Forecasting Mortality Rates with the Adaptive Spatial Temporal Autoregressive Model

Yanlin Shi*

Department of Actuarial Studies and Business Analytics, Macquarie University, Sydney, NSW, Australia

Abstract

Accurate forecasts of mortality rates are essential to various demographic research like population projection, and the pricing of insurance products such as pensions and annuities. Recent studies have considered a spatial-temporal autoregressive (STAR) model for the mortality surface, where mortality rates of each age depend on the historical values of itself (temporarily) and the neighboring cohorts ages (spatiality). This model has sound statistical properties including co-integrated dependent variables and existence of closed-form solutions. Despite its improved forecasting performance over the famous Lee-Carter (LC) model, the constraint that only the effects of the same and neighboring cohorts are significant can be too restrictive. In this study, we adopt a data-driven adaptive weighted structure and propose the adaptive STAR (ASTAR) model. Retaining all desirable features of the STAR, our model uniformly outperforms the LC and STAR counterparties for forecasting accuracy, when mortality data aged 0–100 of the United Kingdom, France, Italy, Spain and Japan over 1950–2016 are considered. Two sensitivity tests and additional simulation results also lead to robust conclusions. The proposed ASTAR model is therefore a widely useful tool in modelling and forecasting mortality rates in other contents, and may be extensible to multi-population modelling.

JEL Code: C18, C32, C52, C53

Keywords: Mortality forecasting, Adaptive weights, Cohort effects, Age-coherence, Lee-Carter model

*Tel.: +61 2 9850 4750

Email address: yanlin.shi@mq.edu.au (Yanlin Shi)

1. Introduction

Life expectancies have consistently improved around the world, which are predominately caused by the mortality improvements (Giacometti et al., 2012). This has spurred serious concerns about the corresponding mortality and longevity risks. Mortality (longevity) risk describes the fact that people are surviving shorter (longer) than expected (Feng and Shi, 2018). For instance, advances in medical science, technological improvements and lifestyle changes may very likely result in higher mortality improvements, which increases the exposure to longevity risk. For the demographic research, actuate mortality forecasting is critical to practice such as population projections. For the insurance industry, mis-estimating such improvements will lead to inaccurate premium ratemaking and growing insolvency risks.

Mortality modeling and forecasting have become standard mitigation tools to reduce mortality and longevity risks. Among existing methods, one popular stream is based on the seminal work of Lee and Carter (1992), which is known as the famous Lee-Carter (LC) model. Many extensions of the LC model have been extensively discussed in the literature – for examples, see Booth et al. (2002), Renshaw and Haberman (2003), Booth et al. (2006), Girosi and King (2007), Renshaw and Haberman (2006) and Barrieu et al. (2012). Another important stream focuses on applying the high-dimensional vector-autoregressive (VAR) models, which can allow more flexible temporal modeling than the LC approach (Li and Lu, 2017; Guibert et al., 2019; Feng et al., 2020; Chang and Shi, 2020). Our research contributes to significant extensions of those VAR-type models.

There are two major concerns of all VAR-type models for mortality modeling and forecasting. In terms of statistical modelling, the issue that the number of variables (e.g. age) is usually greater than the sample size (e.g. year) is widely concerned. Consequently, there are insufficient data to fit a full VAR-type model. Hence, the first concern is an appropriate dimension-reduction approach. A recent paper of Li and Lu (2017) address this by focusing on the cohort effects and adopting a spatial temporal autoregressive (STAR) model with restrictive coefficient matrix, such that only mortality rates for neighboring ages could interact. Second, VAR-type models need to be stationary to produce meaningful estimates and forecasts. Considering that the age-wide mortality rates are non-stationary time series, Li and Lu (2017) employ an age-coherent structure to impose a powerful constraint on the coefficient matrix and impose smoothness penalties. Hence, the forecast mortality rates will be smoothed and those of neighboring age groups will not diverge in the long run. Also, the proposed STAR

model has many attractive statistical properties, such as co-integration of dependent variables and closed-form solutions.

Despite the effectiveness of this approach, some new concerns are brought up. The major issue is that the STAR considers the sparsity of the coefficient matrix in a relatively ad-hoc manner. Within such a framework, only lagged mortality rates of ages $x-2$, $x-1$, and x can affect contemporary rates of age x. However, assuming only significant effects between those neighboring younger cohorts can be too restrictive because information from other younger age groups might further contribute to the mortality forecasting. To see this, we consider the logged total mortality rates of the United Kingdom (UK) of ages 0–100 over 1950–2016. The averaged correlations between those rates of the current and younger cohorts are plotted in Figure 1 [1] Clearly, those correlations demonstrate a declining trend slower than a geometric pattern but faster than a unit-root fashion. This is known as the hyperbolic decay in econometric literature (see, for example, Bollerslev and Mikkelsen (1996), Davidson (2004), Granger and Hyung (2004), Li et al. (2015), Feng and Shi (2017) and Ho and Shi (2020), among others). Such a trend would not be sufficiently captured by a usual autoregressive structure (Baillie et al., 1996). Consequently, an alternative parametric structure to the existing ad-hoc approach is of critical interest to comprehensively describe this trend.

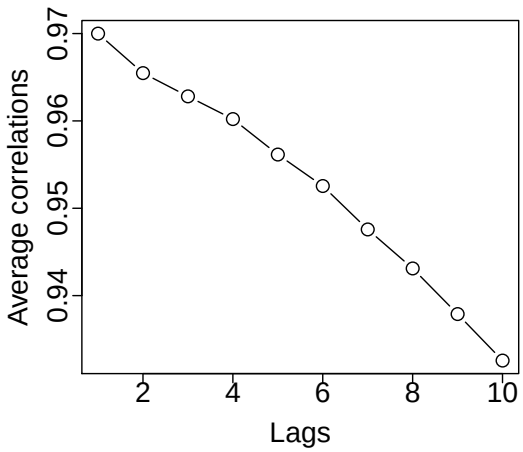


Figure 1: Averaged correlations between logged mortality rates of the current and younger (lagged) cohorts: UK

In order to address those concerns, we propose a novel adaptive spatial temporal autoregressive model (ASTAR). Instead of imposing an ad-hoc lag 2 structure, we aim to include all younger ages (i.e., from 0 to $x-1$) in modelling the mortality rate of age x.

[1] For instance, the first lag is the average correlations between logged mortality rates of ages 1–100 over 1951–2016 and those of ages 0–99 over 1950–2015.

However, such an extension is not straightforward. In the mortality modelling practice, the issue of $p >> n$ described above disables the direct fitting of even a restricted VAR-model, such that free parameters are only at the lower-triangle of the coefficient matrix. A recent study by Guibert et al. (2019) adopts a fully data-driven approach with the elastic-net (ENET) penalty estimation method. However, the price of such flexibility is the loss of age-coherence and the questionable interpretation of the estimates. For example, the empirical analysis in Guibert et al. (2019) indicated that the mortality improvements of age 45 were significantly influenced by the lagged evolution of age 95, which is counterintuitive. To reach a meaningful sparsity of the coefficient matrix, our ASTAR model utilizes the empirical trend observed in Figure 1. More specifically, we extend the STAR model by imposing an adaptive weight to each of the younger cohort, based on the corresponding empirical correlation. Such a data-driven approach comprehensively complements the ad-hoc lag 2 structure as adopted in Li and Lu (2017). Intuitively, all rates of younger ages (e.g., $x - k$) may impact those of age x, but the influences will reduce according to the empirical correlation observed between ages $x-k$ and x. Such a structure results in a more informative modelling framework, whereas the parameters to be estimated are kept minimal to retain all advantages of the STAR model, including age-coherence, closed-form solutions and low computational cost and difficulty.

