4.16	6.46	2.78	2.92	2.74	2.74	2.74	2.74	3.32	2.82	2.82	**2.71**	2.74	3.23	2.73
2	6.1	4.1	5.28	7.44	4.24	4.06	3.66	3.66	3.66	3.66	4.31	3.81	3.8	**3.63**	3.66	4.07	3.64
3	7.78	5.62	6.89	8.75	6.53	5.89	4.99	4.99	4.99	4.99	5.76	5.39	5.34	4.96	4.99	5.22	**4.91**
4	9.9	7.68	8.92	10.51	9.71	8.36	6.8	6.79	6.8	6.8	7.67	7.37	7.53	6.76	6.78	6.99	**6.62**
5	12.31	10.02	11.23	12.53	13.54	11.29	8.86	8.84	8.86	8.86	9.82	9.63	9.81	8.7	8.83	8.82	**8.53**
6	15.14	12.59	13.97	14.98	18.16	14.83	11.32	11.29	11.32	11.32	12.36	11.83	12.14	10.9	11.26	10.78	**10.75**
7	18.46	15.64	17.16	17.99	23.71	19.02	14.28	14.24	14.28	14.28	15.38	14.92	15.65	13.74	14.19	13.48	**13.27**
8	22.37	19.15	20.93	21.58	30.23	23.94	17.84	17.78	17.84	17.84	18.99	18.33	19.43	16.85	17.69	16.89	**16.58**
9	26.79	22.75	25.03	25.62	37.4	29.33	21.74	21.67	21.74	21.74	23	21.4	23.35	19.89	21.52	20.1	**19.8**
10	31.64	26.39	29.41	30.16	45.5	35.16	26.03	25.95	26.04	26.02	27.41	24.84	26.61	23.13	25.72	23.69	**22.94**
11	37.06	30.57	34.45	35.39	54.4	41.95	30.91	30.8	30.92	30.9	32.3	28.71	30.6	26.96	30.5	27.98	**26.49**
12	42.54	34.65	39.3	40.63	63.24	48.46	35.7	35.57	35.71	35.68	37.02	32.38	34.65	30.12	35.17	31.01	**29.94**
13	49.37	39.35	45.48	47.24	73.83	56.31	41.73	41.58	41.75	41.72	42.99	37.07	39.2	35.89	41.08	35.39	**34.37**
14	56.53	45.18	52.19	54.19	84.98	64.53	48.24	48.06	48.26	48.22	49.1	43.97	46.29	41.23	47.46	42.04	**39.78**
15	64.04	51.32	59.1	61.45	96.51	72.81	54.99	54.78	55.01	54.97	55.3	49.01	49.06	46.4	54.1	46.44	**45.4**
Mean	**26.99**	**21.87**	**24.9**	**26.33**	**37.65**	**29.26**	**21.99**	**21.92**	**22**	**21.98**	**22.98**	**20.76**	**21.75**	**19.46**	**21.71**	**19.74**	**19.05**
% Gain	**29.42**	**12.88**	**23.49**	**27.64**	**49.4**	**34.89**	**13.37**	**13.08**	**13.4**	**13.35**	**17.1**	**8.26**	**12.42**	**2.1**	**12.26**	**3.51**	**0**
Average Rank	**15.33**	**9.4**	**13.47**	**15**	**15.8**	**14.6**	**8.27**	**5.87**	**8.27**	**7.07**	**11.73**	**6.8**	**8.8**	**2.07**	**4.73**	**4.67**	**1.13**

Note:
% Gain $= \left(1 - \frac{\text{Mean of the Best combined Method}}{\text{Mean of the Typical Method}}\right) \times 100\%$.

Table 6: Percentage improvement over the best model combination for females: MSEs ($\times 10^3$) of female mortality rates averaged over all the countries for one-year-ahead to 15-year-ahead forecasts by model and forecast horizon. The SRLa approaches produce gains in point forecast accuracy of 19% to 90% over the six individual mortality models for females. The underlined and bolded values correspond to the smallest MSE within the forecast horizon.

Horizon	LC	RH	APC	CBD	M7	PLAT	SMA	BMAB	BMABC	BMAC	MCSV	MCSC	SRL	SRN	SRR	SRLa	SRE
1	5.99	4.1	5.67	26.46	5.34	4.04	4.34	4.36	4.33	4.33	5.08	3.91	3.87	**3.83**	4.27	4.77	4.15
2	7.12	5.21	6.7	27.79	11.9	4.85	5.3	5.31	5.29	5.3	5.89	4.98	4.89	**4.81**	5.24	5.56	5.15
3	8.39	6.71	8.16	29.09	22.71	6.27	6.73	6.73	6.73	6.73	6.92	6.31	6.19	**5.94**	6.65	6.34	6.4
4	9.84	8.56	9.67	30.57	37.66	8	8.62	8.6	8.61	8.61	8.19	7.28	7.71	**7.27**	8.46	7.39	7.75
5	11.51	10.85	11.44	32.13	56.8	10.15	10.87	10.82	10.86	10.85	9.64	8.94	9.64	8.95	10.6	**8.75**	9.32
6	13.25	13.32	13.28	33.95	80.3	12.53	13.49	13.41	13.47	13.45	11.17	10.48	11.72	10.43	13.01	**10.04**	10.66
7	15.25	16.37	15.63	36.04	108.66	15.63	16.75	16.63	16.73	16.69	13.1	12.9	14.15	12.44	15.99	**11.94**	12.86
8	17.41	19.74	17.92	38.37	140.27	18.84	20.32	20.15	20.28	20.23	15.21	14.76	16.88	14.65	19.16	**14.11**	15.05
9	19.84	23.33	20.44	40.79	174.87	22.36	24.22	23.99	24.15	24.08	17.42	17.23	19.22	16.8	22.53	**16.3**	17.24
10	22.38	27.07	22.9	43.45	215.32	26.08	28.54	28.24	28.46	28.35	19.71	19.94	21.25	19.03	26.22	**18.58**	19.44
11	25.33	31.17	25.88	46.43	256.39	30.34	33.27	32.9	33.14	33.02	22.4	23.17	24.7	22.12	30.21	**21.42**	22.05
12	27.99	35.27	28.07	48.98	301.09	33.89	37.72	37.27	37.55	37.4	24.7	25.73	28.31	25.35	33.84	**24.33**	25.25
13	31.14	39.89	31.45	52.28	349.63	38.78	43.23	42.7	43	42.83	27.65	29.48	33.26	28.65	38.41	**27.43**	28.08
14	34.53	45.58	34.86	55.58	397.85	43.55	48.98	48.37	48.7	48.54	**30.99**	33.83	37.64	32.53	43.42	31.34	31.18
15	37.72	51.95	38.41	58.75	451.77	48.51	54.92	54.22	54.56	54.45	**34.39**	38.05	42.45	36.49	48.76	35.46	35.3
Mean	**17.41**	**19.74**	**17.92**	**38.37**	**140.27**	**18.84**	**20.32**	**20.15**	**20.28**	**20.23**	**15.21**	**14.76**	**16.88**	**14.65**	**19.16**	**14.11**	**15.05**
% Gain	**18.95**	**28.5**	**21.24**	**63.22**	**89.94**	**25.09**	**30.55**	**29.95**	**30.4**	**30.24**	**7.2**	**4.36**	**16.37**	**3.64**	**26.33**	**0**	**6.25**
Average Rank	**9.53**	**9.87**	**9.93**	**16.2**	**16.67**	**7.6**	**13.8**	**11.53**	**12.6**	**12**	**5.67**	**3.93**	**5.67**	**2.4**	**8.73**	**3.13**	**3.73**

