


RESEARCH ARTICLE

Forecasting mortality rates with a coherent ensemble averaging approach

Le Chang¹  and Yanlin Shi^{2,*}

¹Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, ACT 2601, Australia and ²Department of Actuarial Studies and Business Analytics, Macquarie University, Sydney, NSW 2019, Australia

*Corresponding author. E-mail: yanlin.shi@mq.edu.au

Received: 23 May 2022; **Revised:** 28 September 2022; **Accepted:** 12 October 2022;

First published online: 25 November 2022

Keywords: Mortality forecasting; ensemble averaging; age coherence; smoothness penalty

JEL codes: C18; C32; C52; C53

Abstract

Modeling and forecasting of mortality rates are closely related to a wide range of actuarial practices, such as the designing of pension schemes. To improve the forecasting accuracy, age coherence is incorporated in many recent mortality models, which suggests that the long-term forecasts will not diverge infinitely among age groups. Despite their usefulness, misspecification is likely to occur for individual mortality models when applied in empirical studies. The reliableness and accuracy of forecast rates are therefore negatively affected. In this study, an ensemble averaging or model averaging (MA) approach is proposed, which adopts age-specific weights and asymptotically achieves age coherence in mortality forecasting. The ensemble space contains both newly developed age-coherent and classic age-incoherent models to achieve the diversity. To realize the asymptotic age coherence, consider parameter errors, and avoid overfitting, the proposed method minimizes the variance of out-of-sample forecasting errors, with a uniquely designed coherent penalty and smoothness penalty. Our empirical data set include ten European countries with mortality rates of 0–100 age groups and spanning 1950–2016. The outstanding performance of MA is presented using the empirical sample for mortality forecasting. This finding robustly holds in a range of sensitivity analyses. A case study based on the Italian population is finally conducted to demonstrate the improved forecasting efficiency of MA and the validity of the proposed estimation of weights, as well as its usefulness in actuarial applications such as the annuity pricing.

1. Introduction

The ongoing improvements in human life expectancies over the past few decades have introduced outstanding challenges in predicting mortality scenarios. As plotted in Figure 1(a), for the Italian population, consistent mortality improvements are observed for ages 0–100 over the investigated period 1950–2016. This issue is concerning, since accurate forecasts are essential to the preparing of plans by governments, the designing of pension schemes and annuity products, and the reserving by insurance companies. Specifically, the longevity risk, such that mortality rates are underestimated, can oblige more-than-expected costs for life annuity providers and defined-benefit pension funds. This risk therefore may negatively influence the global longevity risk market for pension liabilities, the size of which is around 60–80 trillion USD as of 2018 (Blake *et al.*, 2019).

To understand and reduce the risks related to mortality and longevity, mortality modeling and forecasting have become standard mitigation tools. Among existing methods, one popular stream is based on the seminal work of Lee and Carter (1992), which is widely known as the Lee–Carter (LC) model. Influential extensions on LC have been proposed in the actuarial literature. Popular models include the Renshaw–Haberman (RH) model (Renshaw and Haberman, 2006), functional demographic model,

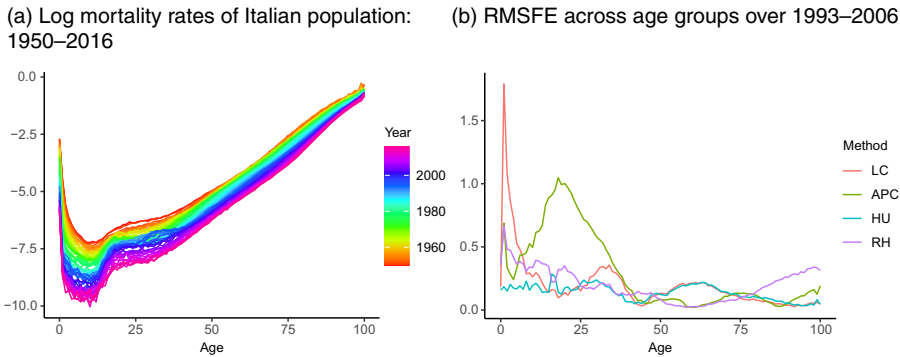


Figure 1. Italian mortality data and preliminary analyses.

or HU model proposed in Hyndman and Ullah (2007) and age-period-cohort (APC) model studied in Cairns *et al.* (2009). Despite their popularity, a major drawback of LC and its extensions is the lack of age coherence in forecasting. In other words, even for adjacent ages, forecasts produced by the LC model may differ infinitely in the long run and are not biologically reasonable.

Existing literature has proposed a range of parametric specifications to resolve this issue. An influential work by Li and Lu (2017) develops the spatial–temporal autoregressive (STAR) model. The desirable age coherence is realized by its established cointegration feature. Another novel attempt is the age-coherent extensions of the LC model using the hyperbolic (LC-H) and geometric (LC-G) convergences in the relative speeds of age-specific mortality decline (Gao and Shi, 2021). A similar principal is adopted to the sparse VAR (SVAR) model by Li and Shi (2021), and we name the SVAR model using hyperbolic and geometric convergences in the long-term mean by SVAR-H and SVAR-G, respectively. Despite the realized age coherence, drawbacks still inevitably exist for those models. For instance, the STAR model employs an ad hoc and thus restrictive sparsity structure to resolve the inherent high-dimensionality ($p \gg N$) issue of the mortality data (Guibert *et al.*, 2019). Same as in LC, LC-H and LC-G employ an inflexible temporal parametric approach (Li and Lu, 2017). SVAR-H and SVAR-G work with differenced log mortality rates (i.e., mortality improvements) that may lose information. Given the diversity, scarcity, and long-spanning nature of the mortality data, any individual model could be misspecified. Consequently, it is not realistic to expect that a single model would provide optimal forecasts in all cases (Bates and Granger, 1969; Kleijn and Van Dijk, 2006; Genre *et al.*, 2013).

To overcome such an outstanding issue, this paper aims to employ the ensemble averaging, also known as the model averaging (MA) approach, to combine the coherent mortality models. MA is a sophisticated and well-developed approach in machine learning (Ley and Steel, 2009; Amini and Parmeter, 2012; Lessmann *et al.*, 2012; Bork *et al.*, 2020; Bravo *et al.*, 2021), which has been widely applied in recent economics and finance research and practices (see, for example, Eicher *et al.*, 2011; Mirestean and Tsangarides, 2016; Shiraya and Takahashi, 2019; Baechle *et al.*, 2020; du Jardin, 2021, among others). With respect to mortality data, Shang (2012) and Kontis *et al.* (2017) have employed various MA strategies, such as the Bayesian model averaging (BMA). A recent study by Kessy *et al.* (2021) examines the stacked regression ensembles.

Among those studies, the same weight is assigned to all ages within an individual model to ease the computation and improve the stability in forecasting. However, mortality models usually behave differently dependent on the age groups. To see this, we model the Italian data over 1950–1992 and forecast rates spanning 1993–2006. The out-of-sample root of mean squared forecast errors (RMSFE) at each age is then produced by averaging errors over all forecasting steps. The results of the LC and its extensions are plotted in Figure 1(b). Clearly, LC and APC models are less favorable over young age groups and should be assigned lower weights. Consequently, age-specific weights are more desirable to achieve the optimal forecasting performance.

In this paper, we employ the classic global minimum variance portfolio (GMVP) strategy (Markowitz, 1952) to choose the weights of the proposed MA approach. The implementation of GMVP without restriction, however, can be dangerous in mortality forecasting, due to the inherent issues of the GMVP solution. For one thing, GMVP ignores the parameter error, and using the in-sample variance may lead to an overfitting issue when forecasting mortality rates. For another, if a dominant model exists, GMVP solution may lead to a “winner-take-all” issue, such that most weights are attributed to this dominant model.

To combat against those issues, we employ two strategies. First, following the approach of Shi (2022a), the variance is minimized for out-of-sample forecast errors, rather than for in-sample residuals. It is well known that out-of-sample error, such as MSFE, consists of both process and parameter risks. Hence, the first issue of GMVP is well addressed via this strategy. Second, to realize the desirable diversity, we employ ten models in the ensemble space: STAR, LC-H, LC-G, SVAR-H, SVAR-G, LC, SVAR, APC, HU, and RH models. More importantly, the diversified age-specific weights are permitted to allow for dynamics as evidenced in Figure 1(b).

Further, two additional penalties are considered in our MA approach to employ strong and reasonable prior information of the mortality data. Specifically, as described above, long-run mortality forecasts should be age coherent. To achieve this, a coherent penalty is imposed to reduce weights of age-incoherent models. Moreover, due to the heterogeneity in out-of-sample forecasts among models, drastic changes of weights from one age to another may result in abrupt differences between forecast mortality rates. For instance, the forecast rate of age 50 in some year might be higher than that of age 51. In life insurance practice, this unreasonably suggests that younger policyholders will pay for higher premiums than older policyholders. To resolve this issue, we impose a smoothness penalty in the optimization to reduce abrupt changes in weights of the adjacent ages in the same model. Both penalties are then selected using a hold-out-sample strategy. In addition, a non-negativity constraint is applied to improve interpretability of weights.

Altogether, we consider eleven models in this paper: STAR, LC-H, LC-G, SVAR-H, SVAR-G, LC, SVAR, APC, HU, RH, and MA. The empirical data sourced from Human Mortality Database (2019) of ten European countries are examined. The ages are one-year groups of 0–100, and the timespan is 1950–2016. We first present the baseline results of ten-step-ahead forecasts, using the popular performance measure RMSFE over both ages and years. Three sets of sensitivity analyses are also conducted, by altering individually the forecasting step, temporal coverage, and age groups. A comparison of the proposed MA approach with the influential model confidence set (MCS) proposed in Hansen *et al.* (2011) is further conducted. To systematically demonstrate its usefulness in economic and financial practices, the proposed MA approach is employed to illustrate fixed-term annuities pricing in a separate case study.

The contributions of this paper are fivefold. First, this study is among the first to comprehensively consider both age-coherent and non-age-coherent specifications and adopt age-specific weights in mortality forecasting using the ensemble averaging strategy. The adopted objective function effectively achieves the benefit of GMVP solution. The diversity and overfitting issues of GMVP are simultaneously addressed, via working with out-of-sample forecast errors and imposing two penalties to honor the reliable prior information of mortality data. The estimated weights are therefore optimal by balancing the randomness of small sample sizes, consistency of estimation, and adopting useful prior information. Essentially, our MA method is a supervised machine learning technique. Compared to other techniques, such as the neural networks, the MA approach is more transparent and does not suffer the “black-box” issue. Second, the proposed MA method provides asymptotically age-coherent forecasts. This is both theoretically verified and empirically visualized in the long-term forecast of our case study. Thus, the desirable age coherence feature in the ensemble is not lost when the MA is executed, even when age-incoherent models are included in the ensemble. Third, our empirical results demonstrate that the proposed MA method can significantly improve out-of-sample forecasting performance of all individual models in the ensemble. According to the RMSFE, MA ranks the first for eight out of ten populations. This result is much robust in all sets of sensitivity analyses. Fourth, we consider

other modeling selection/averaging approaches, such as the MCS and BMA. The outperformance of our proposed approach is demonstrated, and the validity of the proposed estimation approach is illustrated by presenting the estimated age-specific weights. Fifth, a case study is provided to demonstrate the effectiveness of the MA method in economic and financial practice. The narrower width of prediction intervals suggests that MA can be more efficient in measuring uncertainties in mortality forecasting. This is critical to practices like the annuity pricing, and an illustration is thoroughly conducted over different starting ages and maturity terms. Thus, the proposed MA approach can be a widely useful tool in forecasting mortality data for actuarial products. Implications on other actuarial modeling practices are also briefly outlined.