To demonstrate the effectiveness of the proposed model, we provide empirical evidence using the total mortality rates of four out of the five largest European countries: UK, France, Italy and Spain.[2], as well as those of Japan for comparison. All data are sourced from Human Mortality Database (2019). Using the crude rates ranging from 1950 to 2016, we systematically compare the forecasting performance of the LC, STAR and ASTAR models on the age groups of 0–100. As measured by the root of mean squared error (RMSE), the ASTAR model consistently outperforms LC and STAR up to the 16-step-ahead forecasting horizons for all the five countries. This conclusion consistently holds when age (40–91) and training period (1960–2000) are changed individually. Additional simulation studies are further conducted, suggesting the robustness of our results. Finally, a long-term analysis further supports the argued age coherence of ASTAR and provides addition evidence of its more accurate out-of-sample forecasts than those of the LC model.

[2]The data of West Germany are only available from 1956 and are therefore not included in our research.

The contributions of this paper are mostly attributed to the introduction of the data-driven adaptive weight to the mortality forecasting. The proposed ASTAR model significantly complements the recent study of Li and Lu (2017), and largely improves the flexibility to model effects of all younger cohorts. Also, merits of the STAR are all retained in the new model, including the desirable co-integration features and existence of closed-form solution. In addition, we systematically study the forecasting performance of the ASTAR model for mortality data of five frequently investigated countries. Its superiorities over LC and STAR indicate the potential and usefulness of our method to model and forecast mortality rates in other contents. For instance, the current framework may be extended to the multi-population modelling, which has become increasingly popular among the mortality research.

The rest of this paper is organized as follows. In Section 2, we describe the LC model. The STAR and proposed ASTAR models are discussed in Section 3. We present statistical properties and estimation procedure of the ASTAR model in Section 4. In Section 5, we conduct empirical and simulation studies with the single population case. Finally, Section 7 concludes this paper.

2. The Lee–Carter model

The Lee–Carter (LC) model is proposed by combining a demographic model of mortality with time-series methods of forecasting. More specifically, central mortality rate $m_{x,t}$ at age x and year t is assumed to follow the specification shown below:

$$\ln m_{x,t} = a_x + b_x k_t + \varepsilon_{x,t} \tag{1}$$

where a_x is the average pattern of mortality by age across years, b_x is the relative speed of change at each age x, k_t is an index of the level of mortality at time t, $\varepsilon_{x,t}$ is the residual at age x and time t. As for their explanations, the random term $\varepsilon_{s,t}$ reflects a particular age-specific historical influence. Coefficients a_x are age-specific constants that describe the general shape of the age-mortality profile. Index k_t serves to capture the main temporal level of mortality. In terms of estimation, b_x and k_t are calculated by singular value decomposition (SVD) as suggested by Trefethen and Bau (1997). In order to obtain a unique solution, parameters b_x and k_t should satisfy the constraints that b_x sum to 1 and k_t sum to 0. The second constraint implies that estimates of parameters a_x are given by the averages of $\ln m_{x,t}$ over time. Additionally, k_t is adjusted by refitting to total observed deaths. This adjustment gives greater weights to ages at which deaths

are high, thereby partly counterbalancing the effect of using logarithm of rates in the LC model (Booth et al., 2006).

To forecast future mortality rates, Lee and Carter (1992) assume that a_x and b_x remain constant over time. The time factor k_t, on the other hand, is intrinsically viewed as a random walk with drift process as follows:

$$\hat{k}_t = \hat{k}_{t-1} + d + e_t \tag{2}$$

where d is the average annual change in $\hat{k}_t$, and e_t are independent and identically distributed Gaussian sequences with null mean. The expected h-step-ahead forecast log mortality rate can be approximated by:

$$\ln \hat{m}_{x,T+h} = \hat{a}_x + \hat{b}_x \left(\hat{k}_T + h \frac{(\hat{k}_T - \hat{k}_1)}{(T-1)} \right) \tag{3}$$

where T is the maximum of t.

3. The vector autoregressive (VAR) model

Different to factor model like LC, another popular stream to study and forecast mortality rates is the VAR model. However, the application of VAR model to mortality data brings up two issues. First, VAR model requires the dependent variables to be stationary. Without modification or constraints, $\ln m_{x,t}$ is clearly trending and therefore non-stationary. Second, there are more unknown parameters (p) than observations (T) in the standard VAR framework. Suppose we have N age groups, even in the simplest VAR(1) case, for each $\ln m_{x,t}$, all N lagged log mortality rates need to be included. Thus, the total number of parameters to be estimated is $p = N(N+1)$ (including N intercepts). Considering that we usually only have dozens of yearly data to work with, the $p >> NT$ issue will arise for an intermediate N such as 50.

3.1. The spatial temporal autoregressive (STAR) model

To address those two issues, Li and Lu (2017) propose a STAR model. On the temporal dimension, it considers the Granger causality and co-integration to resolve the stationarity problem. As for the age groups, STAR model utilizes the sparse spatial information to reduce the dimensionality of p. Let $y_{x,t} = \ln m_{x,t}$, this leads to the follow

specification.[3]

$$
\begin{aligned}
y_{1,t} &= m_1 + y_{1,t-1} + \varepsilon_{1,t} \\
y_{2,t} &= m_2 + (1-\alpha_2) y_{2,t-1} + \alpha_2 y_{1,t-1} + \varepsilon_{2,t} \\
y_{i,t} &= m_i + (1-\alpha_2-\beta_i) y_{i,t-1} + \alpha_i y_{i-1,t-1} + \beta_i y_{i-2,t-1} + \varepsilon_{i,t}
\end{aligned}
\tag{4}
$$

where $i = 3, 4, ..., N$, and $t = 1, 2, ..., T$. $\varepsilon_{i,t}$ is assumed to follow a multi-Gaussian distribution with $\mathbf{0}$ $(N \times 1)$ means and $\mathbf{\Sigma}$ $(N \times N)$ variance-covariance matrix. Rewritten in a VAR(1) form, we have

$$
\mathbf{Y_t} = \mathbf{M} + \mathbf{B}\mathbf{Y_{t-1}} + \boldsymbol{\varepsilon_t} \tag{5}
$$

where $\mathbf{Y_t}$=$(y_{1,t}, y_{2,t}, ..., y_{N,t})'$, $\mathbf{M}$=$(m_1, m_2, ..., m_N)'$, $\boldsymbol{\varepsilon_t}$=$(\varepsilon_{1,t}, \varepsilon_{2,t}, ..., \varepsilon_{N,t})'$, and

$$
\mathbf{B} = \begin{bmatrix}
1 & 0 & 0 & \cdots & \cdots \\
\alpha_2 & 1-\alpha_2 & 0 & \cdots & \cdots \\
\beta_3 & \alpha_3 & 1-\alpha_3-\beta_3 & 0 & \cdots \\
\vdots & \vdots & \ddots & \ddots & 0 \\
\cdots & 0 & \beta_N & \alpha_N & 1-\alpha_N-\beta_N
\end{bmatrix} \tag{6}
$$

This specification leads to attractive features. First, as shown in Li and Lu (2017), all neighboring age pairs $y_{i,t}$ and $y_{i+1,t}$ are co-integrated with order (1,-1). This successfully solves the stationarity issue and therefore leads to age-coherence. In other words, forecast rates of neighboring age groups will not diverge in the long run. Second, the total number of parameters is largely reduced to $p = 3N - 3$ and no longer greater than NT.