Note:
% Gain $= \left(1 - \frac{\text{Mean of the Best combined Method}}{\text{Mean of the Typical Method}}\right) \times 100\%$.

SRE and SRLa achieve the smallest over all MSEs on average for males and females, respectively. Their respective smallest mean values of MSEs are 0.019 and 0.0141 for males and females, respectively.

Tables 5 and 6 further show the gain in prediction accuracy of individual mortality models with respect to the model combination with the smallest mean of MSEs across the forecasting horizon. The SRE and SRLa indicate the gains in prediction accuracy over the six individual mortality models of 13% to 49% and 19% to 90% for males and females, respectively. The model averaging methods based on the stacked regression ensemble perform competitively in both males and females. The top three performing mortality models in terms of average ranking are stacked regression ensembles, namely SRE, SRN, and SRLa; and SRN, SRLa, and SRE for males and females, respectively. Additionally, the model combination attain a lower model selection risk than the individual mortality models. The range of the MSEs of individual mortality methods is 0.0158 and 0.1229 for males and females, respectively, whereas the corresponding range for the model combination is 0.00393 and 0.00621 for males and females, respectively. Therefore, model combination methods improve the reliability of the mortality rate forecasts because by averaging, their range is bounded. That is, model combination sets a lower bound on point forecast accuracy (Shang and Booth 2020). It

is evident from Tables 5 and 6 that stacked regression ensembles are better capable of generating more accurate mortality rate forecasts and capturing model choice uncertainty than the existing individual mortality models and other model combination methods. This supports the significant potential of using a stacked regression ensemble in mortality modelling.

The bottom panel of Figure 11 presents heat maps showing the average ranks of various mortality models across different forecasting horizons, genders, and countries. The stacked regression methods, namely SRN, SRE, and SRLa are consistently ranked lower across the countries for both males and females. This confirms that stacked regression methods perform consistently well across the countries. The Bayesian model averaging methods perform poorly across the countries compared to both the model confidence set and stacked regression ensemble. Even when the model combination approaches are not top-ranked for any country, they never rank the least accurate. For example, the simple model averaging out-performs some well-known mortality models such as LC for males across most of the countries. This confirms the potential of choosing the model combination over the individual mortality models. The RH is one of the highly ranked among the individual mortality models but lacks consistent performance across all the countries. PLAT and M7 perform poorly across countries for both males and females, respectively. However, similar to the findings in Shang and Haberman (2018), PLAT outperforms all other models for the Japanese female population. Figure 11 also visually shows that LC performs consistently better in females than males in alignment with the findings in Atance, Debón, and Navarro (2020).

The top and middle panels of Figure 11 present heat maps visualizing the ranking of various mortality models at one-year-ahead and 15-year-ahead forecasting horizons for males and females across different countries. We observe that complex mortality models such as PLAT and M7 perform relatively better in the short-term than in the long-term for both males and females across the countries. This is because complex mortality models tend to overfit and fail to generalize well in the new unknown future data, resulting in low predictive power for longer forecasting horizons. Simple mortality models such as APC and LC tend to have better performance in longer horizons than in shorter horizons. RH performs relatively better in both short-term and long-term horizons than other individual mortality models. Similar to Kontis et al. (2017), model combinations perform consistently better in the short-term and long-term forecasting horizons than the individual mortality models. Generally, there is no individual or combined method that performs the best across all the countries. This shows that countries have different mortality characteristics, and hence the performance of a mortality model varies across the countries.

Figure 11 further shows that the SRE and SRN are the over all best models but SRE is more stable than SRN. Also, SRE performs better in longer horizon than SRN. The SRE is more stable than SRN because it concurrently stabilizes and slowly shrinks the weights assigned to different mortality models. The better performance of SRN than other meta-learners is consistent with prior studies which suggest non-negative least square regression as the best choice as the meta-learner in stacked regression ensembles (Breiman 2004). This is because non-negative least square regression imposes non-negative condition to the weights that guarantee the best interpolating model combination. Stacked regression approaches outperform other model combination approaches. This is because the individual mortality models are weighted differently depending on their ability to capture the mortality characteristics of a particular population over the forecasting horizons. This allows the mortality forecasters to develop a mortality model customized to a particular population to achieve reduced out-of-sample errors.

The MCSC is also a competitive model combination. MCSC achieve better performance than MCSV because MCSC is developed such that the superior individual mortality models are chosen via cross-validation rather than using the single validation set. Also, the outstanding performance of MCSC, SRN, SRE, and SRLa signify the potential of allowing the individual mortality models to change depending on the forecasting horizon. Therefore, these methods often combine few selected individual mortality models at a particular horizon. Again, SRR proves to assign more consistent weights to the individual mortality models than SMA and hence we suggest it as the best alternative to SMA. All the model combination approaches perform better than individual mortality models. The predictive accuracy of model combination is also less variable than the forecasts from the individual mortality models across forecast horizons, countries, and genders. The choice of which model combination to use will depend on ease of implementation, computational time available, and the degree of accuracy required.