The rest of this paper is organized as follows. In Section 2, we review the specification and features of the LC model and briefly outline its extensions. Five coherent alternative models are explained in Section 3. The MA approach is proposed in Section 4. We conduct empirical studies with robustness checks and additional analysis in Section 5. Section 6 presents a case study, and Section 7 concludes.

2. The Lee–Carter model

Among the existing models, the seminal work proposed in Lee and Carter (1992), known as the Lee–Carter (LC) model, is one of the most popular approaches. Essentially, LC belongs to the family of factor models and decompose mortality rates into age-dependent and time-series factors. Specifically, the logged central mortality rate of age x in year t , or $\ln m_{x,t}$, is specified as follows:

$$\ln m_{x,t} = a_x + b_x k_t + \varepsilon_{x,t}, \tag{2.1}$$

where $x \in (1, \dots, N)$ and $t \in (1, \dots, T)$. In terms of the factors, a_x is the average age trend of $\ln m_{x,t}$ across t , b_x is the relative speed of decline in $\ln m_{x,t}$ at each x , and k_t is a time-dependent index at each t . Additionally, $\varepsilon_{x,t}$ is the residual at age x and time t . It is usually assumed that $\varepsilon_{x,t}$ follows a multi-Gaussian distribution with zero means and independence on the temporal dimension.

Regarding the estimation of LC, the singular value decomposition (SVD) is the most frequently employed technique. The detailed procedure to implement the SVD is described in Trefethen and Bau (1997). It is worth noting that there are nonunique solutions to produce \hat{b}_x and \hat{k}_t via the SVD. Necessary constraints are therefore imposed to obtain unique estimates, such that $\sum_{x=1}^N \hat{b}_x = 1$ and $\sum_{t=1}^T \hat{k}_t = 0$.

Remark 1. Much of the popularity of LC estimated by SVD is its self-explanatory parameters. Intuitively, a_x describes the overall, or “average,” pattern of the historical logged mortality rates on the age dimension. b_x and k_t are orthogonal norms, which represent the relative dynamics of $\ln m_{x,t}$ on the age and temporal dimension, respectively.

To forecast the future logged mortality rates, or $\ln \hat{m}_{x,T+h}$, \hat{a}_x and \hat{b}_x , are kept constant in Lee and Carter (1992). The temporal factor \hat{k}_t , however, is intrinsically viewed as a random walk with drift process, such that

$$\hat{k}_t = \hat{k}_{t-1} + d_k + e_t, \tag{2.2}$$

where d_k is the average change in \hat{k}_t , and e_t are independently and identically distributed (iid) Gaussian sequences with zero means. Using (2.2) and the property of a random walk process, the h -step-ahead forecast \hat{k}_{T+h} is produced as $\hat{k}_T + hd_k \forall h \geq 1$. Thus, the h -step-ahead forecast of logged mortality rate can be obtained as follows.

$$\ln \hat{m}_{x,T+h} = \hat{a}_x + \hat{b}_x(\hat{k}_T + hd_k).$$

Based on the LC model, many extensions using the factor framework are proposed. For instance, Renshaw and Haberman (2006) include the cohort effect in the LC specification, which considers stochastic features of people born at different time periods (denoted as the RH model). An age-period-cohort (APC) model is derived from RH by constraining the age-specific loadings of temporal and cohort

effects (Cairns *et al.*, 2009). Hyndman and Ullah (2007) extend the LC by introducing more principal components in the decomposition of age and temporal effects, which are fitted via a functional modeling approach (denoted as the HU model).

3. Coherent mortality models

One of the major drawbacks of the LC model is the lack of age coherence in forecasts (see, for instance, Gao and Shi, 2021, among others). The same issue also applies to its popular extensions, including RH, APC, and HU models. The concept of age coherence is first proposed in the influential work of Li and Lu (2017), which ensures biologically reasonable forecasts of mortality rates in the long run. Intuitively, age coherence means that predicted mortality rates (of the same population) across ages should not differ indefinitely. Consistent with Li and Lu (2017), Gao and Shi (2021) and Li and Shi (2021), we formally define age coherence as follows.

Definition 1. *Age coherence means that for the h -step-ahead forecasts, $|\ln \hat{m}_{i,T+h} - \ln \hat{m}_{j,T+h}| = O_p(1)$, $\forall i, j \in (1, \dots, N)$. For the concept of $O_p(\cdot)$, given a set of random variables X_n and a corresponding set of constants a_n , if $X_n = O_p(a_n)$, then for any $\varepsilon > 0$, there exists a finite $M > 0$ and a finite $N > 0$ such that $P(|X_n/a_n| \geq M) \leq \varepsilon \forall n > N$. That is, when $h \rightarrow \infty$, $|\ln \hat{m}_{i,T+h} - \ln \hat{m}_{j,T+h}|$ will not diverge to infinity.*

Remark 2. Note that age coherence can also be defined using the original mortality scale, such that for the h -step-ahead forecasts, we have $|\hat{m}_{i,T+h}/\hat{m}_{j,T+h}| = O_p(1)$, $\forall i, j \in (1, \dots, N)$. To cope with the original scale, a Poisson distribution (assumed for death counts) associated with a log-linear regression is often employed, and the risk exposure (population counts) is usually employed as the offset factor (Brouhns *et al.*, 2002).

To address the lack of age coherence in the LC model, we describe five recently proposed alternatives in this section. Among them, two are based on the LC framework, whereas three are developed from the vector-autoregressive (VAR) model.

3.1. The spatial-temporal autoregressive (STAR) model

Despite its popularity, a general VAR model cannot be directly fitted to a typical mortality data set for two reasons. First, a VAR model requires a stationary response variable, while mortality rates over time are clearly trending and therefore nonstationary. Second, the unknown parameters (p) exceed the observations (T) in a nonconstrained VAR framework. For instance, with an intermediate number of age groups (N), say 50, the $p = 50 \times 51 \gg T$ issue will arise, given that observations are available only for a few dozens of years.

To resolve those issues, Li and Lu (2017) propose the spatial-temporal autoregressive (STAR) model to study and forecast mortality data. On the temporal dimension, it considers the Granger causality and cointegration to resolve the stationarity problem. On the age dimension, the STAR model adopts the sparse spatial information to reduce the dimensionality of p with the follow specification:

$$\begin{aligned} y_{1,t} &= \alpha_1 + y_{1,t-1} + \epsilon_{1,t} \\ y_{2,t} &= \alpha_2 + \beta_{2,1}y_{1,t-1} + (1 - \beta_{2,1})y_{2,t-1} + \epsilon_{2,t} \\ y_{i,t} &= \alpha_i + \beta_{i,i-2}y_{i-2,t-1} + \beta_{i,i-1}y_{i-1,t-1} + (1 - \beta_{i,i-2} - \beta_{i,i-1})y_{i,t-1} + \epsilon_{i,t}, \end{aligned} \quad (3.1)$$

where we let $y_{i,t} = \ln m_{i,t}$ for simplicity, $i = 3, 4, \dots, N$, and $t = 1, 2, \dots, T$. The residual sequence $\epsilon_{i,t}$ is similarly assumed as in the LC model.¹

¹For illustration purpose, note that the STAR and all other VAR-type models investigated in this study consider only one lag in the temporal specification.

Rewriting (3.1) in a VAR(1) form, we have that

$$y_t = \alpha + \mathbf{B}y_{t-1} + \epsilon_t, \tag{3.2}$$

where $y_t = (y_{1,t}, y_{2,t}, \dots, y_{N,t})'$, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)'$, $\epsilon_t = (\epsilon_{1,t}, \epsilon_{2,t}, \dots, \epsilon_{N,t})'$, and

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \dots & \dots \\ \beta_{2,1} & 1 - \beta_{2,1} & 0 & \dots & \dots \\ \beta_{3,1} & \beta_{3,2} & 1 - \beta_{3,1} - \beta_{3,2} & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \dots & 0 & \beta_{N,N-2} & \beta_{N,N-1} & 1 - \beta_{N,N-2} - \beta_{N,N-1} \end{bmatrix} \tag{3.3}$$

To ensure invertibility and interpretability, it is required that $0 < \beta_{i,i-2} < 1$, $0 < \beta_{i,i-1} < 1$, and $0 < \beta_{i,i-2} + \beta_{i,i-1} < 1$ for all $i > 2$. When $i = 2$, we need that $0 < \beta_{2,1} < 1$.

The interpretation of the STAR model parameters is straightforward. For age $i > 2$, $(1 - \beta_{i,i-2} - \beta_{i,i-1})$ is the usual temporal effect of the first time lag $y_{i,t-1}$ on $y_{i,t}$. As for the mortality practice, cohort effects are more important, which are measured by $\beta_{i,i-1}$ (the same cohort) and $\beta_{i,i-2}$ (the younger neighboring cohort). The relevant positive constraints not only ensure the invertibility (and thus stationarity) of $y_{i+1,t} - y_{i,t}$ but also the interpretability of temporal and cohort effects. A negative measure is usually unexplainable, and nor it is revealed in the existing literature.

In addition to the interpretability, the specification described in (3.2) and (3.3) has attractive statistical features. As shown in Li and Lu (2017), with the assumption that all $y_{i,t}$ are $I(1)$ (commonly used in mortality research), all neighboring age pairs $y_{i,t}$ and $y_{i+1,t}$ are cointegrated with order $(-1, 1)$. This approach successfully resolves the stationarity issue and results in age coherence. More specifically, cointegration suggests that $y_{i+1,t} - y_{i,t}$ is stationary, which leads to a constant long-run mean. Further, the total number of parameters is largely reduced from $N^2 + N$ to $3N - 3$, which is feasible to solve under a general penalized least squares framework with closed-form solutions.

Remark 3. To reduce the randomness of estimates owing to limited data availability, Li and Lu (2017) conduct the estimation in a penalized least squares fashion, which imposes a Tikhonov (L_2) regularization in the loss function. In other words, three smoothing penalties (corresponding to α_i , $\beta_{i,i-1}$ and $\beta_{i,i-2}$, respectively, for age i) need to be preselected before a solution can be derived. Essentially, this assumes that the coefficients across ages should change smoothly. See Section 3.4 for how those penalties are selected.

The forecasting of STAR is therefore performed in an iterative fashion, where

$$\begin{aligned} \hat{y}_{t+1} &= \hat{\alpha} + \hat{\mathbf{B}}y_t \\ \hat{y}_{t+h} &= \hat{\alpha} + \hat{\mathbf{B}}\hat{y}_{t+h-1} \end{aligned} \tag{3.4}$$

and $h > 1$. Due to the established co-integration feature, $\hat{y}_{i,t+h} - \hat{y}_{j,t+h}$ is stationary for all $i, j = 1, \dots, N$ and $i \neq j$, and $h \geq 1$. This then straightforwardly ensures the age coherence in forecasting.

3.2. Age-coherent extensions of the LC model

Other than working with the VAR framework, a recent study by Gao and Shi (2021) proposes two effective age-coherent extensions of the LC model. Both are motivated by Li *et al.* (2013) to rotate $\hat{b}_{x,h}$ gradually over time. The rotation is supported by the fact that mortality decline decelerates at younger ages and accelerates at old ages (Li *et al.*, 2013). Also, to realize age coherence in the long run, Gao and Shi (2021) require that $\hat{b}_{x,h}$ eventually converge (rotate) to a constant $1/N$ for all ages (a flat line). The rationale of $1/N$ comes from the constraint that $\sum_{x=1}^N \hat{b}_{x,h} = 1$ when $h \geq 1$ and $\hat{b}_{i,h} = \hat{b}_{j,h}$ when $h \rightarrow \infty$.

Specifically, the first extension adopts the autoregressive (AR) framework. Under the AR(1) scenario, the time-varying $\hat{b}_{x,h}$ (represented by $\hat{b}_{x,h}^G$) is specified below.

$$\hat{b}_{x,h}^G = r_x^l (\hat{b}_{x,h-1}^G - 1/N) + 1/N = (r_x^l)^h (\hat{b}_x - 1/N) + 1/N, \tag{3.5}$$

where $h \geq 1$ and $\hat{b}_{x,0}^G = \hat{b}_x$. To ensure the stationarity, the coefficient r_x^l strictly falls between 0 and 1. In such a case, $\hat{b}_{x,h}^G$ converges to the long-run mean $1/N$ geometrically. For a larger (smaller) r_x^l , the speed of convergence is slower (faster).