The forecasting is performed in an iterative fashion, where

$$
\begin{aligned}
\hat{\mathbf{Y}}_{\mathbf{t+1}} &= \hat{\mathbf{M}} + \hat{\mathbf{B}}\mathbf{Y_t} \\
\hat{\mathbf{Y}}_{\mathbf{t+h}} &= \hat{\mathbf{M}} + \hat{\mathbf{B}}\hat{\mathbf{Y}}_{\mathbf{t+h-1}}
\end{aligned}
\tag{7}
$$

and $h > 1$. Also, to ensure the age-coherence in the short run, Li and Lu (2017) conduct

[3]Note that the STAR and ASTAR models discussed in this paper only consider one lag in the VAR specification. Including more lags is of great interest to further improve the forecasting accuracy, but may introduce more difficulty in the explanations of parameters. A comprehensive analysis on this is out of the scope of this paper and remains for future works.

the estimation in a penalized leasts-squares (PLS) fashion. The details are discussed in Section 4.

3.2. The adaptive spatial temporal autoregressive (ASTAR) model

Despite the effectiveness of the STAR model, the sparsity of the coefficient matrix $\mathbf{B}$ is quite restrictive. For each $y_{i,t}$, by enforcing all coefficients of $y_{i-k,t}$ $(k > 3)$ to be exactly 0, this specification ignores all potential cohorts effects for those younger cohorts. In other words, only the same and neighboring cohort effects are considered in the STAR model. This is a relatively ad-hoc structure and clearly would not sufficiently capture the hyperbolic trend observed in Figure 1.

In order to retain the advantages of the STAR model, we develop a more comprehensive extension which adaptively considers the impacts of all younger cohorts. More specifically, the coefficient matrix is modified to be

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots \\ \beta_2 w_{21} & 1-\beta_2 & 0 & \cdots & \cdots \\ \beta_3 w_{32} & \beta_3 w_{31} & 1-\beta_3 & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \beta_N w_{N,N-1} & \cdots & \beta_N w_{N,2} & \beta_N w_{N,1} & 1-\beta_N \end{bmatrix} \tag{8}$$

Contrast to (6), we distribute the impact of the age $i-2$ and $i-1$ ($\alpha_i + \beta_i$ as in (6)) to all ages $i-k$ $(0 < k < i-1)$.

The distribution to age $i-k$ is based on the corresponding weight w_{ik}, which is calculated as follows.

1. The sample correlations $(r_{x,k})$ of $y_{x,t}$ and $y_{x-k,t-1}$ are calculated for all positive lag k and ages $x = k+1, k+2, ..., N$.
2. For each k, derive the averaged r_k as $1/N \sum_{x=1}^{N} r_{x,k}$, which is analogous to the autoregressive function (ACF) at kth lag used in the time series analysis.
3. For each age $i \geq 2$, we rescale r_k as w_{ik} such that $\sum_{k=1}^{i-1} w_{ik} = 1$. More specifically, $w_{ik} = r_k / \sum_{l=1}^{i-1} r_l$ for $1 \leq k \leq i-1$.

The merits of this adaptive weight are two-fold. First, all w_{ik} are created via a data-driven non-parametric approach, which does not require additional assumptions. Second, the use of those weights essentially utilize the information of all younger cohorts $(y_{i-k,t-1})$ adaptively. Thus, this structure would be expected to outperform the more constrained sparse coefficient matrix $\mathbf{B}$ as considered in Li and Lu (2017). Also, as

discussed in Section 4, this more flexible structure only requires two tunning parameter, which is actually less computational intensive than the STAR model.

To demonstrate the second merit, note that for $y_{i,t}$, the STAR model considers two explanatory variables ($y_{i-1,t-1}$ and $y_{i-2,t-1}$), apart from the autoregressive lag $y_{i,t-1}$. In contrast, the ASTAR model only requires one additional variable $y^w_{i,t-1}$, which, however, adaptively contains all information of the younger ages (from 0 to $i-1$), defined as $\sum_{k=1}^{i-1} y_{i-k,t-1}w_{i,k}$. More specifically, one can show that for the ASTAR model,

$$\begin{aligned} y_{i,t} &= m_i + (1-\beta_i)y_{i,t-1} + \beta_i \sum_{k=1}^{i-1} y_{i-k,t-1}w_{i,k} + \varepsilon_{i,t} \\ &= m_i + (1-\beta_i)y_{i,t-1} + \beta_i y^w_{i,t-1} + \varepsilon_{i,t} \end{aligned}$$

As the weights are purely data-driven, $y^w_{i,t-1}$ created in such a fashion is more informative than the lagged rates of just the closest two younger ages ($i-1$ and $i-2$). Hence, ASTAR is expected to improve STAR for the forecast accuracy, as more comprehensive information is now utilized. The forecasting with ASTAR is conducted in the same way as described in (7). Its prediction interval can be produced via simulation, by assuming that $\varepsilon_{i,t}$ follows a multi-Gaussian distribution, as considered in the LC and STAR. More specifically, the sample covariance matrix of the estimated $\varepsilon_{i,t}$ are used to perform the simulation. The lower and upper bounds of the prediction interval are then just the percentiles of the simulated rates. The estimation and other technical features of the ASTAR model are similar to those of STAR and are discussed in Section 4.

4. Technical details of the ASTAR model

As the ASTAR is a more flexible extension of the STAR model, the estimation would be performed in the same fashion. All the technical/statistical features of STAR can be also retained.

4.1. Stationarity of the ASTAR model

As described in (8), the coefficient matrix $\mathbf{B}$ is constrained so that each row sums to exactly 1. More generally speaking, we can rewrite the ASTAR model as follows:

$$
\begin{aligned}
y_{1,t} &= m_1 + y_{1,t-1} + \varepsilon_{1,t} \\
y_{2,t} &= m_2 + (1 - b_{21}) y_{2,t-1} + b_{21} y_{1,t-1} + \varepsilon_{2,t} \\
y_{3,t} &= m_3 + (1 - b_{31} - b_{32}) y_{3,t-1} + b_{31} y_{1,t-1} + b_{32} y_{2,t-1} + \varepsilon_{3,t} \\
y_{i,t} &= m_i + (1 - \sum_{l=1}^{i-1} b_{il}) y_{i,t-1} + \sum_{l=1}^{i-1} b_{il} y_{l,t-1} + \varepsilon_{i,t}
\end{aligned}
\tag{9}
$$

where $i > 3$. Compared to (8), we have that $b_{ik} = \beta_i w_{i,i-k}$ and $\sum_{l=1}^{i-1} b_{il} = \beta_i$.

Proposition 1. *Indeed under the specification (9) and assume that all residuals $\varepsilon_{i,t}$ are stationary and all $0 < b_{ij} < 1$, different component processes $y_{i,t}$ and $y_{j,t}$ are co-integrated, with co-integration vector (1, -1).*

Proof. See Appendix A. □

Therefore, as long as β_i for $i = 2, 3, .., N$ in (8) all fall in the range (0,1), $y_{i,t}$ is co-integrated as those of the STAR model. This successfully resolves the stationarity issue for VAR-type models, and ensures the age-coherence in the long run.