The results in Tables 5 and 6 do not confirm whether the attained gain in the forecasting accuracy by the mortality models are statistically different. Thus, to validate that, we apply the Friedman test on the MSEs of different mortality models averaged at each forecasting horizon across all the countries. The $p-$values are 9.508×10^{-35} and 2.095×10^{-31} for males and females, respectively. As these $p-$values are less than customary 5%, we confidently reject the null hypothesis that the forecasting accuracy of the mortality models is the same. We proceed with the Nemenyi test to identify explicitly which models are statistically different from each other. Figure 12 visualizes the

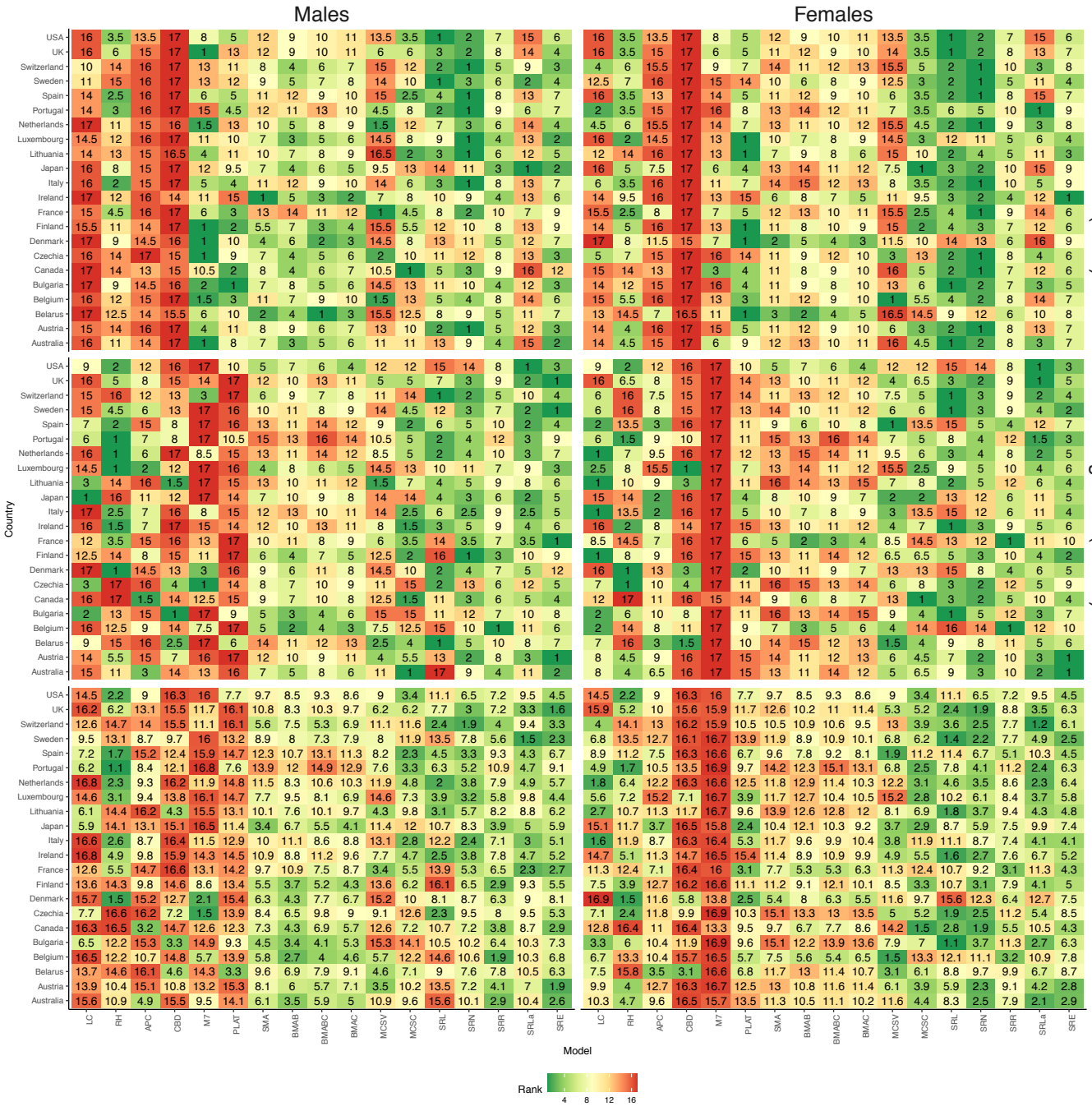

Figure 11: Heat maps showing the average ranks of different mortality models based on their MSEs over different forecasting horizons for males (Left panel) and females (Right panel) across different countries. The mean rank for each model and country is given, with the lowest representing the most accurate mortality model. The display ranges from green (best model) to red (worst model).