The second extension employs the hyperbolic decay, which may achieve the final convergence slower than the AR specification. This concept is often refer to as long memory in time-series analysis for economic and financial practices (see, for example, Smallwood and Norrbin, 2006; Ho and Shi, 2020, among others). When assumed to decay hyperbolically, the time-varying $\hat{b}_{x,h}$ (represented by $\hat{b}_{x,h}^H$) is described by

$$\hat{b}_{x,h}^H = \delta_h(d_x^l) (\hat{b}_x - 1/N) + 1/N, \tag{3.6}$$

where

$$\delta_h(d_x^l) = \frac{k - 1 + d_x^l}{k} \delta_{h-1}(d_x^l) \text{ and } \delta_0(d_x^l) = 1.$$

The hyperbolic parameter, or d_x^l , needs to fall between 0 and 1 for a stationary rotation. In this case, since $\delta_h(d_x^l) \rightarrow 0$ when $h \rightarrow \infty$, $\hat{b}_{x,h}^H$ will eventually converge to $1/N$ and thus leads to the age-coherent forecasts. The speed of decay is slower (faster) for larger (smaller) d_x^l .

Note that both r_x^l and d_x^l are age dependent, and their estimation is not a trivial issue due to the large sample size on the age dimension. Gao and Shi (2021) employ the inversed Epanechnikov kernel to reduce the complexity of estimation. More specifically, let τ defined as a scaled index x/N with $x \in \{1, \dots, N\}$, when evaluated at N , the inversed Epanechnikov kernel is determined by $1 - K_{b^w}(\tau - 1) = 1 - K(\frac{\tau-1}{b^w})$, where $K(\tau - 1) = 0.75(1 - (\tau - 1)^2)_+$. The parameter b^w is the corresponding kernel bandwidth and constrained to fall in the range of (0, 1]. Hence, the parametric structures of r_x^l and d_x^l are simplified to follow the pattern of $1 - K_{b^w}(\tau - 1)$ across ages.

Remark 4. The rationale of employing the inversed Epanechnikov kernel is thoroughly discussed in Gao and Shi (2021). In short, the bandwidth determines a ‘‘cutoff’’ age, and the kernel describes a pattern consistent with the sense that mortality decline is more difficult for older ages (Li *et al.*, 2013), starting from this cutoff. For instance, with ages 0–100 and a bandwidth of 0.7, all ages younger than 70 will share the same mortality decline speed. For ages older than 70, this decline is more difficult for an age further away from 70. Essentially, only two parameters are needed when (3.5) or (3.6) is employed: the decay parameter of the first age (r_1^l or d_x^l) and the kernel bandwidth b^w . See Section 3.4 for how those parameters are selected.

Once r_x^l and d_x^l become available, the forecast logged mortality rates of the LC extension with $\hat{b}_{x,h}^G$ (LC-G) and that with $\hat{b}_{x,h}^H$ (LC-H) are

$$\begin{aligned} \ln \hat{m}_{x,T+h}^G &= \hat{a}_x + \hat{b}_{x,h}^G (\hat{k}_T + h \hat{d}_k) \\ \ln \hat{m}_{x,T+h}^H &= \hat{a}_x + \hat{b}_{x,h}^H (\hat{k}_T + h \hat{d}_k), \end{aligned} \tag{3.7}$$

where \hat{a}_x and \hat{k}_t are identical to those of the original LC model. In other words, the in-sample fits of LC-G and LC-H are the same as those described in (2.1). Recall that when $h \rightarrow \infty$, $\hat{b}_{x,h}^G$ or $\hat{b}_{x,h}^H$ converges to $1/N$. Thus, following (3.7), using either LC-G or LC-H, forecast rates of different ages will not differ infinitely and are thus age coherent.

3.3. The sparse VAR (SVAR) model and coherent extensions

Despite its attractive features, STAR’s parametric structure as of (3.3) is ad hoc and relatively inflexible. Specifically, the effects of all cohorts other than the same and next younger one are forced to be zero. This may limit a comprehensive analysis of the effects of all possible cohorts.

To address this, Guibert *et al.* (2019) recently work on the mortality improvements denoted by $\Delta y_{i,t} = y_{i,t} - y_{i,t-1}$ and propose a sparse VAR (SVAR) model. Since $y_{i,t}$ is assumed I(1), the differentiation resolves the stationarity issue of a VAR system. Moreover, the SVAR model adopts a pure data-driven method to select nonzero coefficients via the ENET penalty estimation, rather than working with the ad hoc structure as in (3.3). The SVAR model is specified below.

$$\Delta \mathbf{y}_t = \mathbf{a} + \sum_{v=1}^V \mathbf{B}_v \Delta \mathbf{y}_{t-v} + \boldsymbol{\epsilon}_t, \tag{3.8}$$

where $\Delta \mathbf{y}_t = (\Delta y_{1,t}, \Delta y_{2,t}, \dots, \Delta y_{N,t})'$, V is the preselected AR lags and the sparsity of \mathbf{B}_v is determined by the ENET penalty without any constraints. The forecast is performed similarly to (3.4) as a usual VAR-type model for $\Delta \mathbf{y}_t$, and the forecast for the original logged mortality rate is computed as $\widehat{\mathbf{y}}_{t+h} = \mathbf{y}_t + \sum_{l=1}^h \Delta \widehat{\mathbf{y}}_{t+l}$.

However, the desirable feature of age coherence is lost when working with mortality improvements directly, as in the SVAR model. That is, as long as the estimated long-term mean of $\Delta y_{i+1,t} - \Delta y_{i,t}$ (denoted by $\Delta \widehat{\mu}_{i+1}$) is not $O_p(1/h)$, the long-run forecast $\widehat{y}_{i+1,t+h} - \widehat{y}_{i,t+h} = y_{i+1,t} - y_{i,t} + h \Delta \widehat{\mu}_{i+1}$ will still grow to reach infinity, when $h \rightarrow \infty$.

To produce age-coherent forecasts, a recent study of Li and Shi (2021) extends the approach of Gao and Shi (2021) to the SVAR framework. In particular, the intercept $\hat{\alpha}_x$ is allowed to be time-varying for age x , and Li and Shi (2021) adopt the hyperbolic decay as follows:

$$\hat{\alpha}_{x,h}^H = \delta_h (d_x^s) (\hat{\alpha}_x - \hat{\alpha}_x^*) + \hat{\alpha}_x^*,$$

where $\hat{\alpha}_x$ is the estimate of a usual SVAR model as of (3.8), and $\hat{\alpha}_x^*$ is the long-term mean of $\hat{\alpha}_x$. This specification is named SVAR with hyperbolic decay, or SVAR-H model. A similar specification can then be derived following the geometric decay as in (3.5):

$$\hat{\alpha}_{x,h}^G = r_x^s (\hat{\alpha}_{x,h-1}^G - \hat{\alpha}_x^*) + \hat{\alpha}_x^*,$$

which is named SVAR with geometric decay, or SVAR-G model. The age-dependent r_x^s and d_x^s can be specified using the inversed Epanechnikov kernel, as for LC-G and LC-H models.

Forecasts of $\Delta \widehat{\mathbf{y}}_{T+h}$ using the SVAR-G or SVAR-H model are the same as in (3.4). Logged mortality rates can then be forecast by

$$\widehat{y}_{x,T+h} = y_{x,T} + \sum_{k=1}^h \Delta \widehat{y}_{x,T+k}.$$

Remark 5. Since $\Delta y_{x,t}$ is stationary, $\widehat{y}_{x,T+h}$ will approach $y_{x,T} + h \widehat{\mu}_x$ in the long run, where $\widehat{\mu}_x$ is the estimated mean of $\Delta y_{x,t}$. In a matrix form, it is easy to see that

$$\boldsymbol{\mu} = [I - \mathbf{B}_1 - \dots - \mathbf{B}_V]^{-1} \boldsymbol{\alpha}. \tag{3.9}$$

Further, recall that when $h \rightarrow \infty$, $\hat{\alpha}_{x,T+h}^G$ or $\hat{\alpha}_{x,T+h}^H$ converges to $\hat{\alpha}_x^*$. To achieve age coherence, $\hat{\alpha}_x^*$ is obtained such that all elements of $\boldsymbol{\mu}$ are identical. In practice, estimates of $\hat{\alpha}_x^*$ are produced as follows:

$$\hat{\alpha}_x^* = [I - \widehat{\mathbf{B}}_1 - \dots - \widehat{\mathbf{B}}_V] \widehat{\boldsymbol{\mu}}^*,$$

where $\widehat{\mathbf{B}}_v$ are in-sample estimates of a usual SVAR model, and $\widehat{\boldsymbol{\mu}}^*$ consists of identical elements of $\widehat{\mu}_x^*$, which can be estimated as the average of $\widehat{\mu}_x$ over all ages. $\widehat{\mu}_x$ is obtained from the in-sample estimates of a usual SVAR model with (3.9).

Note that there are three tuning parameters in both SVAR-H and SVAR-G models: the sparsity penalty parameter (λ_s), the hyperbolic or geometric decay parameter of the first age group (d_1^s or r_1^s), and the kernel bandwidth (b^w). See Section 3.4 for how those parameters are selected.

3.4. Tuning parameters selection

All the five age-coherent mortality models described in this section require some preselected tuning parameters, either before the model is fitted (STAR, SVAR-H, and SVAR-G) or at the forecasting stage (LC-H, LC-G, SVAR-H, and SVAR-G). Due to the time-series nature, the usual cross-validation technique to select tuning parameters is not directly applicable.

In practice, a popular strategy for time-series data is to adopt the expanding-window approach explained in Hyndman and Athanasopoulos (2018) to collect short-term out-of-sample forecasts. However, since age coherence is a long-term property, we follow Gao and Shi (2021) and adopt the hold-out-sample approach to select tuning parameters in all cases. Specifically, the selection aims to minimize

$$\text{RMSFE} = \sqrt{\frac{1}{N(T/4)} \sum_{i=1}^N \sum_{h=1}^{T/4} (\ln \hat{m}_{i,3T/4+h} - \ln m_{i,3T/4+h})^2}, \quad (3.10)$$

where RMSFE is the root of mean squared forecasting errors, and the evaluation period is given by the last fourth ($[3T/4 + 1, T]$) of the data in our study. A high-level summary of the selection procedure is listed in Table 1 of Supplement Material for each model.

4. The proposed ensemble averaging approach

Despite the achieved desirable age coherence, among the five investigated models in Section 3, it is usually difficult to pick up one single model that uniformly outperforms the rest. This may be attributed to the fact that each model has its merits and drawbacks. For instance, the STAR model does not consider the empirically sensible decline in decaying speed for older ages. In contrast, the cohort effect may be “ignored” in extensions of LC and SVAR. Mortality data are often observed for a long time span and over countries/regions. The cross-sectional (temporal) heterogeneity may favour certain assumptions over some populations (during some periods). In addition, despite the long-term issue of age incoherence, influential models, such as LC, SVAR, APC, HU, and RH, may demonstrate favorable performance for certain age groups/populations in short term. Consequently, no single model may be well specified to capture the features and dynamics in mortality modeling entirely.

Among the existing literature, ensemble averaging or MA is an effective strategy to combat against such uncertainty (see, for example, Ley and Steel, 2009; Amini and Parmeter, 2012; Lessmann *et al.*, 2012; Bork *et al.*, 2020; Bravo *et al.*, 2021, among others). In this section, we propose an effective MA strategy using the variance-optimization-weights to realize age-coherent forecasts. Technical details are also provided.