4.2. Estimation with PLS

To ensure the age-coherence in the short run, Li and Lu (2017) introduce smoothing parameters for αs and βs in (6) and obtain the estimates via PLS. Following the same design, given the data-driven adaptive weights w_{ik}, we have the objective function of ASTAR as described below.

$$
\begin{aligned}
LF = &\sum_{i=2}^{N} \sum_{t=2}^{T} \left[y_{i,t} - m_i - (1 - \beta_i) y_{i,t-1} - \sum_{l=1}^{i-1} \beta_i w_{i,i-l} y_{l,t-1} \right]^2 \\
&+ \sum_{t=2}^{T} (y_{1,t} - y_{1,t-1} - m_1)^2 + \lambda_m \sum_{i=2}^{N} (m_i - m_{i-1})^2 \\
&+ \lambda_\beta \sum_{i=3}^{N} (\beta_i - \beta_{i-1})^2
\end{aligned}
\tag{10}
$$

where λ_m and λ_β are pre-selected smoothing parameters of $\mathbf{M}$ and $\boldsymbol{\beta}$ respectively, and $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_N)'$. The estimates of $\boldsymbol{M}$ and $\boldsymbol{\beta}$ can then be derived by minimizing LF. With greater values of λ_m (λ_β), the fitted $\mathbf{M}$ ($\boldsymbol{\beta}$) will be smoother across ages. when both λ_m and λ_β are equal to 0, this is a special case of PLS with no penalty and reduces to the usual ordinary lease square case. With smoother changed $\mathbf{M}$ and $\boldsymbol{\beta}$, the forecast $\hat{\mathbf{Y}}_{\mathbf{t+h}}$ are expected to change smoother from $i-1$ to i. In other words, mortality rates will not

diverge for neighboring ages even in the short run. Similar to STAR, as quadratic errors are employed, the estimates of all parameters have closed-form solutions, the existence of which is proved in Appendix B.

Altogether, there are two tuning parameters for our ASTAR model, λ_m and λ_β. They need to be selected before performing the estimation. A common solution for this is to employ cross-validation. However, due to the time-series nature, related method such as leave-one-age-group-out is not applicable for the ASTAR model. Hence, we employ the procedure discussed in Hyndman and Athanasopoulos (2018) to perform cross-validation, which is also known as 'evaluation on a rolling forecasting origin.' The basic algorithm is explained below:

1. Identify the first training sample (e.g. $y_{i,2}$, $y_{i,3}$,...,$y_{i,0.7T}$ for $i = 1, 2, ..., N$) out of the the entire dataset;
2. Given a set of d, λ_m and λ_β, use the training sample to fit the ASTAR model and obtain the 1-step-ahead forecast $\hat{y}_{i,0.7T+1}$;
3. Extend the training set to include $y_{i,0.7T+1}$ and refit the ASTAR model to obtain the 1-step-ahead forecast $\hat{y}_{x,0.7T+2}$;
4. Repeat steps 2–3 until $\hat{y}_{i,T}$ is generated; and
5. Calculate the root of mean squared error (RMSE) as

$$\sqrt{\frac{1}{0.3T \times N}\sum_{i=1}^{N}\sum_{h=1}^{0.3T}(y_{i,0.7T+h} - \hat{y}_{i,0.7T+h})^2}$$

λ_m and λ_β are then chosen as those with the smallest RMSE via a grid search, where the potential ranges of λ_m and λ_β are both $[0,\infty)$.[4]

In the case of the STAR model, there are actually one more tuning parameter (smoothing penalties for ms, αs and βs as in (4)) to choose. Since both models have closed-form solutions, the ASTAR model essentially results in less computational cost and intensity. In other words, via adopting the data-driven adaptive weights, the proposed ASTAR model enables us to examine more comprehensive cohort effects of younger ages with less computational expense.

[4]We limit the upper bound to 10^5 in our empirical analyses.

5. Empirical application

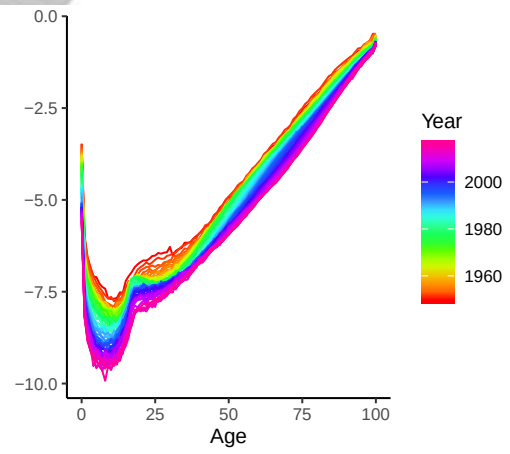
In this paper, we focus on uni-sex[5] mortality data of four out of the five largest European countries: the United Kingdom (UK), France, Italy and Spain. Data of Japan are also included for comparison. All data are obtained from the Human Mortality Database (2019). Following Booth et al. (2006), we choose an opportune range of data starting from 1950 to 2016 in order to have a reliable and complete dataset. Age groups 0–100 are included in the sample. The crude total mortality rates are studied, and the log rates are plotted in Figure 2 across all investigated years. Consistent improvements over time are observed for all the five countries. It can also be seen that the Italian and Spanish data are relatively rougher than the rest, with large variations for ages 20–40. For instance, mortality rates of age 25 in the 1990s are higher than those in the 1980s.

To illustrate the powerfulness of our proposed model, we consider the training sample of 1950–2000 and forecast the mortality rates for 2001–2016. To illustrate the assigned weights w_{ik}, for all the five countries, we plot those distributed to each younger cohort at age 10 in Appendix C. Among the 10 available weights, we can see that the derived patterns among all countries are very similar. Also, consistent with Figure 1, those weights do not reduce geometrically with the increment of lags. The estimates of $1 - \beta_i$ (period effects) are presented in Appendix D, along with 95% confidence intervals.[6] Despite some differences, all curves start from around 1 at age 0 and almost monotonically decline. After reaching the bottom age (around 70 for UK and France, 95 for Italy and Japan, and 25 for Spain), estimates begin to increase for all countries. Roughly speaking, except for Spain, those patterns suggest that the period (cohort) effects are lower (higher) for older ages. For UK and France, after around age 70, the period (cohort) effects of the oldest ages start to increase (decline). As for the magnitudes, with a close to 1 maximum in all cases, the minimums of the period effects range from 0.25 (Italy) to 0.45 (Spain).

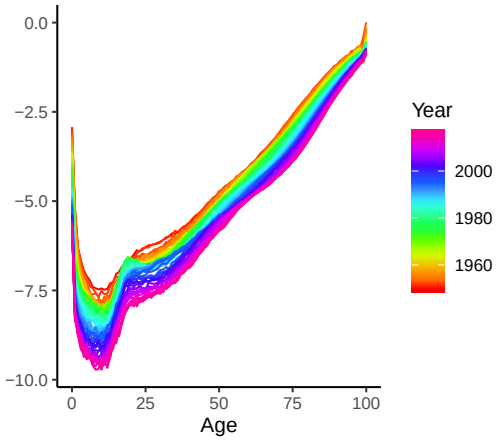
Based on the obtained estimates, the out-of-sample results of the LC, STAR and ASTAR models are compared. Two sets of relevant sensitivity analyses are then conducted. The first set shrinks the sample to model ages 41–90 only, as done in Giacometti et al. (2012). The second set considers a shorter training period over 1960–2000. Following that, we perform a simulation study to check the robustness of our empirical results.

[5] We have also modeled the female and male data separately. The results are consistent with those present in this section, and are available upon request.