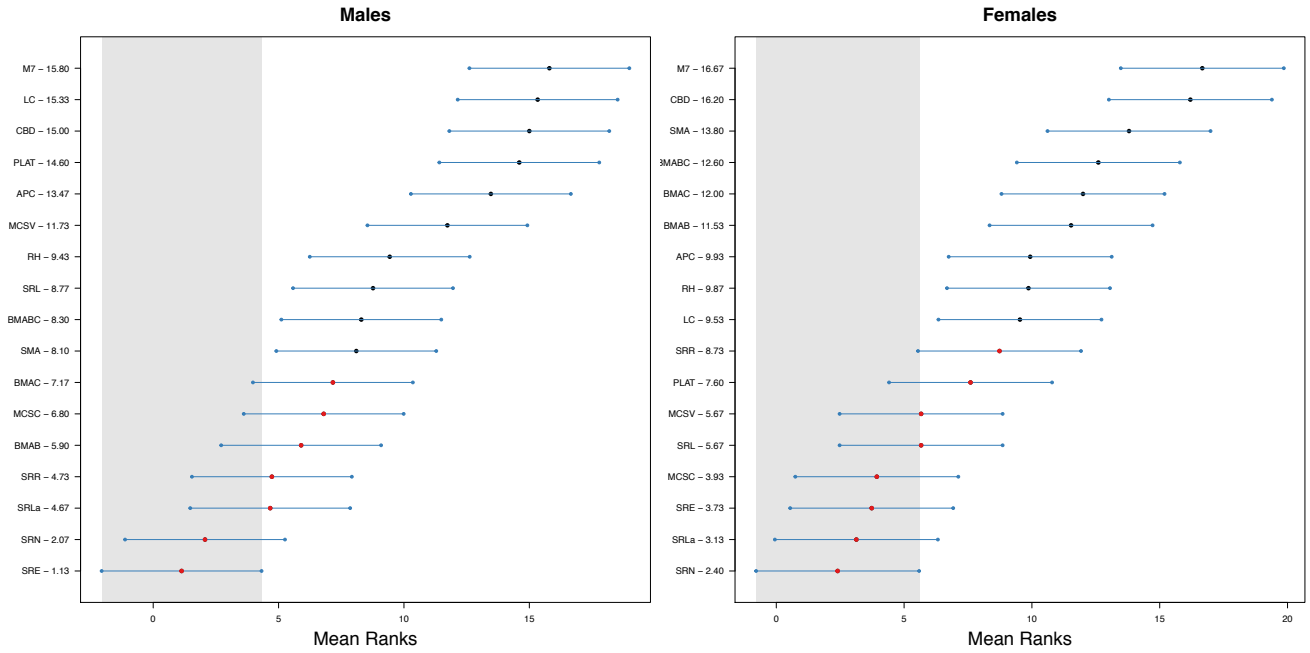

Figure 12: Multiple comparison with the best style plot for the mortality models considered for males (Left panel) and females (Right panel). The average ranks are computed according to MSEs across the forecasting horizons and countries at a customary 5% level of significance. Any mortality model with mean rank (plotted with •) outside the grey bounds indicating significant differences. The number along the $y-$ axis represents the average ranking of a mortality model over the forecasting horizons. The smaller numbers represent models with better predictive accuracy.

Nemanyi test results. SRE achieved better predictive performance than all the mortality models for males, but its predictive performance is not statistically different from SRN. For females, SRN perform significantly better than the rest of the forecasting mortality methods examined in terms of MSEs but its forecasting accuracy is not significantly different from SRE, SRLa, and MCSC. These results align with the results presented in Tables 5 and 6.

8. Conclusion and Future Research

In this paper, we have introduced and applied a stacked regression ensemble approach for combining mortality models. In our empirical application, we find that stacking six individual mortality models increases the predictive accuracy of mortality rate forecasts. The stacked regression ensemble methods, namely SRE and SRN outperform both individual and combined mortality methods for males and females, respectively. SRE and SRN are capable of optimally combining the features of multiple individual mortality models, which capture the mortality dynamics of particular mortality data better than individual mortality models. The success of these methods is attributable to their ability to simultaneously estimate the optimal combining weights that depend upon the length of the forecasting horizon and implicitly selecting the accurate and diverse individual mortality models to combine from the list of models considered. The choice of the lasso, non-negative least squares, and elastic regressions as the meta-learners automatically trim the mortality models which produce similar mortality rate forecasts or have the least forecasting accuracy at each forecasting horizon. Therefore, model selection is crucial in developing a competitive model combination. This is confirmed by the outstanding performance of lasso, non-negative least squares, elastic regressions, and the model confidence set variant based on cross-validation.

We also show that the optimal weights for combining the individual mortality models vary depending on the meta-learner, forecasting horizon, country, and gender. Mortality forecasters need to develop a model combination customized to the forecasting horizon and mortality data. This is contrary to the weights estimated using the Bayesian model averaging methods, simple model averaging, and model confidence set that vary less among the methods and the forecasting horizons. Also, estimating weights or choosing the individual mortality models via cross-validation proves to be a crucial step. Cross-validation allows the mortality forecasters to incorporate future mortality data uncertainty in the weights or model selection. Finally, the stacked regression methods like SRE and SRN are statistically different from other mortality methods across the countries and genders. However, both SRLa and MCSC are almost as good as SRE and SRN.

Future research could include more mortality models outside the family of the GAPC models. The mortality models to be included should capture different mortality data features such as linearity, non-linearity, trends, data size, and statistical considerations such as ensuring the smoothness of death rates, capturing outliers, and giving the recent mortality data more weights than the very past mortality data. The models should also capture unexpected events such as COVID-19 pandemic. This is expected to increase the forecasting accuracy of the model combinations, especially for higher mortality countries like in central and eastern Europe. However, even without considering mortality models other than GAPC, it is still prudent to combine the GAPC models as their combination reduces model selection risks and gives better and consistent mortality rate forecasts than selecting the single best mortality model. Finally, this study focuses on central projection; therefore, an important future improvement could be developing model combination approaches that can simultaneously generate central mortality projections and their corresponding probabilistic distributions.

Acknowledgments

This research is funded by the Australian Research Council Centre of Excellence in Population Ageing Research (CEPAR) project number CE110001029.

References

Ahlburg, Dennis A. 1995. "Simple Versus Complex Models: Evaluation, Accuracy, and Combining." *Mathematical Population Studies* 5 (3): 281–90. https://doi.org/10.1080/08898489509525406.

Ahrens, Achim, Christian B. Hansen, and Mark E Schaffer. 2019. "Iassopack: Model Selection and Prediction with Regularized Regression in Stata." IZA Discussion Papers 12081. Bonn: Institute of Labor Economics (IZA). http://hdl.handle.net/10419/193375.

Akaike. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19 (6): 716–23. https://doi.org/10.1109/TAC.1974.1100705.

Aldave, Roberto, and Jean-Pierre Dussault. 2014. "Systematic Ensemble Learning for Regression." *arXiv:1403.7267 [Stat]*, March. http://arxiv.org/abs/1403.7267.

Atance, David, Ana Debón, and Eliseo Navarro. 2020. "A Comparison of Forecasting Mortality Models Using Resampling Methods." *Mathematics* 8 (9): 1550. https://doi.org/10.3390/math8091550.

Barigou, Karim, Pierre-Olivier Goffard, Stéphane Loisel, and Yahia Salhi. 2021. "Bayesian Model Averaging for Mortality Forecasting Using Leave-Future-Out Validation." *Https://Hal.archives-Ouvertes.fr/Hal-03175212/Document*, 31.

Bates, J. M., and C. W. J. Granger. 1969. "The Combination of Forecasts." *Journal of the Operational Research Society* 20 (4): 451–68. https://doi.org/10.1057/jors.1969.103.