4.1. Background: A solution from global minimum variance portfolio

In his seminal work, Markowitz (1952) discusses a solution to achieve global minimum variance for a portfolio, or the GMVP solution. The idea can be straightforwardly applied to forecasting error minimization using an MA model. Suppose that there exist a finite number ($J > 2$) of models, the forecasting error of an MA approach is

$$e_{M,t} = w_1 e_{1,t} + w_2 e_{2,t} + \dots + w_J e_{J,t},$$

Table 1. *Out-of-sample forecasting performance.*

	STAR	LC-H	LC-G	SVAR-H	SVAR-G	LC	SVAR	APC	HU	RH	MA
Austria	0.1992	0.1758	0.1841	0.2019	0.2025	0.2338	0.2040	0.3800	0.1764	0.2323	0.1750
Denmark	0.2556	0.3282	0.3276	0.3154	0.3156	0.3394	0.3207	0.2961	0.3219	0.3549	0.2902
UK	0.1004	0.1299	0.1322	0.1087	0.1084	0.1505	0.1000	0.2702	0.1286	0.0959	0.0866
Finland	0.2517	0.2434	0.2457	0.2811	0.2813	0.2641	0.2849	0.4143	0.2529	0.2829	0.2387
France	0.0985	0.1406	0.1435	0.0919	0.0930	0.1686	0.0948	0.2727	0.1104	0.1807	0.0794
Italy	0.1374	0.1233	0.1309	0.1218	0.1231	0.2072	0.1237	0.4068	0.1528	0.2369	0.1155
Netherlands	0.1434	0.1517	0.1546	0.1552	0.1553	0.1735	0.1430	0.2352	0.1831	0.1718	0.1248
Norway	0.2643	0.2281	0.2334	0.2875	0.2874	0.2898	0.3100	0.3284	0.2819	0.2694	0.2359
Spain	0.1562	0.1851	0.1899	0.1578	0.1575	0.2425	0.1581	0.3045	0.1508	0.3369	0.1368
Switzerland	0.2270	0.2717	0.2728	0.2477	0.2472	0.2901	0.2494	0.2681	0.2376	0.3094	0.2238
Mean	0.1834	0.1978	0.2015	0.1969	0.1971	0.2360	0.1989	0.3176	0.1996	0.2471	0.1707

where $e_{j,t}$ is the forecast error at time t for model j ($j \in \{1, 2, \dots, J\}$), w_j is the assigned weight for the j th model and $e_{M,t}$ is then the weighted summation of forecast errors of all J models, or the forecast error of the MA approach. Denote that $\mathbf{w} = (w_1, \dots, w_J)'$, the GMVP solution of the optimal \mathbf{w}^* is

$$\mathbf{w}^* = \Sigma_e^{-1} \boldsymbol{\iota} (\boldsymbol{\iota}' \Sigma_e^{-1} \boldsymbol{\iota})^{-1},$$

where Σ_e is a $J \times J$ variance-covariance matrix of forecast errors, and $\boldsymbol{\iota}$ is a $J \times 1$ vector of ones.

Remark 6. Note that when forecast errors of some models are strongly (and positively) correlated, it is possible to have negative weights in \mathbf{w}^* . Under a financial portfolio optimization scenario, this means a short position should be held for some assets. In the mortality case, although negative weights are sensible from a model combination perspective, their interpretation is not as straightforward. Also, negative weights might cause instability in long-term forecasting. To avoid this, the non-negative constraints should be added to the optimization issue that is usually faced in a GMVP problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}' \Sigma_e \mathbf{w} \\ \text{subject to} \quad & \boldsymbol{\iota}' \mathbf{w} = 1 \\ & \mathbf{w} \geq \mathbf{0}. \end{aligned}$$

This is a quadratic optimization issue with equality and inequality constraints, which can be solved via iteration-based algorithms (Gill *et al.*, 2019).

4.2. An MA approach for mortality forecasting

For an MA model used in mortality forecasting, the most essential issue is the objective function to be employed in the optimization. This is because that for the original GMVP solution, there are two major issues. First, since only the global risk is considered, GMVP may result in large weights in few models. This is also known as the “winner-take-all” problem and may negatively affect diversity in an MA approach. Second, the estimate is highly sensitive to parameter estimation errors of the covariance matrix. Consequently, an inappropriate covariance error matrix may lead to unrobust weights in an MA model.

In this paper, diversity is ensured in two ways. First, we employ both age-coherent and age-incoherent models, which include a total of ten specifications: STAR, LC-H, LC-G, SVAR-H, SVAR-H, LC, SVAR, APC, HU, and RH. Second, different from a single weight for all ages used in existing studies (see, for example, Shang, 2012; Kontis *et al.*, 2017; Kessy *et al.*, 2021, among others), age-specific weights are employed in all cases. As preliminarily evidenced in Figure 1(b), not all models can perform uniformly well for all age groups. This setting then effectively prevents the existence of one or few dominant models.

To avoid the potential overfitting issue of in-sample errors, we construct Σ_e using the out-of-sample errors only. Specifically, the objective function will consist of out-of-sample mean squared forecasting errors (MSFE). Since it is well known that MSFE incorporates both process risk and parameter risk, this will effectively address the second problem of GMVP. In addition, we adopt two pieces of prior information in the optimization to further improve reasonableness in mortality forecasting. First, as argued above, the MA approach should (asymptotically) result in age-coherent forecasts. Therefore, the weights assigned to age-incoherent models should be asymptotically approaching zero. A penalty is imposed in the objective function for this aim. Second, mortality rates across neighboring ages tend to demonstrate similar dynamics. However, without further constraints, abrupt changes of forecast rates will be unavoidable, even for neighboring ages, due to potential randomness introduced by small samples. Consequently, inspired by existing studies such as Li and Lu (2017), a smoothness penalty is additionally introduced to the objective function.

The optimization problem to choose optimal weights of an MA mortality model is then stated below. Note that for a mortality data set with N age groups estimated by J models (with the total J_c age-coherent models ordered first and the rest J_n age-incoherent models ordered last), the weight vector to be estimated is $\mathbf{w} = (\mathbf{w}'_1, \dots, \mathbf{w}'_N)'$ (the dimensionality is $(NJ) \times 1$), with $\mathbf{w}_x = (w_{x,1}, \dots, w_{x,J})'$. To further allow for the age coherence penalty, smoothness penalty and non-negative weights, we have the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}'\boldsymbol{\Omega}_e\mathbf{w} + \lambda_1\mathbf{w}'\mathbf{C}_1\mathbf{w} + \lambda_2\mathbf{w}'\mathbf{C}_2\mathbf{w} \\ \text{subject to} \quad & \mathcal{I}\mathbf{w} = \mathbf{1} \\ & \mathbf{w} \geq \mathbf{0}, \end{aligned} \tag{4.1}$$

where $\boldsymbol{\Omega}_e$ is an $(NJ) \times (NJ)$ matrix with $\boldsymbol{\Sigma}_{e_x}$ on the diagonal blocks and elsewhere 0. \mathbf{C}_1 and \mathbf{C}_2 are both $(NJ) \times (NJ)$ ancillary matrices to specify the quadratic age coherence penalty and smoothness penalty, respectively. λ_1 and λ_2 are the associated tuning parameters. \mathcal{I} is another $(NJ) \times (NJ)$ ancillary matrix to impose the equality constraint for each age. Specifically, \mathbf{C}_1 is a diagonal matrix with the following diagonal items:

$$\overbrace{(\underbrace{0, \dots, 0}_{J_c}, \underbrace{1, \dots, 1}_{J_n}, \dots, \underbrace{0, \dots, 0}_{J_c}, \underbrace{1, \dots, 1}_{J_n})}_{N \times J}.$$

The xj th column of \mathbf{C}_2 is

$$\mathbf{c}_{xj} = (\underbrace{0, \dots, 0}_{j-1}, \underbrace{-1}_{xj-1}, \underbrace{0, \dots, 0}_{j-1, -1, \dots}, \underbrace{1}_{j-1}, \underbrace{0, \dots, 0}_{j-1}, \dots)'$$

That is, there are $(N - 1)(-1)$'s and one 1 in each column, with all other elements being 0. An example is provided below to illustrate the specification when $N = 2, J_c = 1$ and $J_n = 1$. Also,

$$\mathcal{I} = \begin{pmatrix} \iota' & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \iota' & \dots & \mathbf{0} \\ \dots & \dots & \dots & \dots \\ \mathbf{0} & \mathbf{0} & \dots & \iota' \end{pmatrix},$$

where ι is a $J \times 1$ vector consisting of 1, as defined previously. The \mathbf{w}^* can then be obtained using a standard quadratic programming algorithm, with preselected λ_1 and λ_2 (Goldfarb and Idnani, 1983). Note that the coherence penalty term is

$$\lambda_1\mathbf{w}'\mathbf{C}_1\mathbf{w} = \lambda_1 \sum_{x=1}^N \sum_{j=1}^{J_n} w_{xj}^2,$$

which reduces the weights for the J_n age-incoherent models with the influence of penalty λ_1 , and the smoothness penalty term is

$$\lambda_2\mathbf{w}'\mathbf{C}_2\mathbf{w} = \lambda_2 \sum_{x=2}^N \sum_{j=1}^J (w_{x-1,j} - w_{xj})^2,$$

which increases the smoothness of assigned weights for the same model between adjacent ages for a larger λ_2 .

Example 1. Consider a simple case with $N = 2$ ages, $J_c = 1$ and $J_n = 1$ models, we have that

$$\Omega_e = \begin{pmatrix} \Sigma_{e_1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{e_2} \end{pmatrix} = \begin{pmatrix} \sigma_{1,1}^2 & \sigma_{1,12} & 0 & 0 \\ \sigma_{1,12} & \sigma_{1,2}^2 & 0 & 0 \\ 0 & 0 & \sigma_{2,1}^2 & \sigma_{2,12} \\ 0 & 0 & \sigma_{2,12} & \sigma_{2,2}^2 \end{pmatrix},$$

$$C_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } C_2 = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}.$$

The optimization problem stated in (4.1) then reduces to

$$\begin{aligned} \min_w \quad & w_{1,1}^2 \sigma_{1,1}^2 + w_{1,2}^2 \sigma_{1,2}^2 + w_{2,1}^2 \sigma_{2,1}^2 + w_{2,2}^2 \sigma_{2,2}^2 + 2w_{1,1}w_{1,2}\sigma_{1,12} + 2w_{2,1}w_{2,2}\sigma_{2,12} \\ & + \lambda_1(w_{1,2}^2 + w_{2,2}^2) + \lambda_2[(w_{1,1} - w_{2,1})^2 + (w_{1,2} - w_{2,2})^2] \\ \text{subject to} \quad & w_{1,1} + w_{1,2} = 1 \\ & w_{2,1} + w_{2,2} = 1 \\ & w_{1,1}, w_{1,2}, w_{2,1}, w_{2,2} \geq 0. \end{aligned}$$

With a larger (smaller) λ_1 , weights of the second model (assumed age-incoherent) will be smaller (larger). Similarly, with a larger (smaller) λ_2 , the $\hat{w}_{1,1}$ and $\hat{w}_{1,2}$ will change to $\hat{w}_{2,1}$ and $\hat{w}_{2,2}$ more (less) smoothly.

Denote that the h -step-ahead forecast of the age x 's logged mortality rate by the j th model as $\hat{y}_{x,j,T+h}$, the forecast of MA is therefore

$$\hat{y}_{x,M,T+h} = \sum_{j=1}^J \hat{w}_{x,j} \hat{y}_{x,j,T+h}.$$

We now demonstrate that this forecast also achieves the desirable age coherence in an asymptotic fashion.

Theorem 1. Assume that the estimated long-run mean of mortality improvements is not increasing with h for all models considered in the MA approach, those estimators of age-coherent models are all asymptotically consistent, h and T go to infinity at the same rate, and the penalty λ_1 goes large with T^2 . Forecasts of the MA approach using estimated weights as the solution to (4.1) are asymptotically age coherent.