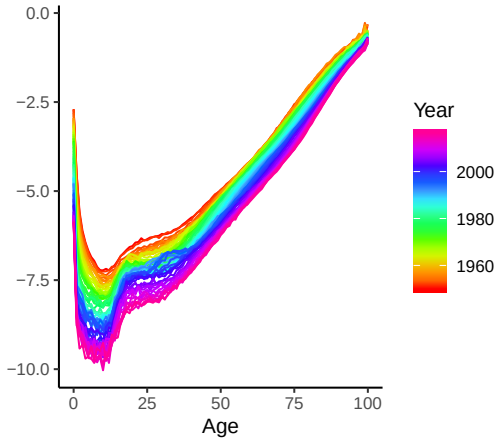
[6] The confidence intervals are produced using 1,000 Bootstrap replicates. The lower and upper bounds are the corresponding 2.5th and 97.5th percentiles, respectively.



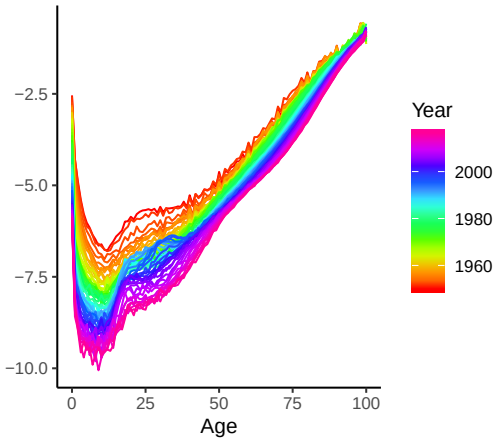
(a) UK



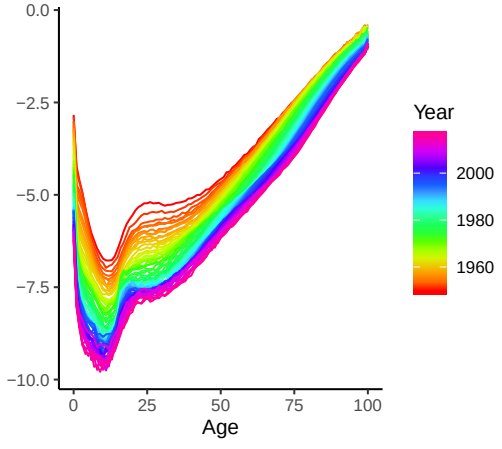
(b) France



(c) Italy



(d) Spain



(e) Japan

Figure 2: Mortality data 1950–2016

Finally, we perform a long-term analysis of the life expectancies at new born.

5.1. Out-of-sample forecasting performance

To compare the forecasting performance across all models, we follow Li and Lu (2017) and employ the RMSE. We consider RMSEs over age groups and at individual time horizons separately and an overall measure as follows:

$$\begin{aligned} RMSE_x &= \sqrt{\frac{1}{16}\sum_{h=1}^{16}(y_{x,T+h} - \hat{y}_{x,T+h})^2} \\ RMSE_h &= \sqrt{\frac{1}{101}\sum_{x=0}^{100}(y_{x,T+h} - \hat{y}_{x,T+h})^2} \\ RMSE_{all,h} &= \sqrt{\frac{1}{101 \times h}\sum_{i=1}^{h}\sum_{x=0}^{100}(y_{x,T+i} - \hat{y}_{x,T+i})^2} \end{aligned} \tag{11}$$

$RMSE_x$ ($RMSE_h$) is the RMSE averaged over all 16 forecasting steps (101 age groups) for age group x (time horizon h). $RMSE_{all,h}$ is the overall measure considering both dimensions up to step h. Relevant results for all the four models are reported in Table 1, as well as in Figures 3 and 4.

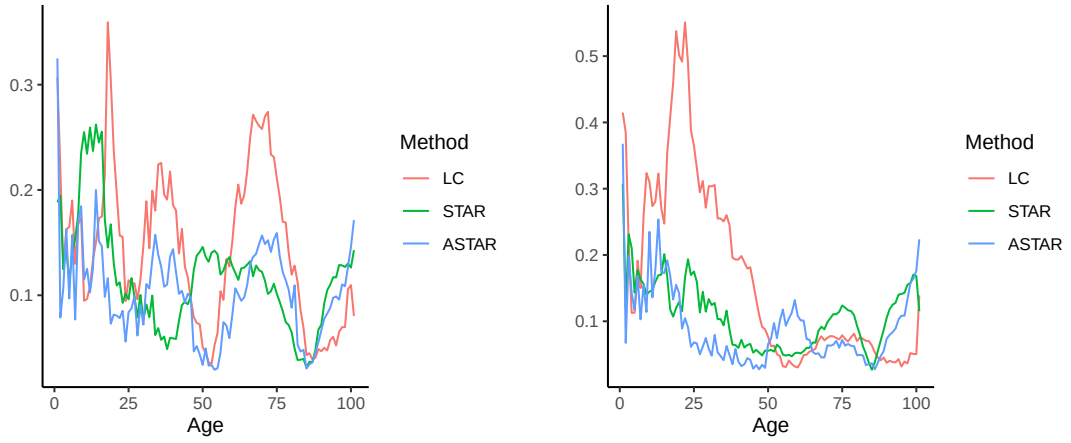
Figure 3 displays the $RMSE_x$. For most countries, both STAR and ASTAR models produce smaller RMSE than the LC model, especially among the young ages. Comparing the two spatial temporal models, despite the mixed observations for specific age groups, the RMSE curves of ASTAR are overall lower than those of STAR for young to middle-aged groups in almost all cases. For instance, ASTAR consistently produces smaller RMSE than STAR across ages 20–50 for the French, Italian and Japanese data. This suggests that for those age groups, allowing for more informative influences from the younger cohorts can improve the forecasting accuracy of mortality rates.

Descriptive statistics of $RMSE_x$ are reported in Table 1. The mean $RMSE_x$ across all age groups of the ASTAR model is around 30%, 45%, 25%, 25% and 60% (15%, 15%, 25%, 10% and 30%) smaller than that of LC (STAR) for UK, France, Italy, Spain and Japan, respectively. Q_1 and Q_3 measures further support that ASTAR model performs almost uniformly the best among the three competing models. Standard deviation of $RMSE_x$ confirms that results of ASTAR are more narrowly spread than those of LC and ASTAR in four out of five cases. Taking the Japanese data for example, the standard deviation of STAR (0.1366) is over 60% smaller than that of LC (0.3605) and around 30% smaller than that of STAR (0.1960). As indicated by $RMSE_{all,16}$, the overall