Berdin, Elia, and Helmut Gründl. 2015. "The Effects of a Low Interest Rate Environment on Life Insurers." *The Geneva Papers on Risk and Insurance - Issues and Practice* 40 (3): 385–415. https://doi.org/10.1057/gpp.2014.38.

Bergmeir, Christoph, Mauro Costantini, and José M. Benítez. 2014. "On the Usefulness of Cross-Validation for Directional Forecast Evaluation." *Computational Statistics & Data Analysis* 76 (August): 132–43. https://doi.org/10.1016/j.csda.2014.02.001.

Blake, David P., Andrew J. G. Cairns, Kevin Dowd, and Amy R. Kessler. 2018. "Still Living with Mortality: The Longevity Risk Transfer Market After One Decade." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3271283.

Booth, H., and L. Tickle. 2008. "Mortality Modelling and Forecasting: A Review of Methods." *Annals of Actuarial Science* 3 (1-2): 3–43. https://doi.org/10.1017/S1748499500000440.

Breiman, Leo. 2004. "Stacked Regressions." *Machine Learning* 24 (1): 49–64. https://doi.org/10.1007/bf00117832.

Brouhns, Natacha, Michel Denuit, and Jeroen K. Vermunt. 2002. "A Poisson Log-Bilinear Regression Approach to the Construction of Projected Lifetables." *Insurance: Mathematics and Economics* 31 (3): 373–93. https://doi.org/10.1016/S0167-6687(02)00185-3.

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 & Insurance* 73 (4): 687–718. https://EconPapers.repec.org/RePEc:bla:jrinsu:v:73:y:2006:i:4:p:687-718.

Cairns, Andrew J. G., David Blake, Kevin Dowd, Guy D. Coughlan, David Epstein, and Marwa Khalaf-Allah. 2011. "Mortality Density Forecasts: An Analysis of Six Stochastic Mortality Models." *Insurance: Mathematics and Economics* 48 (3): 355–67. https://doi.org/https://doi.org/10.1016/j.insmatheco.2010.12.005.

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): 1–35. https://doi.org/10.1080/10920277.2009.10597538.

Clarke, Bertrand. 2004. "Comparing Bayes Model Averaging and Stacking When Model Approximation Error Cannot Be Ignored." *Journal of Machine Learning Research* 4 (4): 683–712. https://doi.org/10.1162/153244304773936090.

Currie, Iain D. 2016. "On Fitting Generalized Linear and Non-Linear Models of Mortality." *Scandinavian Actuarial Journal* 2016 (4): 356–83. https://doi.org/10.1080/03461238.2014.928230.

Demšar, Janez. 2006. "Statistical Comparisons of Classifiers over Multiple Data Sets." *Journal of Machine Learning Research* 7 (1): 1–30. http://jmlr.org/papers/v7/demsar06a.html.

Doumpos, Michael, and Constantin Zopounidis. 2007. "Model Combination for Credit Risk Assessment: A Stacked Generalization Approach." *Annals of Operations Research* 151 (1): 289–306. https://doi.org/10.1007/s10479-006-0120-x.

Dowd, Kevin, Andrew J. G. Cairns, David P. Blake, Guy Coughlan, David Epstein, and Marwa Khalaf-Allah. 2008. "Backtesting Stochastic Mortality Models: An Ex-Post Evaluation of Multi-Period Ahead-Density Forecasts." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1396201.

Friedman, Milton. 1937. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance." *Journal of the American Statistical Association* 32 (200): 675–701. https://doi.org/10.1080/01621459.1937.10503522.

Genre, Véronique, Geoff Kenny, Aidan Meyler, and Allan Timmermann. 2013. "Combining Expert Forecasts: Can Anything Beat the Simple Average?" *International Journal of Forecasting* 29 (1): 108–21. https://doi.org/10.1016/j.ijforecast.2012.06.004.

Graefe, Andreas, J. Scott Armstrong, Randall J. Jones, and Alfred G. Cuzán. 2014. "Combining Forecasts: An Application to Elections." *International Journal of Forecasting* 30 (1): 43–54. https://doi.org/10.1016/j.ijforecast.2013.02.005.

Gunes, Funda, Russ Wolfinger, and Pei-Yi Tan. 2017. "Stacked Ensemble Models for Improved Prediction Accuracy." *Sas*, 1–19.

Hansen, Peter, Asger Lunde, and James Nason. 2011. "The Model Confidence Set." *Econometrica* 79 (2): 453–97. https://doi.org/10.3982/ECTA5771.

Hunt, Andrew, and Andrés M. Villegas. 2015. "Robustness and Convergence in the LeeCarter Model with Cohort Effects." *Insurance: Mathematics and Economics* 64 (September): 186–202. https://doi.org/10.1016/j.insmatheco.2015.05.004.

Hyndman, Rob, Heather Booth, Leonie Tickle, and John Maindonald. 2019. *Demography: Forecasting Mortality, Fertility, Migration and Population Data*. https://CRAN.R-project.org/package=demography.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, eds. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics 103. New York: Springer.

Janssen, Fanny. 2013. "DEMOGRAPHIC RESEARCH VOLUME 29 , ARTICLE 13 , PAGES 323-354 Impact of Different Mortality Forecasting Methods and Explicit Assumptions on Projected Future Life Expectancy : The Case of the Netherlands Lenny Stoeldraijer Coen van Duin Leo van Wissen Table of Contents" 29 (August): 323–54. https://doi.org/10.4054/DemRes.2013.29.13.

———. 2018. "Advances in Mortality Forecasting: Introduction." *Genus* 74 (1). https://doi.org/10.1186/s41118-018-0045-7.

Khairalla, Mergani A, Xu Ning, Nashat T AL-Jallad, and Musaab O El-Faroug. 2018. "Short-Term Forecasting for Energy Consumption Through Stacking Heterogeneous Ensemble Learning Model." *Energies* 11 (6). https://doi.org/10.3390/en11061605.

Kontis, Vasilis, James E. Bennett, Colin D. Mathers, Guangquan Li, Kyle Foreman, and Majid Ezzati. 2017. "Future Life Expectancy in 35 Industrialised Countries: Projections with a Bayesian Model Ensemble." *The Lancet* 389 (10076): 1323–35. https://doi.org/10.1016/S0140-6736(16)32381-9.