Proof. See Supplement Materials A. □

Finally, the selection of λ_1 and λ_2 can still be performed using the hold-out-sample strategy as explained in Section 3.4. The detailed steps to forecast logged mortality rates using the proposed MA approach are listed below.

1. With the training period $[1, \dots, T/2]$, fit each of the J models that are included in the ensemble;
2. Collect out-of-sample forecast errors over $[T/2 + 1, \dots, 3T/4]$, and calculate the sample estimate $\hat{\Sigma}_e$;
3. Perform a grid search of λ_1 and λ_2 : for each candidate, obtain estimates of weights as in (4.1) using $\hat{\Sigma}_e$ obtained in Step 2, fit all individual models over the full training sample $[1, \dots, 3T/4]$, produce the forecasts over the test period $[3T/4 + 1, \dots, T]$, and calculate the corresponding

RMSFE of the MA model using the estimated weights and forecasts of individual models as in (3.10);

4. Select the optimal λ_1 and λ_2 as that minimize the RMSFE;
5. Repeat steps 1-2 to obtain final estimated weights, with selected tuning parameters over training period $[1, \dots, 3T/4]$ and $\widehat{\Sigma}_e$ estimated with out-of-sample forecast errors over $[3T/4 + 1, \dots, T]$; and
6. Conduct the out-of-sample forecasting ($[T + 1, \dots, T + h]$) with each of the J models, and the forecast rate of MA is their weighted summation using weights estimated at step 5.

To study the uncertainties of the forecast, we may follow the usual bootstrap method as explained in Chang and Shi (2021). Simply speaking, in-sample errors can be bootstrapped to produce replicates, which are then fitted following the procedure explained above to forecast new rates. The 2.5th and 97.5th percentiles out of those forecasts can then construct the 95th prediction interval (PI).

Remark 7. The advantages of the proposed MA approach are fourfold. First, different from a “naive” simple average approach, the weights are selected in a way that follows the spirit of GMVP and thus honors the diversity benefits. Second, the employed objective function effectively resolves two major issues of the GMVP solution. In particular, as forecast errors consist of $y_{x,T+h} - \widehat{y}_{xj,T+h}$, the squared loss (variance) to be minimized is composed of out-of-sample MSFE, which considers the parameter risk that GMVP neglects. In addition, the imposed penalties ensure asymptotic age coherence in the long run and simultaneously reduce the risk of overfitting in the short run. Specifically, the employed hold-out-sample strategy to choose a smoothness penalty could effectively eliminate the undesirable abrupt changes in forecasts rates (a signal of overfitting) from one age to another. This may also improve the consistency of weights estimation, since potential biasness can be effectively reduced by minimizing RMSFE over the test sample. Thus, a biasness adjustment factor is not needed in (4.1). Third, the imposed non-negativity constraint improves the interpretability. In short, the proposed MA approach considers both the in-sample fitting and out-of-sample forecasting accuracy, as well as the interpretability issue. Also, the modeling mechanic of our approach is transparent. As a supervised learning technique, this may be considered an advantage of our approach over other competitive but more “black-box” machine learning models, such as the neural networks. Fourth, the curse of high dimensionality will be avoided by working with mortality rates of individual ages. To see this, although Ω_e in (4.1) is high-dimensional, it is a sparse matrix with only nonzeros on the diagonal blocks. Despite the large value of N , each Σ_{e_x} only considers J models ($J \times J$). Since it usually holds that $J \ll T$, Σ_{e_x} can be reliably estimated.

5. Empirical out-of-sample forecasting results

In this study, we focus on the mortality data of ten European countries investigated in the seminal work of Li and Lee (2005): Austria, Denmark, the United Kingdom (UK), Finland, France, Italy, Netherlands, Norway, Spain, and Switzerland. All data are obtained from the Human Mortality Database (2019). Following Booth *et al.* (2006), we choose an opportune range of data, 1950–2016, to have a reliable, complete data set. Similarly, ages from 0 to 110 (the upper limit in Human Mortality Database, 2019) years are included in the sample, where data at older ages (100 and above) are grouped to avoid potentially erratic rates therein. The crude total (uni-sex) mortality rates are studied.

To illustrate the powerfulness of our proposed model, we consider the training sample of 1950–2006 and forecast the mortality rates for 2007–2016 as the baseline results. Out-of-sample forecasts of the eleven investigated models: STAR, LC-H, LC-G, STAR-H, STAR-G, LC, SVAR, APC, HU, RH, and MA are presented and compared. Next, we conduct three sets of relevant sensitivity analyses, including robustness check on the forecasting horizon, sample period, and age groups. Additional analyses with

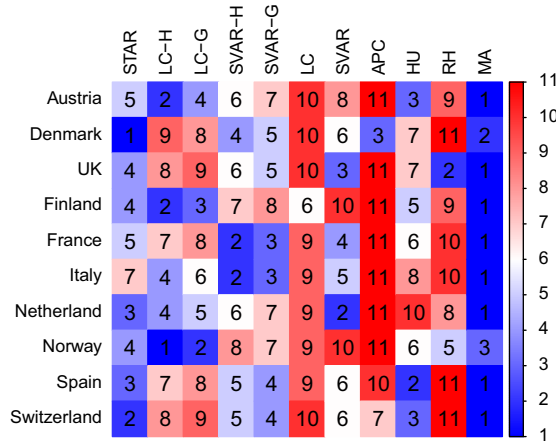


Figure 2. Heatmap of model forecasting accuracy rankings under the baseline scenario.

the model confidence set (MCS) and Bayesian modeling average (BMA) are conducted at the end of this section.²

5.1. Baseline results

To compare the forecasting performance across all models, we follow existing studies such as Li and Lu (2017) and employ the RMSFE examining all age groups and forecasting steps:

$$RMSFE = \sqrt{\frac{1}{101 \times h} \sum_{i=1}^h \sum_{x=0}^{100} (y_{x,T+i} - \hat{y}_{x,T+i})^2}$$

In the baseline case, we select $h = 10$, such that the training sample of 1950–2006 is fitted, and the out-of-sample forecasts are produced over 2007–2016 for each population. The ten individual models: STAR, LC-H, LC-G, STAR-H, STAR-G, LC, SVAR, APC, HU, and RH are first estimated, following necessary tuning parameter selection procedures described in Section 3.4. Weights of those models are then chosen following steps listed in Section 4.2. All forecasts of each of the eleven models are then collected to compute their corresponding RMSFE. This procedure is executed for all the ten examined populations.

All specific RMSFEs are presented in Table 1. To facilitate the comparison, the ranking of each model is displayed in Figure 2, and a black color indicates a higher place. On average, it can be seen that the five age-coherent models beats the five age-incoherent counterparts. More importantly, contrasting the six age-coherent sets of forecasts, those of MA outperform individual models in eight out of the ten populations and are the second or third best for the rest. This strongly supports the dominant forecasting accuracy using our proposed MA technique. As for the specific metrics, on average, the RMSFE of MA is 0.1707, which is almost 25% smaller than that of LC (0.2360). A summary of those baseline RMSFEs can be found in Table 2, which is presented when the MCS analysis is performed.

We now compare the individual models included in the ensemble. On average, STAR produces the lowest RMSFE, and results of the other four age-coherent models are relatively close to each other. Specifically, the average ranking of STAR is 3.6, and those of LC-H, SVAR-H, and SVAR-G are close to 5, whereas LC-G ranks somewhat lower at 7. Further, RMSFEs of SVAR-H and SVAR-G are much similar throughout all cases. Those of LC-H and LC-G are also not too far away from each other, although

²We have also conducted a simulation study using the age-coherent models only, the results of which are consistent with the baseline observations and available upon request.

Table 2. Model confidence sets: selected models.

	STAR	LC-H	LC-G	SVAR-H	SVAR-G	LC	SVAR	APC	HU	RH
Austria		X	X							
Denmark	X								X	
UK				X	X					
Finland		X	X							
France	X	X	X	X	X		X			
Italy	X	X	X	X	X		X		X	
Netherlands	X	X	X						X	
Norway		X	X						X	
Spain	X				X					
Switzerland		X	X							

the difference is visually larger than those in SVAR-H and SVAR-G. For the five age-incoherent models, SVAR and HU result in similar overall performance, whereas LC, APC, and RH rank at the end of all considered models.

5.2. Sensitivity analyses

To check the robustness of baseline results, we consider three sets of sensitivity analyses individually:

1. Forecast steps are increased to $h = 16$, as examined in Gao and Shi (2021);
2. The starting year is truncated to 1960; and
3. The age range is reduced to 0–89.

The ranking of all models examined in each set is plotted in Figure 3(a)–(c), and corresponding RMSFEs are summarized in Table 2 in the Supplement Materials. The ranking plot suggests that our proposed MA approach works consistently well by altering individually the forecasting horizon, weights construction, sample period, and age groups. It almost uniformly ranks among the top 3 models and produces the smallest average RMSFE in all scenarios. Nevertheless, it is worth noting that STAR model ranks the second best in all cases, whereas LC, APC, and RH are the three least preferred models. The overall performance of the five age-coherent models is uniformly better than that of the five age-incoherent models, by averaging the RMSFEs across relevant models and populations.

In summary, we employ the popular criterion RMSFE and demonstrate that the proposed MA approach can overall improve the forecasting performance of all models included in the ensemble. This conclusion is robust when various sensitivity analyses are conducted with individual adjustment being made.

5.3. Comparison with the model confidence set and Bayesian modeling average approaches

In their seminal work, Hansen *et al.* (2011) propose an MCS approach that chooses the subset of superior models out of a range of candidate. Those superior models are identified by testing if they can be assumed to present equal predictive ability at a given confidence level. Such a test may be performed for any loss functions, including the popular squared losses employed in this paper. Briefly speaking, the MCS test works using bootstrap samples sequentially, by eliminating the worst model at each stage. The procedure stops until the null hypothesis of equal predictive ability of the remaining model cannot be rejected.

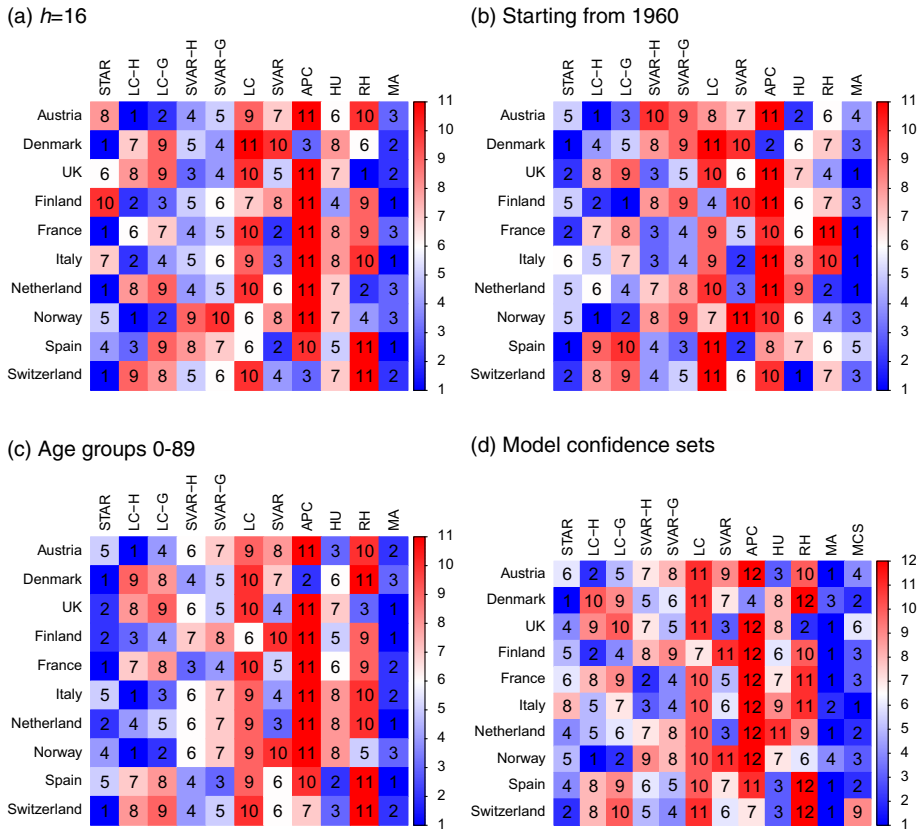


Figure 3. Heatmaps of model forecasting accuracy rankings under additional scenarios.