Table 1: Summary of RMSE over age groups

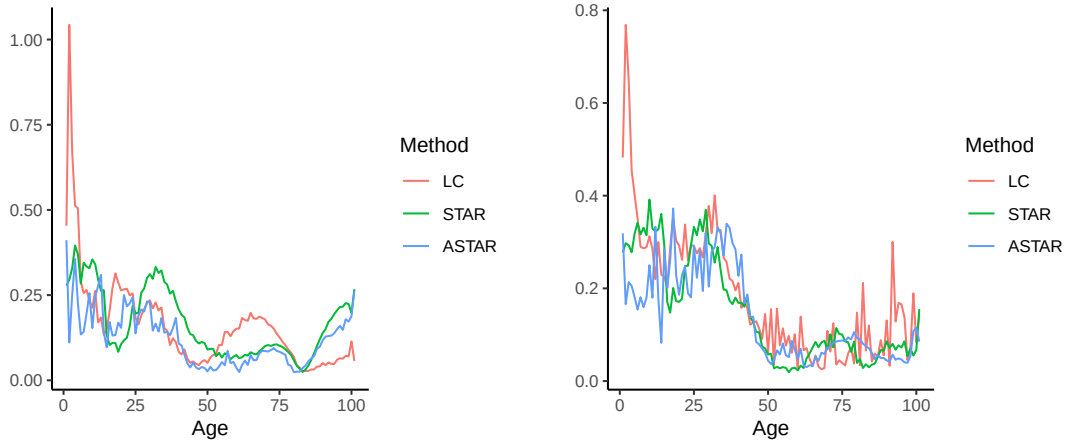
Model	$RMSE_{all,16}$	Mean	*Std. Dev.*	Q_1	Q_3
Panel A: UK					
LC	0.1619	0.1447	0.0729	0.0859	0.1931
STAR	0.1280	0.1172	0.0519	0.0850	0.1362
ASTAR	**0.1119**	**0.1024**	**0.0454**	**0.0745**	**0.1295**
Panel B: France					
LC	0.2158	0.1653	0.1394	0.0532	0.2600
STAR	0.1184	0.1074	**0.0500**	0.0606	0.1411
ASTAR	**0.1074**	**0.0907**	0.0578	**0.0493**	**0.1128**
Panel C: Italy					
LC	0.2140	0.1612	0.1414	0.0648	0.2050
STAR	0.1928	0.1663	0.0979	0.0834	0.2573
ASTAR	**0.1462**	**0.1225**	**0.0802**	**0.0536**	**0.1692**
Panel D: Spain					
LC	0.2280	0.1828	0.1369	0.0665	0.2864
STAR	0.1854	0.1493	0.1105	0.0630	0.2491
ASTAR	**0.1703**	**0.1402**	**0.0970**	**0.0576**	**0.2052**
Panel E: Japan					
LC	0.4449	0.3605	0.2621	0.1180	0.5636
STAR	0.2416	0.1960	0.1421	**0.0566**	0.3048
ASTAR	**0.1693**	**0.1366**	**0.1005**	0.0646	**0.2042**

Note: this table displays the RMSE over age groups for the 16-step-ahead forecasts of total mortality rates of UK, France, Italy, Spain, Japan. $RMSE_{all,16}$ is the overall RMSE across all ages and time horizons. Mean, *Std. Dev.*, Q_1 and Q_3 are the sample mean, standard deviation, first quartile and third quartile of the RMSEs over age groups, respectively. Bold numbers represent the smallest RMSEs among three models. LC, STAR and ASTAR stand for Lee-Carter, spatial temporal autoregressive and adaptive spatial temporal autoregressive models, respectively. Bold numbers represent the smallest value among three models.



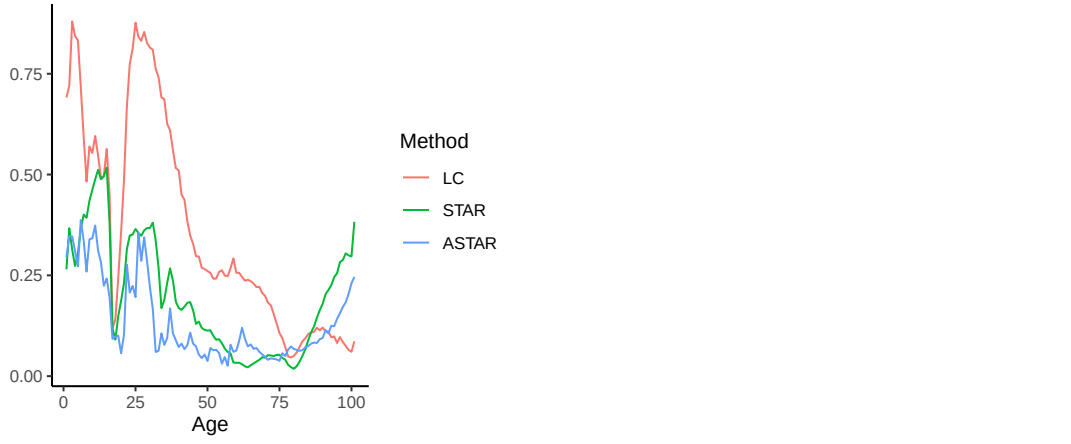
(a) UK

(b) France



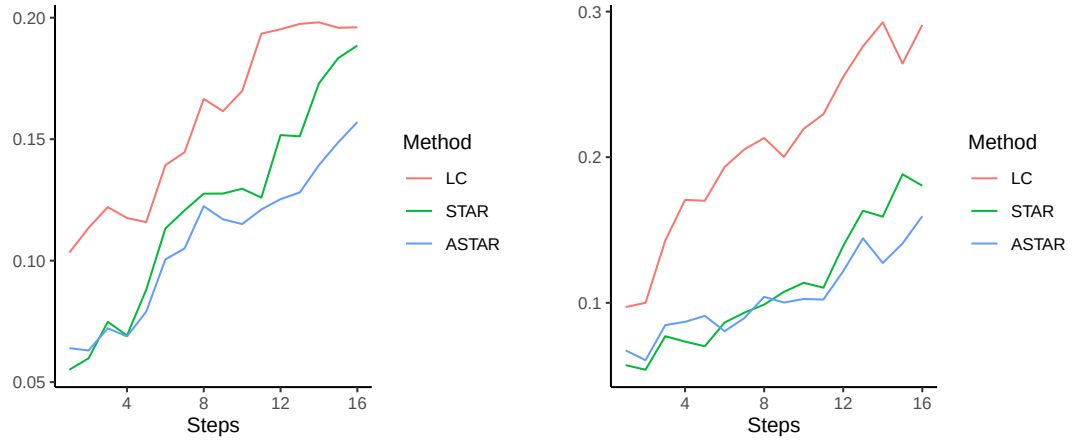
(c) Italy

(d) Spain



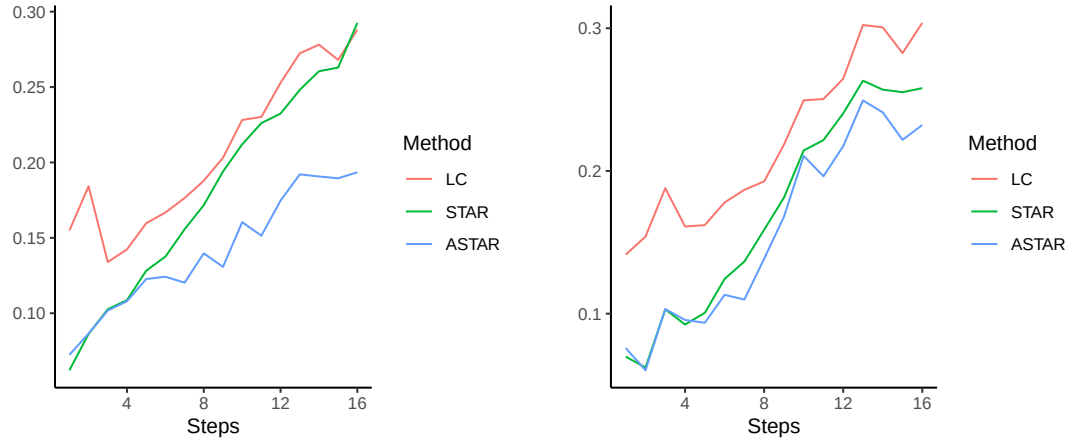
(e) Japan

Figure 3: RMSE over ages



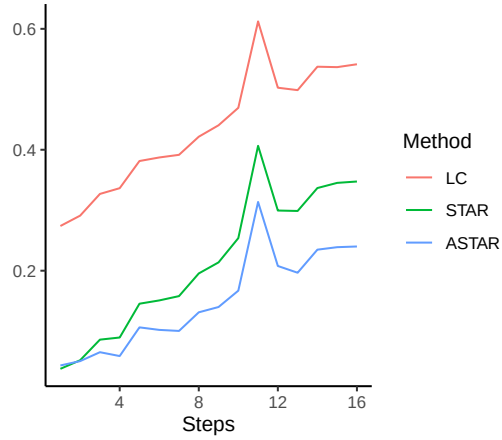
(a) UK

(b) France



(c) Italy

(d) Spain



(e) Japan

Figure 4: RMSE over forecasting steps

performance of our proposed ASTAR model also beats the rest.