Kourentzes, Nikolaos, Devon Barrow, and Fotios Petropoulos. 2019. "Another Look at Forecast Selection and Combination: Evidence from Forecast Pooling." *International Journal of Production Economics* 209 (March): 226–35. https://doi.org/10.1016/j.ijpe.2018.05.019.

Kourentzes, Nikolaos, and Fotios Petropoulos. 2017. "Forecasting with R A Practical Workshop International Symposium on Forecasting 2017 Forecasting with R." International Symposium on Forecasting 2016.

Lee, Ronald D ., and Lawrence R Carter. 1992. "Modeling and Forecasting U . S . Mortality Author (S): Carter Published by : American." *Statistical Association Stable* 87 (419): 659–71.

Ma, Zhongchen, and Qun Dai. 2016. "Selected an Stacking ELMs for Time Series Prediction." *Neural Processing Letters* 44 (3): 831–56. https://doi.org/10.1007/s11063-016-9499-9.

Makridakis, Spyros, and Michèle Hibon. 2000. "The M3-Competition: Results, Conclusions and Implications." *International Journal of Forecasting* 16 (4): 451–76. https://doi.org/10.1016/S0169-2070(00)00057-1.

Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos. 2019. "The M4 Competition: 100,000 Time Series and 61 Forecasting Methods." *International Journal of Forecasting*, July, S0169207019301128. https://doi.org/10.1016/j.ijforecast.2019.04.014.

Oeppen, James, and James Vaupel. 2002. "Broken Limits to Life Expectancy." *Science, V.296, 1029-1031 (2002)* 296 (January).

Piłatowska, Mariola. 2009. "Combined Forecasts Using the Akaike Weights." *Dynamic Econometric Models* 9 (0): 5. https://doi.org/10.12775/DEM.2009.001.

Plat, Richard. 2009. "On Stochastic Mortality Modeling." *Insurance: Mathematics and Economics* 45 (3): 393–404. https://doi.org/10.1016/j.insmatheco.2009.08.006.

Puurula, Antti, Jesse Read, and Albert Bifet. 2014. "Kaggle LSHTC4 Winning Solution." *arXiv:1405.0546 [Cs]*, May. http://arxiv.org/abs/1405.0546.

Rabbi, Ahbab Mohammad Fazle, and Stefano Mazzuco. 2018. "Mortality and Life Expectancy Forecast for (Comparatively) High Mortality Countries." *Genus* 74 (1): 18. https://doi.org/10.1186/s41118-018-0042-x.

Ray, Evan L., and Nicholas G. Reich. 2018. "Prediction of Infectious Disease Epidemics via Weighted Density Ensembles." *PLOS Computational Biology* 14 (2): e1005910. https://doi.org/10.1371/journal.pcbi.1005910.

Reich, Nicholas G., Craig J. McGowan, Teresa K. Yamana, Abhinav Tushar, Evan L. Ray, Dave Osthus, Sasikiran Kandula, et al. 2019. "Accuracy of Real-Time Multi-Model Ensemble Forecasts for Seasonal Influenza in the U.S." Edited by Virginia E. Pitzer. *PLOS Computational Biology* 15 (11): e1007486. https://doi.org/10.1371/journal.pcbi.1007486.

Renshaw, A. E., and S. Haberman. 2006. "A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors." *Insurance: Mathematics and Economics* 38 (3): 556–70. https://doi.org/10.1016/j.insmatheco.2005.12.001.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *Ann. Statist.* 6 (2): 461–64. https://doi.org/10.1214/aos/1176344136.

Shang, Han Lin. 2012. "Point and Interval Forecasts of Age-Specific Life Expectancies: A Model Averaging Approach." *Demographic Research* 27: 593–644. https://doi.org/10.4054/DemRes.2012.27.21.

———. 2015. "Statistically Tested Comparisons of the Accuracy of Forecasting Methods for Age-Specific and Sex-Specific Mortality and Life Expectancy." *Population Studies* 69 (3): 317–35.

Shang, Han Lin, and Heather Booth. 2020. "Synergy in Fertility Forecasting: Improving Forecast Accuracy Through Model Averaging." *Genus* 76 (1): 27. https://doi.org/10.1186/s41118-020-00099-y.

Shang, Han Lin, and Steven Haberman. 2018. "Model Confidence Sets and Forecast Combination: An Application to Age-Specific Mortality." *Genus* 74 (1). https://doi.org/10.1186/s41118-018-0043-9.

Sharabiani, Mansour T A, and Alireza S Mahani. 2016. "Multi-Stage Heterogeneous Ensemble Meta-Learning with Hands-Off User-Interface and Stand-Alone Prediction Using Principal Components Regression: The R Package EnsemblePCReg," 32.

Shaw, Chris. 2007. "Fifty Years of United Kingdom National Population Projections: How Accurate Have They Been?" *Population Trends*, 16.

Sill, Joseph, Gabor Takacs, Lester Mackey, and David Lin. 2009. "Feature-Weighted Linear Stacking." *arXiv:0911.0460 [Cs]*, November. http://arxiv.org/abs/0911.0460.

SriDaran, Dilan, Michael Sherris, Andrés Villegas, and Jonathan Ziveyi. 2021. "A Group Regularisation Approach for Constructing Generalised Age-Period-Cohort Mortality Projection Models." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3790991.

Stock, James H., and Mark W. Watson. 2004. "Combination Forecasts of Output Growth in a Seven-Country Data Set." *Journal of Forecasting* 23 (6): 405–30. https://doi.org/10.1002/for.928.

Tashman, Leonard J. 2000. "Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review." *International Journal of Forecasting* 16 (4): 437–50. https://doi.org/10.1016/S0169-2070(00)00065-0.

University of California Berkeley and Max Planck Institute for Demographic Research. 2020. "Human Mortality Database." *Https://Www.mortality.org/*.

Villegas, Andrés M., Vladimir K. Kaishev, and Pietro Millossovich. 2018. "StMoMo: An R Package for Stochastic Mortality Modelling." *SSRN Electronic Journal*, no. 1992. https://doi.org/10.2139/ssrn.2698729.

Wagenmakers, Eric-Jan, and Simon Farrell. 2004. "AIC Model Selection Using Akaike Weights." *Psychonomic Bulletin & Review* 11 (1): 192–96. https://doi.org/10.3758/BF03206482.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wolpert, David H. 1992. "Stacked Generalization." *Neural Networks* 5: 241–59.