To employ the MCS for mortality forecasting, we produce squared forecast errors over the last 1/4 of the training period (i.e., 1993–2006) for each individual model. Similar to Kessy *et al.* (2021), we choose a confidence level of 10%, and the selected models for each population are presented in Table 2. It is interesting to note that LC-H and LC-G are the two mostly selected model in the MCS, followed by STAR. On the other hand, LC, APC, and RH models are not picked in any of the ten populations.

Assuming the equal weights for all ages and selected models, we construct this new MA approach via the MCS method. Together with the baseline results, the RMSFEs are summarized in Table 3, and individual ranks are illustrated in Figure 3(d). Clearly, our proposed MA approach beats the MCS according to all the five metrics presented in Table 3. For RMSFE of individual populations, our approach results in more accurate forecasts in seven out of ten countries. Additionally, given that MCS significantly reduces the number of selected models in the ensemble, we conclude that our approach is preferred for both the achieved diversity and resulted forecasting accuracy.

Alternative to the GMVP and MCS, another potentially useful model averaging approach is the Bayesian modeling average (BMA). Specifically, weights of the BMA are essentially posterior model probabilities. Detailed calculations of those weights can be found in Raftery *et al.* (1997), and a full BMA usually requires a computational intensive approach, such as the Markov chain Monte Carlo. Fortunately, in practice, seminal works including Bates and Granger (1969) have suggested that weights of a full BMA can be well approximated by employing a much simpler Akaike or Bayesian information criterion. Applications of such an approximation to mortality data can be found in Shang (2012). However, Wagenmakers and Farrell (2004) point out that if one model in the ensemble is dominant according to the used information criterion, the posterior probability of a BMA may tend to be close to

Table 3. Summary of all investigated models including the MCS.

	Mean	Std. Dev.	Median	Q ₁	Q ₃
STAR	0.1834	0.0642	0.1777	0.1389	0.2455
LC-H	0.1978	0.0680	0.1804	0.1434	0.2396
LC-G	0.2015	0.0665	0.1870	0.1462	0.2426
SVAR-H	0.1969	0.0814	0.1798	0.1302	0.2727
SVAR-G	0.1971	0.0812	0.1800	0.1311	0.2727
LC	0.2360	0.0614	0.2381	0.1819	0.2834
SVAR	0.1989	0.0871	0.1810	0.1286	0.2760
APC	0.3176	0.0628	0.3003	0.2709	0.3671
HU	0.1996	0.0702	0.1798	0.1513	0.2491
RH	0.2471	0.0807	0.2532	0.1936	0.3027
MA	0.1708	0.0728	0.1560	0.1088	0.2368
MCS	0.1802	0.0711	0.1633	0.1200	0.2417
BMA	0.1792	0.0677	0.1630	0.1255	0.2373

Note: bold numbers are the smallest quantity for each statistic across the twelve models.

one, also known as the “winner-take-all” issue. Consequently, we consider an alternative BMA technique discussed in Kontis *et al.* (2017), and age-specific weights are approximated by

$$\hat{w}_{x,j}^b \approx \frac{e^{-0.5|PB_{x,j}|}}{\sum_{j=1}^J e^{-0.5|PB_{x,j}|}},$$

where $PB_{x,j} = \sum_{i=1}^h (\hat{y}_{x,j,T+i} - y_{x,j,T+i})/h$ is the projection bias for logged mortality rates of age x produced by model j . Intuitively, due to the closeness of mortality forecasts, no $PB_{x,j}$ could be universally small or close to 0 to cause a dominant weight in a BMA mortality model.

Using the age-specific BMA weights of all models, the RMSFE results are summarized in the last row of Table 3. Overall, the results of BMA outperform all but those of our MA approach. Two conclusions could be drawn. First, BMA forecasts could serve as an additional robustness check to our baseline results. The altered factor is the weighting strategy in the model averaging. Second, the overall outperformance of MA over BMA supports the superior effectiveness of our proposed approach. Despite that no dominant weights may exist in the employed BMA, such an approach is “deterministic” and formula based. Thus, it cannot cope with the coherent and smoothness penalties. Consequently, the desirable asymptotic age coherence will be lost for the resulting forecasts, and abrupt changes in mortality forecasts are inevitable, even for neighboring age groups. This might also contribute to its observed inferior forecasting performance to our MA approach.

6. A case study and illustration on annuity pricing

In this section, we comprehensively present a case study by investigating the Italian population using the examined models. Curves presented in Figure 1 have visualized the dynamics of associated mortality rates. To illustrate the practical usefulness of the proposed MA approach, we demonstrate the age-specific forecasting results and its long-term forecasting results. We also validate the proposed objective function of the MA method by contrasting relevant results with those of BMA. An application on annuity pricing can be found in end of this section.

6.1. Forecasting results

Instead of examining a single metric, we employ the RMSFE over the age dimension to compare forecasting performance of each model. Consistent with our baseline setting discussed in Section 5.1, the

Table 4. Summary of RMSFE over age groups for the Italian population.

	Mean	Std. Dev.	Q ₁	Q ₃
STAR	0.1075	0.0859	0.0273	0.1583
LC-H	0.1074	0.0608	0.0460	0.1612
LC-G	0.1106	0.0703	0.0447	0.1580
SVAR-H	0.0931	0.0792	0.0292	0.1518
SVAR-G	0.0948	0.0808	0.0294	0.1557
LC	0.1911	0.3245	0.0553	0.2503
SVAR	0.0930	0.0821	0.0345	0.1567
APC	0.2945	0.2820	0.0650	0.4802
HU	0.1158	0.1002	0.0365	0.1903
RH	0.2003	0.1271	0.1399	0.2425
MA	0.0859	0.0604	0.0260	0.1416
BMA	0.0958	0.0783	0.0287	0.1423

Note: bold numbers are the smallest quantity for each statistic across all models.

training period of 1950–2006 is employed to fit in-sample estimates, and out-of-sample forecasts are calculated for the test period of 2007–2016.³ At each of the 101 one-year age groups (0–100), we calculate the age-specific RMSFE as

$$RMSFE_x = \sqrt{\frac{1}{10} \sum_{h=1}^{10} (y_{x,T+h} - \hat{y}_{x,T+h})^2}.$$

The results are summarized in Table 4.

In Table 4, it can be seen that the proposed MA approach still leads to the smallest average RMSFE over ages, with the narrowest variation. Specifically, the mean (standard deviation) of MA's RMSFE is 0.0859 (0.0604), over 50% (80%) smaller than that of LC. Thus, for the Italian population, MA can improve the forecasting accuracy of individuals models in the ensemble.

We explore out-of-sample interval estimates in Figure 4, where the data investigated are the logged mortality rates averaged out of all ages. The results of MA and SVAR-H, the top two best models as evidenced in Table 4 are contrasted, and the 95% PIs of MA and SVAR-H are constructed using bootstrap replicates as outlined in Section 4.2. Clearly, compared to the mean estimates of SVAR-H, those of MA are overall closer to the true values spanning 2007–2016. This is consistent with our observations in Table 4. Also, although both PIs manage to cover the true values at all forecasting steps, the width of MA's PI is uniformly narrower than that of SVAR-H. This implies that MA is potentially more efficient than the SVAR-H model. Such a finding is consistent with our discussion in Section 3. Specially, MA with weights estimated via the forecast error optimization may improve the forecasting uncertainty over all individual models.

Finally, we display the long-term forecasts as of 2051, a 35-step-ahead horizon in Figure 5. The full data of 1950–2016 are fitted to provide the estimates of parameters, and forecasts of MA and LC are generated and contrasted. Compared to the true values of 2016, forecasts of LC in 2051 demonstrate little improvements on both the very old ages and “accidental hump” (ages 20–25). Such results are against the observations in Figure 1(a), which suggests significant improvements even for those ages. In contrast, due to its age-coherence feature, forecasts of MA show more consistent improvements over all

³Note that the data range of 1950–2016 is consistent with our setting in Section 5. This selection is identical to those adopted in related works, such as Guibert *et al.* (2019), Gao and Shi (2021) and Li and Shi (2021). The relevant forecasting and modeling results in those papers and ours are therefore directly comparable. It is worth mentioning that the data availability in Human Mortality Database has been extended to 2019, at the time of writing this paper. Future studies are advised to adopt the up-to-date mortality rates for novelty.

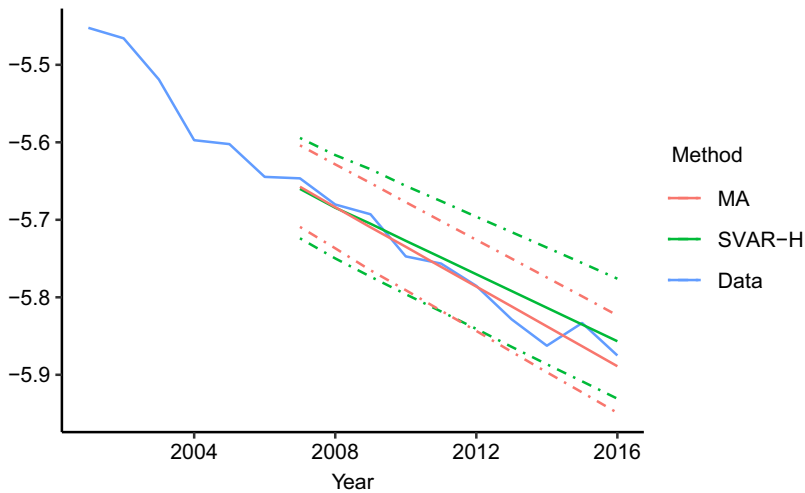


Figure 4. Forecast versus actual Italian log mortality rates: 2001–2016.

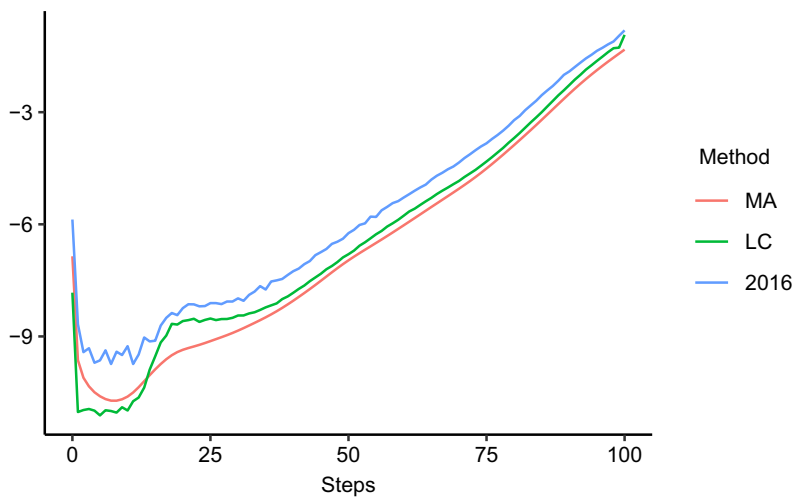


Figure 5. Actual and Long-term forecast Italian log mortality rates: 2016 vs 2051.

ages. Also, even for a considerably long period of 35 years, no abrupt changes are observed at any ages for the MA model. This supports the effectiveness of the smoothness penalty of λ_2 employed in (4.1).