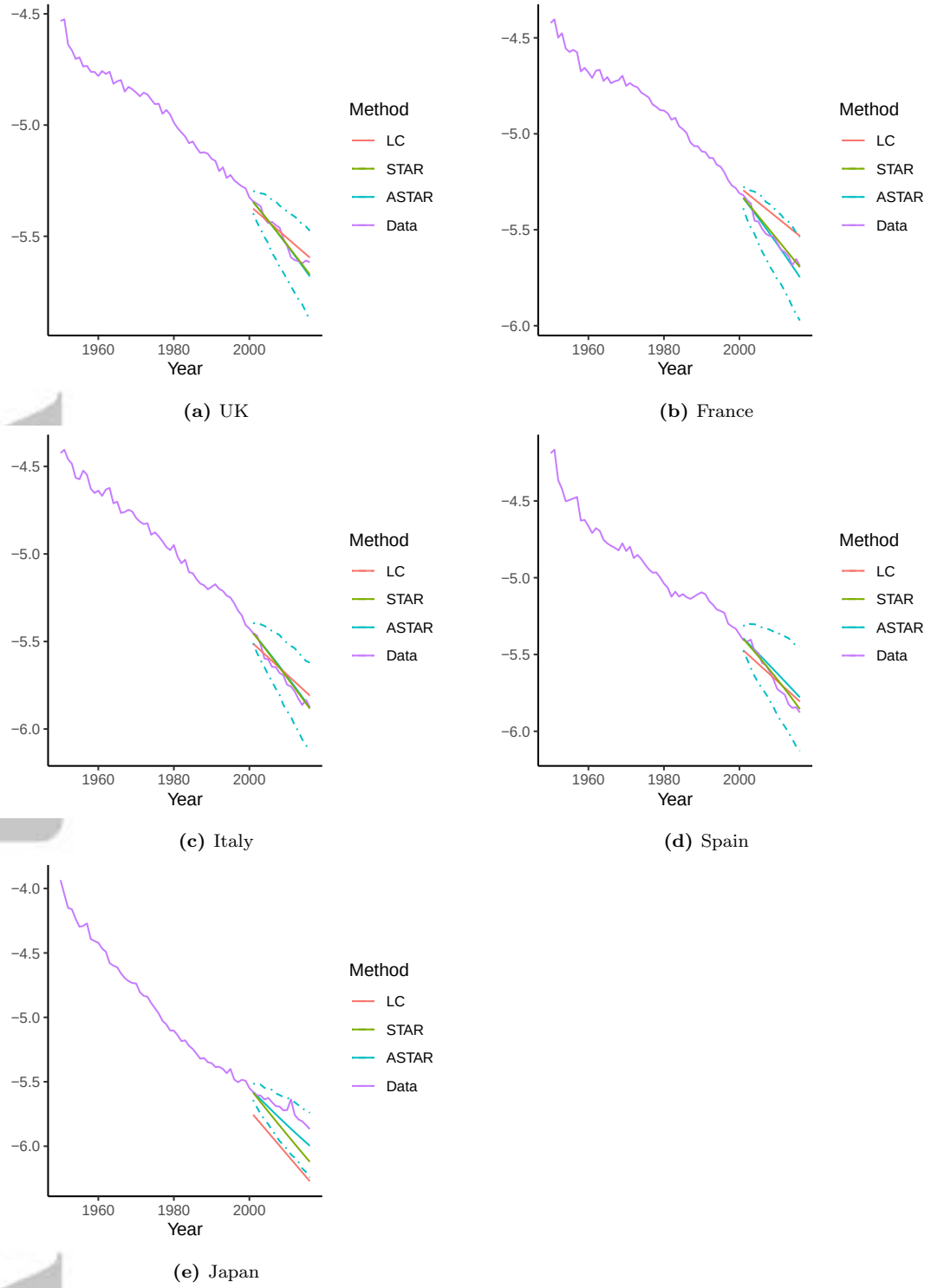
Figure 4 plots the $RMSE_h$ at individual forecasting steps ranging from 1 (2001) to 16 (2016). Distinct differences among all the three models can be observed, especially at larger steps. In general, spatial temporal models almost uniformly beat the LC in all scenarios. Comparing with STAR model, $RMSE_h$ of the ASTAR is smaller in the vast majority of cases. Further, with the growth of h (especially from the 6th step onwards), the increment in $RMSE_h$ is slower for the ASTAR than that of LC and STAR, suggesting its better performance in the long run.

To visually compare the forecast rates against the actual rates, we follow Li and Lu (2017) and plot the temporal rates averaged across all ages 0–100. The actual data from 1950–2016 and the forecast rates over 2001–2016 produced by all the three modes are plotted in Figure 5. In addition to the point estimates, we include the 95% prediction intervals (PIs) of our ASTAR model, based on simulated multi-Gaussian distribution with 5,000 replicates, as described in Section 3.2. For the point estimates, the spatial temporal models produce forecasts that are much closer to the true data than the LC for all countries. Those averaged point estimates of STAR and ASTAR are almost identical for UK and Italy. ASTAR produces visually much better estimates than STAR for French and Japanese data. As for the interval estimates, almost all true data fall into the 95% PIs generated by the ASTAR model, except one point for the Japanese data in 2010, which is distinctively different from other observations (an 'outlier'). This is in-line with the expected confidence level of the designed PIs (i.e., only 1 violation out of the 80 cases).

To check the sensitivity of our baseline results presented above, we consider two scenarios. In the first case, we shrink the data to cover ages 41–90 as done by Giacometti et al. (2012), and in the second one we consider a shorter training period over 1960–2000. We report $RMSE_{all,16}$ produced by the three models in Table 2. It can be seen than except for the Spanish data in the first scenario, the proposed ASTAR model still uniformly outperforms STAR and LC models. Hence, we conclude that the forecasting superiority of the ASTAR model over LC and STAR is insensitive to age and period variations.

5.2. Simulation results

To further examine the forecasting effectiveness of the ASTAR model shown above, we follow Feng and Shi (2018) and perform simulation studies in this section. For all the five countries, we generate 1,000 replicates. Those replicates are generated by weighted



(a) UK

(b) France

(c) Italy

(d) Spain

(e) Japan

Figure 5: Forecast vs actual mortality rates averaged across 0–100

Table 2: Results of sensitivity tests

	41-90			1960–2000		
	LC	STAR	ASTAR	LC	STAR	ASTAR
UK	0.1559	0.1065	**0.0957**	0.1504	0.1247	**0.1081**
France	0.0810	0.0770	**0.0686**	0.2181	0.1154	**0.1131**
Italy	0.1151	0.0942	**0.0638**	0.2046	0.1968	**0.1556**
Spain	0.0955	**0.0765**	0.0789	0.3212	0.2146	**0.2099**
Japan	0.2191	0.0954	**0.0693**	0.2388	0.1974	**0.1636**

Note: this table displays the $RMSE_{all,16}$ of forecasts in two sets of sensitivity tests. The first set considers ages 41–90. The second one uses training period over 1960–2000. Bold numbers represent the smallest value among three models.