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2018. "Using Stacking to Average Bayesian Predictive Distributions (with Discussion)." *Bayesian Analysis* 13 (3). https://doi.org/10.1214/17-BA1091.

Appendix

Appendix A. Multiple Comparison of Forecasting Performance

Friedman's test compares the average ranking of M mortality models across multiple data. The forecasting mean squared errors are set in a matrix consisting of h rows representing different forecasting horizons and M columns which represent different mortality models. The mean squared errors for different models over different forecasting horizons are then ranked in ascending order, and the mean rank for each column is computed. If two or more mortality models tie, the average ranks are assigned. When the forecast mean squared errors differ among methods, there will be a significant difference in the sum of the ranks at least for one column. We can evaluate if the differences are statistically different by setting up the null and alternative hypotheses

$$H_0 : \text{The forecast accuracy is the same for all the models,}$$
$$H_1 : \text{The forecast accuracy differs between models.}$$

The decision rule is to reject H_0 if the Friedman test statistic, F, given as

$$F = \frac{12}{Mh} \sum_{m=1}^{M} R_m^2 - 3h(M+1),$$

is greater than a critical value of $\alpha\%$ level of significance. Here, R_m is the sum of ranks in column m and F is chi-squared distributed with $M-1$ degrees of freedom given that $h > 10$ and $M > 5$ (Demšar 2006).

If we reject the null hypothesis, H_0, at a given significance level α, then we apply the Nemenyi test to identify which models are statistically different from each other (Demšar 2006). The Nemenyi test is a two-sided procedure with the null hypothesis that the two mortality models yield similar average ranks. The Nemenyi test uses a critical difference

as the threshold to identify which models are statistically different from each other. Demšar (2006) calculate the critical difference (CD) at a particular significance level α as

$$\text{CD} = q_\alpha \sqrt{\frac{M(M+1)}{6h}},$$

where q_α is the studentized range statistic divided by $\sqrt{2}$, M is the number of models used in the comparison, and h is the number of data sets (forecasting horizons) applied. If the critical difference between the average ranking of the models is greater than the computed critical difference, then the mortality models perform statistically different from each other.

Appendix B. Combining the Mortality Models Using the CoMoMo Package

CoMoMo is a user-friendly open-source R package that combines different mortality models using weights generated by different model combination methods. CoMoMo combines the mortality rate forecasts from the generalized age-period-cohort models implemented in StMoMo package (Villegas, Kaishev, and Millossovich 2018). CoMoMo has methods for estimating model combination weights using Simple Model Averaging, Bayesian Model Averaging, Model Confidence Set, and Stacked Regression Ensemble. CoMoMo uses the block cross-validation approach implemented in the StMoMo package. Currently, CoMoMo is available in Github at https://github.com/kessysalvatory/CoMoMo. We shall make the CoMoMo available in the CRAN after peer review of this paper. In what follows, we use England and Wales male mortality data to describe the steps required to implement different model combination in CoMoMo.

1. Install the CoMoMo package

The CoMoMo development version can be installed with the following commands:

```
devtools::install_github("amvillegas/StMoMo", ref = "GroupLasso", force = TRUE)
devtools::install_github("kessysalvatory/CoMoMo")
```

2. Download the mortality data

To fit the mortality models to England and Wales males data, we need to download the data from the Human Mortality Database. We do that using the demography package (Hyndman et al. 2019).

```
library(demography); library(StMoMo)
MorData <- hmd.mx(country = 'GBRTENW', username = username, password = password)
DataStMoMo <- StMoMoData(MorData, "male")
agesFit <- 50:89; yearsFit <- 1960:1990
nAg <- length(agesFit); nYr <- length(yearsFit)
```

The username and password above are for the Human Mortality Database and should be replaced appropriately.

3. Define the mortality models

We can specify the list of the mortality models that we want to combine. Currently, we can only select the generalized age-period-cohort mortality models supported by StMoMo.

```
LC <- lc(); APC <- apc(); CBD <- cbd(link = "log"); M7 <- m7(link = "log")
RH <- rh(approxConst = TRUE); PLAT <- plat()
models <- list("LC" = LC, "RH" = RH, "APC" = APC, "CBD" = CBD, "M7" = M7, "PLAT" = PLAT)
```

4. Generate the metadata

We generate the metadata for the stacked regression ensemble using block cross-validation described in Subsection 6.1. First, we train the mortality models via cross-validation to produce the cross-validated forecasts at the forecasting horizon h. We then combine the mortality rate forecasts and observed mortality rates to form metadata. We use the function stackMetadata(models, data = NULL, Dxt = NULL, Ext = NULL, ages.fit = NULL, years.fit = NULL, ages = NULL, years = NULL, h = NULL) to generate the metadata which is of class stackmeta as follows:

```
library(CoMoMo)
metaData <- stackMetadata(models, data = DataStMoMo, ages.fit = agesFit, years.fit = yearsFit,
                          h = 15)
```

5. Compute the weights

We can compute the weights using different model combinations methods as follows:

- Stacked Regression Ensemble

We can use different meta-learners as explained in Subsection 6.2 to estimate the weights. The `CoMoMo` package currently supports linear (`Linear`), non-negative least regression (`nnls`), ridge (`Ridge`), lasso (`Lasso`), and elastic net (`Elastic`) regressions as the meta-learners. The `nnls` is the default meta-learner because it has proven to have outstanding performance than most other meta-learners. We use the function `stack()`, which takes an argument of the class `stackmeta` and a meta-learner to compute the weights as follows:

```
# When normalize = TRUE all the weights sum to a unit
stack_nnls_weight <- stack(metaData, metalearner = "nnls", normalize = TRUE)
# When normalize = False all the weights may not sum to a unit
stack_elastic_weight<- stack(metaData, metalearner = "Elastic", normalize = FALSE)
```

We can plot the weights using the function `plot()` and it supports all the features of `ggplot2` (Wickham 2016).

```
library(ggplot2)
plot(stack_nnls_weight) + labs(x = "Forecasting Horizon", y = "Weight") +
theme(panel.grid.major = element_blank(), panel.background = element_blank(),
axis.line = element_line(colour = "black")) + theme(legend.text = element_text(size = 10)) +
theme(plot.title = element_text(size = 10, hjust = 0.5)) + theme(legend.position="bottom")+
scale_colour_manual(values = c("brown", "blue", "red","purple","orange","skyblue4")) +
scale_linetype_manual(values = c("solid", "dashed", "dotted", "solid", "dotdash", "twodash"))
```