6.2. Validation of the proposed weighting method

In Figure 1(b), we present the $RMSFE_x$ of age-incoherent models over the out-of-sample period 1993–2006. Those corresponding forecast errors are the core part in the loss to be minimized in (4.1). Consistent with this, we present $RMSFE_x$ of age-coherent models over the same period in Figure 6. To facilitate the comparison, $RMSFE_x$ of five age-incoherent models are grouped together, and their average value is also plotted (denoted as NC). Summarizing from the two figures, there are two major findings. First, as stated in Section 1, no single model can uniformly outperform the rest across all age groups. This validates the necessity of age-specific weights as employed in our MA approach. Second, the $RMSFE_x$

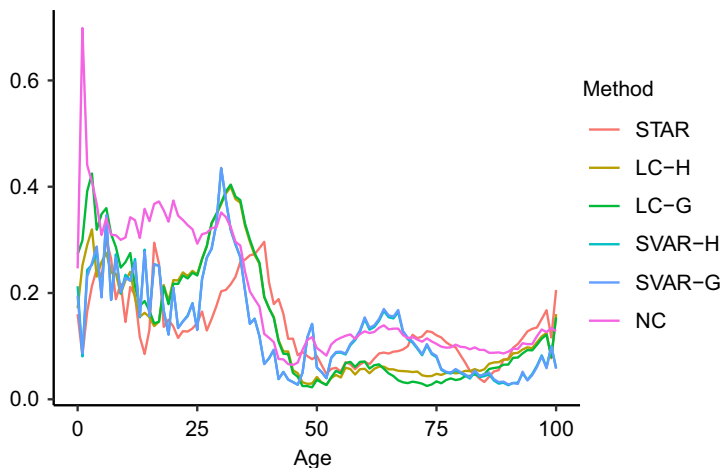


Figure 6. $RMSFE_x$ of fitted models for Italian population: 1993–2006.

of age-incoherent model is overall higher than that of age-coherent models. Specifically, age-incoherent $RMSFE_x$ is quite high over young ages 0–25 and is still among the worst for ages beyond 50. This suggests that with the imposed coherent penalty λ_1 in (4.1), the estimated weights of age-incoherent models may be much lower than those of age-coherent models.

In Figure 7(a), we plot the final estimated weights of the MA model. The dynamics are much consistent with our findings in Figure 6. First, the total weights of age-incoherent models are small, due to their unfavorable $RMSFE_x$ across ages discussed above. Also, fitted weights demonstrate much consistent patterns as their $RMSFE_x$. For instance, the weights of SVAR-H and SVAR-G models are increasing over the oldest age groups, for their smallest $RMSFE_x$ at those ages. Additionally, all weights vary smoothly across ages, due to the impact of smoothness penalty λ_2 . More importantly, it is verified that no single model dominates in the identified age-specific weights, which supports the diversity realized by our MA approach. Last but not least, since SVAR-H and SVAR-G models produce much similar $RMSFE_x$ at all ages, their corresponding weights are close to each other.

We now contrast the MA results to those of BMA. Recall that in Section 5.3, an issue is identified for BMA, such that abrupt changes are unavoidable, due to the formula-based nature of BMA. To see this, we plot estimated weights of BMA in Figure 7(a). Clearly, those weights are identical to the proportions of $RMSFE_x$ of an individual model taking in the total $RMSFE_x$ at the same age, as shown in Figures 1(b) and 6. Consequently, rough dynamics are observed for all fitted weights, and sharp changes are observed in multiple ages. For instance, the weight of LC model in BMA drops from 0.1 to 0.05 from age 0 to 1, and then bounces back to 0.09 for age 2. Further, despite their overall unfavorable performance, age-incoherent models are still assigned considerably large weights. Averaging over all ages, weights of each model are close to 0.1, the value of which would have been assumed for a simple average method. As for the forecasting accuracy, the $RMSFE_x$ estimated by the BMA approach is reported in Table 4, the results are outperformed by the proposed MA counterparts.

In conclusion, to realize the coherent and smoothness penalties, a squared loss on weights will need to be implemented in the objective function. Consequently, the GMVP structure as adopted in (4.1) is a simple but effective approach, compared to other alternatives such as the BMA.⁴ Besides, the incapability to incorporate those penalties will be faced by a novel approach proposed in Kessy *et al.* (2021). Although this stacked regression employ an Elastic NET (ENET) loss, the inclusion of the smoothness

⁴Note that if the simple formula approximation is not employed, a BMA approach may incorporate the penalties. However, the computational cost would be much higher than our MA counterpart, due to the complexity in the associated Markov Chain Monte Carlo algorithm.

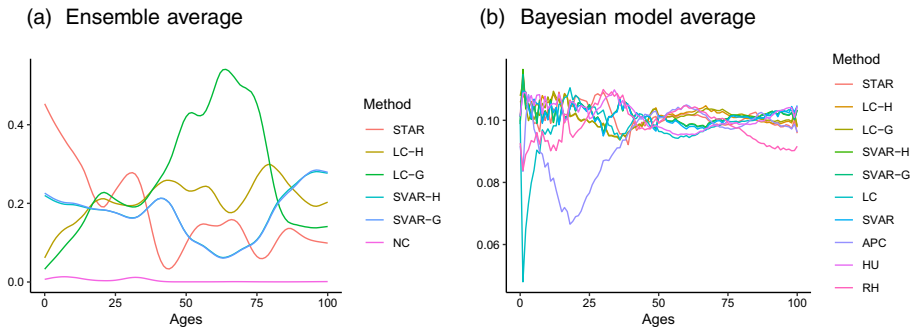


Figure 7. Fitted weights of the ensemble average and Bayesian model average approaches.

penalty is not straightforward. This is due to the fact that the smoothness penalty is not a standard L_2 -type loss, and a more complicated computational approach will need to be implemented (Chang and Shi, 2021). The imposed equality and inequality constraints will further increase the computational cost.

6.3. A financial application: Fixed-term annuity pricing

To demonstrate the practical usefulness of the MA approach, a financial application is presented in this section. Following Shi (2022b), we consider the pricing of fixed-term annuities using the MA technique. This product has attracted a growing number of policyholders world-widely, especially those planning for their retirements. Comparing to the lifetime annuities, fixed-term products pay a predetermined and guaranteed income of higher level with deferrable options (Shang and Haberman, 2017). A cohort approach, as adopted in Shi (2022b), is employed here to price annuities. To be consistent with our previous analyses, the maximal survival age is limited to 100.

The specific pricing scheme is explained below. First, the τ year survival probability of a person aged x at $t = 0$ is

$${}_{\tau}p_x = \prod_{j=1}^{\tau} {}_1p_{x+j-1},$$

where ${}_{\tau}p_{x+j-1} = e^{-m_{x+j-1}}$ and m_{x+j-1} can be obtained from mortality forecasts. Hence, an annuity with a T -year maturity and written for an x -year-old person including a benefit \$1 per year and conditional on the survival is priced as

$$a_x^T(m_{x,1:T}) = \sum_{\tau=1}^T P_B(0, \tau) E(I(T_x > \tau) | m_{x,1:T}),$$

where $I(\cdot)$ is the indication function, T_x is the survival time and $P_B(0, \tau)$ is the price of a τ -year bond with interest rate set to the yield of annuity. Since $E(I(T_x > \tau) | m_{x,1:T}) = {}_{\tau}p_x(m_{x,1:T})$, the fixed-term annuity price is fully determined by the underlying yield and mortality rates. Thus, to accurately price the product, it is critical to providing precise forecasts of $m_{x,1:T}$, potentially over a long period. That is, the mortality experiences of policyholders need to be accurately projected to minimize the mis-pricing risks.

For illustration purpose, we employ up to 35-step-ahead forecasts (from 2017 to 2051) of the proposed MA model. The estimated prices of fixed-term annuities as of 2016 are presented in Table 5. The calculation is conducted for the Italian population up to the 30-year-maturity, starting from age 65. We follow Shi (2022b) and present results of the four ages: 65, 70, 75 and 80. Both point and 95% interval estimates are reported. For simplicity, a constant interest rate of 3% is assumed in all case. Thus, the only source of uncertainty is limited to the precision of mortality forecasting. Evidenced in Figure 4, the

Table 5. *Predicted fix-term annuity prices for the Italian population.*

Age	Measure	$T = 5$	$T = 10$	$T = 15$	$T = 20$	$T = 25$	$T = 30$
65	Mean	4.530	8.408	11.712	14.502	16.813	18.663
	LB	4.528 (−0.04%)	8.403 (−0.06%)	11.699 (−0.12%)	14.474 (−0.19%)	16.705 (−0.64%)	18.436 (−1.22%)
	UB	4.532 (0.04%)	8.413 (0.06%)	11.725 (0.11%)	14.526 (0.17%)	16.890 (0.46%)	18.844 (0.97%)
70	Mean	4.499	8.330	11.559	14.229	16.360	17.940
	LB	4.495 (−0.08%)	8.318 (−0.14%)	11.535 (−0.20%)	14.136 (−0.65%)	16.150 (−1.28%)	17.589 (−1.96%)
	UB	4.503 (0.08%)	8.341 (0.13%)	11.582 (0.20%)	14.300 (0.50%)	16.534 (1.06%)	18.247 (1.71%)
75	Mean	4.440	8.177	11.259	13.714	15.533	NA
	LB	4.433 (−0.17%)	8.159 (−0.22%)	11.182 (−0.69%)	13.529 (−1.35%)	15.200 (−2.15%)	NA
	UB	4.448 (0.16%)	8.195 (0.22%)	11.322 (0.56%)	13.872 (1.16%)	15.823 (1.86%)	NA
80	Mean	4.323	7.879	10.705	12.800	NA	NA
	LB	4.310 (−0.29%)	7.824 (−0.70%)	10.554 (−1.42%)	12.499 (−2.35%)	NA	NA
	UB	4.334 (0.27%)	7.932 (0.67%)	10.841 (1.27%)	13.069 (2.10%)	NA	NA

Note: this table displays the forecast fix-term annuity price for the average population as of 2016. The forecast mortality rates range from 2017 to 2051. LB and UB stand for the 2.5th and 97.5th percentiles of the prediction interval, respectively. Mean is the point/mean forecast price. T is the maturity term. Value in bracket is the percentage difference compared to the forecast mean annuity price. We only consider contracts with maturity so that age + maturity ≤ 100 .

proposed MA technique is more efficient in providing PIs. This supports the small widths observed in Table 5. As argued in Fung *et al.* (2015), underpricing of annuities as small as by 0.1% can lead to dramatic shortfall in reserving with a large portfolio. Consequently, precisely and efficiently determine the uncertainty of premium rate is critical for insurers to optimizing their reserves, such that the associated ruin probability is minimized.

Specifically, a prospective lifetable constructed using the mortality forecast by the MA approach may largely reduce the longevity risks. As demonstrated in Figure 7, an age-incoherent model like LC may underestimate mortality improvements of old ages. For pension products, a corresponding prospective lifetable will then result in an underestimated reserve and thus increase the ruin probability. Using an age-coherent model like the MA approach to produce such a prospective lifetable will help address this issue.

7. Concluding remarks

In this study, we propose an effective ensemble or model averaging (MA) approach to study and forecast logged mortality rates. Three key results can be drawn from our research. First, the proposed MA model is effective in mortality forecasting by improving accuracy of all individual models in the ensemble. Altogether, we examine five age-coherent candidates: spatial temporal autoregressive (STAR) model proposed in Li and Lu (2017), Lee–Carter with hyperbolic (LC-H) and geometric (LC-G) decays studied in Gao and Shi (2021), sparse VAR with hyperbolic (SVAR-H) decay examined in Li and Shi (2021) and SVAR model with geometric decays (SVAR-G). Five age-incoherent models are also considered, including the LC model (Lee and Carter, 1992), SVAR model (Guibert *et al.*, 2019), age-period-cohort model (Cairns *et al.*, 2009), functional demographic model (Hyndman and Ullah, 2007) and Renshaw-Haberman model (Renshaw and Haberman, 2006). Our data include 0–100 ages of ten European populations considered in Li and Lee (2005), spanning 1950–2016. Robust results are also observed when the forecasting steps, sample period and covered ages are altered individually. As an alternative model averaging approach, we also explore the model confidence set (Hansen *et al.*, 2011) for mortality forecasting. The improvements in forecasting with our proposed MA approach are consistently demonstrated in all cases.