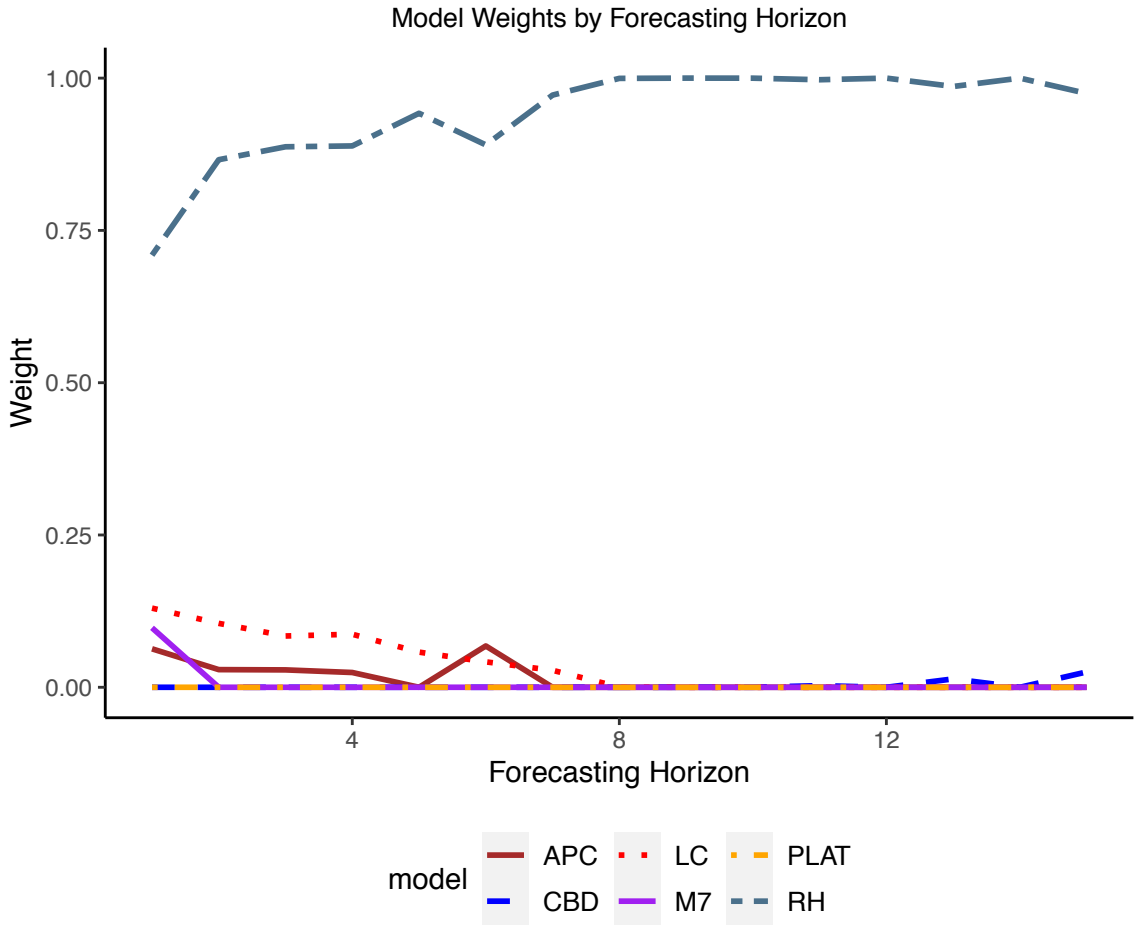

Figure 13: Horizon-specific optimal combining weights learned using non-negative least squares (SRN) for England and Wales males mortality data from 1960 to 1990 and ages 50 to 89.

Note that Figure 13 is similar to the second graph in the third panel of Figure 6.

- Bayesian Model Averaging (BMA)

We estimate the Bayesian model weights using the function `bma(models, method = "cv", data = NULL, Dxt = NULL, Ext = NULL, ages.fit = NULL, years.fit = NULL, ages = NULL, years = NULL, holdout, h = NULL)`. We can use either a single-validation set (`sv`) or a cross-validation approach (`cv`) to estimate the weights. The `holdout` is the amount of the data withheld for estimating the projection bias when we specify the `method = sv`. The argument `holdout` takes a default value of one-third of the training data.

```
bma_weight_val <- bma(models, data = DataStMoMo, ages.fit = agesFit, years.fit = yearsFit,
                      h = 15, method = "sv")
bma_weight_cv <- bma(models, data = DataStMoMo, ages.fit = agesFit, years.fit = yearsFit,
                     h = 15, method = "cv")
```

- Model Confidence Set (MCS)

We estimate the weights using the function `mcs(models, method = "cv", data = NULL, Dxt = NULL, Ext = NULL, ages.fit = NULL, years.fit = NULL, ages = NULL, years = NULL, holdout, h = NULL, B = 5000, l = 3, alpha = 0.1)`. `B` is the bootstrap samples with default sample of `B = 5000`, `l` is the block length with the default value of `l=3` and `alpha` is the level of the test with the default value of `alpha = 0.1`.

```
mcs_weight_val <- mcs(models, data = DataStMoMo, ages.fit = agesFit, years.fit = yearsFit,
                      h = 15, method = "sv")
mcs_weight_cv <- mcs(models, data = DataStMoMo, ages.fit = agesFit, years.fit = yearsFit,
                     h = 15,  method = "cv")
```

6. Fit the mortality models

We use the function `fitCoMoMo(models, data = NULL, Dxt = NULL, Ext = NULL, ages.fit = NULL, years.fit = NULL, ages = NULL, years = NULL)` to fit multiple mortality models as follows:

```
modelFits <- fitCoMoMo(models, data = DataStMoMo, ages.fit = agesFit, years.fit = yearsFit)
```

7. Combine the fitted mortality models and combination weights.

We use the function `CoMoMo()` to combine the fitted mortality models and different combination weights as follows:

```
modcom <- CoMoMo(modelFits, weight = stack_nnls_weight)
```

8. Forecast the mortality rates

We use the function `forecast()` to generate the combined mortality rates using the weights from different combination methods as follows:

```
morFor <- forecast(modcom, h = 15)
```