Second, the technical properties of MA are presented and discussed. The weights in MA are determined by optimizing the resulting variance of out-of-sample errors with non-negative constraints to improve interpretability. The considered out-of-sample error resolves the major issue of the based global minimum variance portfolio approach, such that the parameter risk is not considered. Moreover, since the weights are age specific, with a considerably large size of the ensemble space, the diversity is realized to present a potential “winner-take-all” issue. To further avoid overfitting and abrupt changes of forecast rates over adjacent ages, a smoothness penalty is imposed. Also, we employ a coherent tuning parameter to achieve the desirable age coherence asymptotically. The resulting forecasts of our MA method are then proved to be asymptotically age coherent. For its averaging nature, the MA approach can resolve the misspecification issue that each of the individual models may face. This explains its attractive forecasting performance in the empirical analyses.

Finally, MA can be a useful tool in the practical application. To demonstrate this, we use the Italian population and present a case study. In particular, interval results indicate that the MA approach may result in lower forecasting uncertainty than all individual models. Besides, via an analysis on the estimated age-specific weights, we show that the proposed MA approach is more appropriate than popular alternatives, such as the BMA. A financial application to the fixed-term annuities pricing demonstrates the usefulness in other practices. For instance, practitioners like insurers may benefit from adopting the MA technique to improve the reserving accuracy and thus reduce the ruin probability.

Apart from the improved forecasting accuracy, the proposed MA approach can shed light on two areas to consider for the actuarial practice. First, as plotted in Figure 7(a), age-specific weights illustrate the relative strength of a model for certain ages. Among the Italian population, for the workforce age

groups (30–65), the LC-type age-coherent models are preferred, whereas SVAR-type counterparts take most weights on the oldest ages (90–100). This might suggest that mortality rates of workforce groups are less prone to the cohort effects, whereas oldest ages are associated with more cross-group impacts. Consequently, decisions on risk factors, such as the birth year, may have various impacts on all future ages (some could be insignificant, if the age fall in 30–65), when pricing the premium of the associated pension products. Second, the proposed objective function presents a simple but effective approach to handle the modeling errors. Specifically, due to the small sample size, without constraints, randomness in estimation could be high to reduce the reliability of estimates. Apart from adopting the out-of-sample forecast errors, our approach employs the coherent and smoothness penalties. Those constraints are reflective of prior information to effectively reduce the influence of the randomness introduced by small samples. Implications can be made when computing the solvency capital requirement (SCR) in the European Union region. For instance, to better improve the reserving efficiency, the SCR may honor appropriate prior information that is adopted to reduce modeling error for the internal models employed by an insurer.

There are also some pathways for future research to extend the technical features of the proposed model. For instance, robust optimization technique (Bertsimas *et al.*, 2011) may be implemented when solving (4.1) to optimal weights. This may further improves the robustness of the fitted MA, as the size of out-of-sample forecast error is usually limited. Also, as pointed out in Brouhns *et al.* (2002), mortality models based on logged mortality rates frequently perform better for young ages and worse for old ages, compared to those on the original scale. The reason is that the logarithm of observed force of mortality is much more variable at older ages than at younger ages because of the much smaller absolute number of deaths at older ages. Since life insurance and pension products may be more relevant to mortality patterns in old ages, age-coherent approaches are to be explored using mortality rates at the original scale. For instance, in their seminal work, Brouhns *et al.* (2002) incorporate the LC methodology in the modeling of Poisson-distributed death counts via maximum likelihood estimation. Thus, the LC-H and LC-G approaches may be extended to this framework, to obtain the age-coherent forecasts considering the additional information of exposure to risk. VAR-type models can be developed based on the AR models with Poisson response variables (see, for example, Brandt and Williams, 2001). Such explorations remain for future works.

Acknowledgment. The authors would like to thank the Australian National University and Macquarie University for the research support. We particularly thank the Editor (Montserrat Guillen) and two anonymous referees for providing valuable and insightful comments on earlier drafts. The usual disclaimer applies.

Supplementary material. To view supplementary material for this article, please visit <https://doi.org/10.1017/asb.2022.23>

References

- Amini, S.M. and Parmeter, C.F. (2012) Comparison of model averaging techniques: Assessing growth determinants. *Journal of Applied Econometrics*, **27**(5), 870–876.
- Baechle, C., Huang, C.D., Agarwal, A., Behara, R.S. and Goo, J. (2020) Latent topic ensemble learning for hospital readmission cost optimization. *European Journal of Operational Research*, **281**(3), 517–531.
- Bates, J.M. and Granger, C.W. (1969) The combination of forecasts. *Journal of the Operational Research Society*, **20**(4), 451–468.
- Bertsimas, D., Brown, D.B. and Caramanis, C. (2011) Theory and applications of robust optimization. *SIAM Review*, **53**(3), 464–501.
- Blake, D., Cairns, A.J., Dowd, K. and Kessler, A.R. (2019) Still living with mortality: The longevity risk transfer market after one decade. *British Actuarial Journal*, **24**, 1–80.
- Booth, H., Hyndman, R., Tickle, L. and De Jong, P. (2006) Lee-Carter mortality forecasting: A multi-country comparison of variants and extensions. *Demographic Research*, **15**, 289–310.
- Bork, L., Møller, S.V. and Pedersen, T.Q. (2020) A new index of housing sentiment. *Management Science*, **66**(4), 1563–1583.
- Brandt, P.T. and Williams, J.T. (2001) A linear poisson autoregressive model: The Poisson AR(p) model. *Political Analysis*, **9**(2), 164–184.
- Bravo, J.M., Ayuso, M., Holzmann, R. and Palmer, E. (2021) Addressing the life expectancy gap in pension policy. *Insurance: Mathematics and Economics*, **99**, 200–221.

- Brouhns, N., Denuit, M. and Vermunt, J.K. (2002) A poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, **31**(3), 373–393.
- Cairns, A.J., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A. and Balevich, I. (2009) A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, **13**(1), 1–35.
- Chang, L. and Shi, Y. (2021) Mortality forecasting with a spatially penalized smoothed VAR model. *ASTIN Bulletin: The Journal of the IAA*, **51**(1), 161–189.
- du Jardin, P. (2021) Forecasting corporate failure using ensemble of self-organizing neural networks. *European Journal of Operational Research*, **288**(3), 869–885.
- Eicher, T.S., Papageorgiou, C. and Raftery, A.E. (2011) Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics*, **26**(1), 30–55.
- Fung, M.C., Peters, G.W. and Shevchenko, P.V. (2015) A state-space estimation of the Lee–Carter mortality model and implications for annuity pricing. arXiv preprint arXiv:1508.00322.
- Gao, G. and Shi, Y. (2021) Age-coherent extensions of the Lee–Carter model. *Scandinavian Actuarial Journal*, **2021**(10), 998–1016.
- Genre, V., Kenny, G., Meyler, A. and Timmermann, A. (2013) Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting*, **29**(1), 108–121.
- Gill, P.E., Murray, W. and Wright, M.H. (2019) *Practical Optimization*. Philadelphia: SIAM.
- Goldfarb, D. and Idnani, A. (1983) A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, **27**(1), 1–33.
- Guibert, Q., Lopez, O. and Piette, P. (2019) Forecasting mortality rate improvements with a high-dimensional VAR. *Insurance: Mathematics and Economics*, **88**, 255–272.
- Hansen, P.R., Lunde, A. and Nason, J.M. (2011). The model confidence set. *Econometrica*, **79**(2), 453–497.
- Ho, K.-Y. and Shi, Y. (2020). Discussions on the spurious hyperbolic memory in the conditional variance and a new model. *Journal of Empirical Finance*, **55**, 83–103.
- Human Mortality Database (2019) University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). URL <http://www.mortality.org>.
- Hyndman, R. and Ullah, S. (2007) Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis*, **51**(10), 4942–4956.
- Hyndman, R.J. and Athanasopoulos, G. (2018) *Forecasting: Principles and Practice*. OTexts.
- Kessy, S., Sherris, M., Villegas, A. and Ziveyi, J. (2021) Mortality forecasting using stacked regression ensembles. Available at SSRN 3823511.
- Kleijn, R. and Van Dijk, H.K. (2006) Bayes model averaging of cyclical decompositions in economic time series. *Journal of Applied Econometrics*, **21**(2), 191–212.
- Kontis, V., Bennett, J.E., Mathers, C.D., Li, G., Foreman, K. and Ezzati, M. (2017) Future life expectancy in 35 industrialised countries: Projections with a Bayesian model ensemble. *The Lancet*, **389**(10076), 1323–1335.
- Lee, R.D. and Carter, L.R. (1992) Modeling and forecasting US mortality. *Journal of the American Statistical Association*, **87**(419), 659–671.
- Lessmann, S., Sung, M.-C., Johnson, J.E. and Ma, T. (2012) A new methodology for generating and combining statistical forecasting models to enhance competitive event prediction. *European Journal of Operational Research*, **218**(1), 163–174.
- Ley, E. and Steel, M.F. (2009) On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, **24**, 651–674.
- Li, H. and Lu, Y. (2017) Coherent forecasting of mortality rates: A sparse vector-autoregression approach. *ASTIN Bulletin: The Journal of the IAA*, **47**(2), 563–600.
- Li, H. and Shi, Y. (2021) Mortality forecasting with an age-coherent sparse VAR model. *Risks*, **9**(2), 35.
- Li, N. and Lee, R. (2005) Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, **42**(3), 575–594.
- Li, N., Lee, R. and Gerland, P. (2013) Extending the Lee-Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography*, **50**(6), 2037–2051.
- Markowitz, H. (1952) Portfolio selection. *The Journal of Finance*, **7**(1), 77–91.
- Mirestean, A. and Tsangarides, C.G. (2016) Growth determinants revisited using limited-information Bayesian model averaging. *Journal of Applied Econometrics*, **31**(1), 106–132.
- Raftery, A.E., Madigan, D. and Hoeting, J.A. (1997) Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92**(437), 179–191.
- Renshaw, A.E. and Haberman, S. (2006) A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, **38**(3), 556–570.
- Shang, H. and Haberman, S. (2017) Grouped multivariate and functional time series forecasting: An application to annuity pricing. *Insurance: Mathematics and Economics*, **75**, 166–179.
- Shang, H.L. (2012). Point and interval forecasts of age-specific life expectancies: A model averaging approach. *Demographic Research*, **27**, 593–644.
- Shi, Y. (2022a). *Coherent mortality forecasting with a model averaging approach: Evidence from global populations*. Working paper.
- Shi, Y. (2022b). Forecasting mortality rates with the penalized exponential smoothing state space model. *Journal of the Operational Research Society*, **73**(5), 955–968.

- Shiraya, K. and Takahashi, A. (2019) Pricing average and spread options under local-stochastic volatility jump-diffusion models. *Mathematics of Operations Research*, **44**(1), 303–333.
- Smallwood, A.D. and Norrbin, S.C. (2006) Generalized long memory processes, failure of cointegration tests and exchange rate dynamics. *Journal of Applied Econometrics*, **21**(4), 409–417.
- Trefethen, L.N. and Bau, D. (1997) *Numerical Linear Algebra*. Vol. **50**. Philadelphia: SIAM.
- Wagenmakers, E.-J. and Farrell, S. (2004) AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, **11**(1), 192–196.