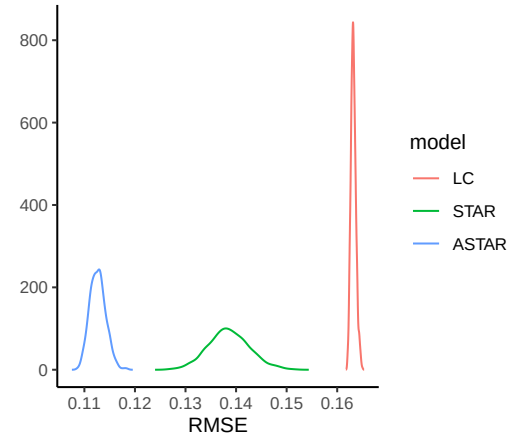
penalized regression splines with a monotonicity constraint (Wood, 1994). This method is frequently used to produce smoothed mortality rates based on the crude data. We firstly fit the entire sample 1950–2000 via the regression splines, the residuals are then collected as the fitted rates for age x at time t subtracted from the true log of mortality rates. Assuming a multi-Gaussian distribution with sample mean and covariances of the collected residuals as the population mean and covariances, respectively, 51×101 errors are then simulated for the data of all the five countries. Further added to the fitted values of the corresponding splines, one complete simulated sequence is produced. Such a procedure is repeated until 1,000 replicates are created. We then follow the same steps as in Section 5.1 to fit the LC, STAR and ASTAR models. Based on the simulated data, the 16-step-ahead forecasts are finally generated in each case, and the $RMSE_{all,16}$ can be calculated using the true sample of 2001–2016. Those produced RMSEs are summarized in Table 3.

The descriptive statistics presented in Table 3 are largely consistent with our previous findings. Although LC models demonstrate the smallest spread, the resulting mean $RMSE_{all,16}$ are much worse than those of the spatial temporal models. In all cases, ASTAR model produces the best average, Q_1 and Q_3 statistics for $RMS_{all,16}$ for all the five sets of simulated replicates. To visually compare the simulation results, we plot the smoothed densities of those RMSEs in Figure 6. Despite the mixed shapes of distributions, the densities generated by the ASTAR are uniformly distributed lower than those of LC and STAR, with distinctive differences in all cases.

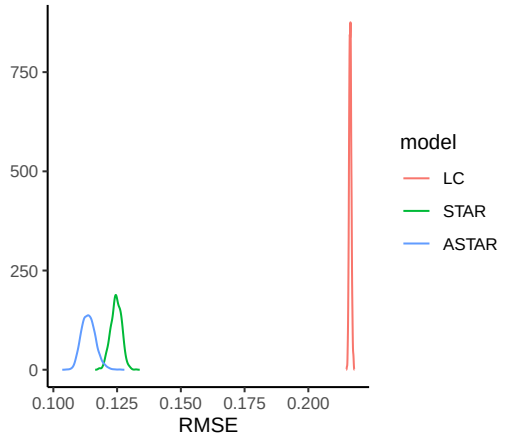
Table 3: Summary of simulation results

Model	Mean	*Std. Dev.*	Q_1	Q_3
Panel A: UK				
LC	0.1632	**0.0005**	0.1629	0.1635
STAR	0.1386	0.0041	0.1359	0.1412
ASTAR	**0.1127**	0.0016	**0.1115**	**0.1137**
Panel B: France				
LC	0.2164	**0.0004**	0.2161	0.2167
STAR	0.1246	0.0021	0.1232	0.1262
ASTAR	**0.1138**	0.0028	**0.1118**	**0.1155**
Panel C: Italy				
LC	0.2130	**0.0004**	0.2128	0.2133
STAR	0.1849	0.0030	0.1828	0.1868
ASTAR	**0.1438**	0.0020	**0.1424**	**0.1450**
Panel D: Spain				
LC	0.2255	**0.0034**	0.2233	0.2276
STAR	0.1916	0.0058	0.1878	0.1952
ASTAR	**0.1608**	0.0078	**0.1553**	**0.1644**
Panel E: Japan				
LC	0.4420	**0.0010**	0.4414	0.4427
STAR	0.2483	0.0045	0.2451	0.2512
ASTAR	**0.1722**	0.0027	**0.1704**	**0.1739**

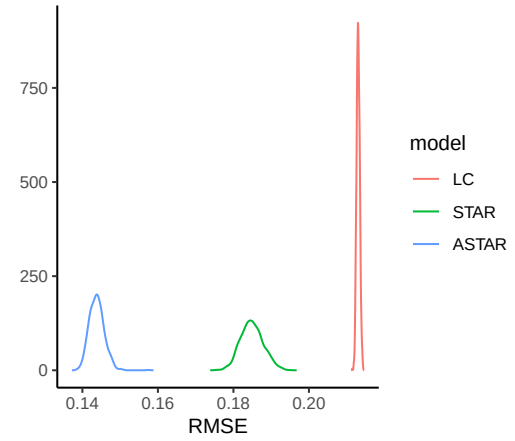
Note: this table displays the $RMSE_{all,16}$ of forecasts based on simulated French and Spanish total mortality rates. Mean, Std. Dev., Q_1 and Q_3 are the sample mean, standard deviation, first quartile and third quartile of the RMSEs over the 1000 simulated replicates. Bold numbers represent the smallest value among three models.



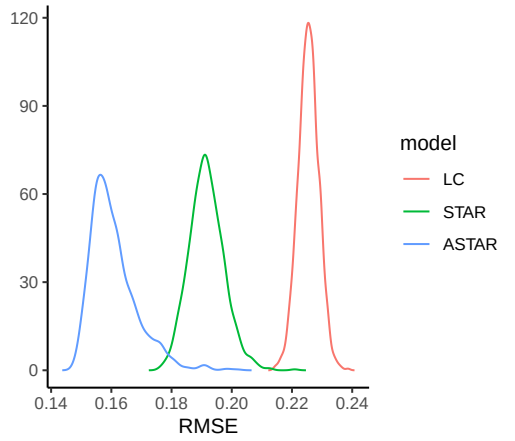
(a) UK



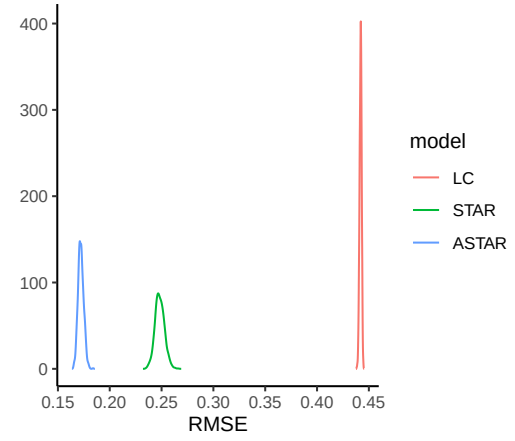
(b) France



(c) Italy



(d) Spain



(e) Japan

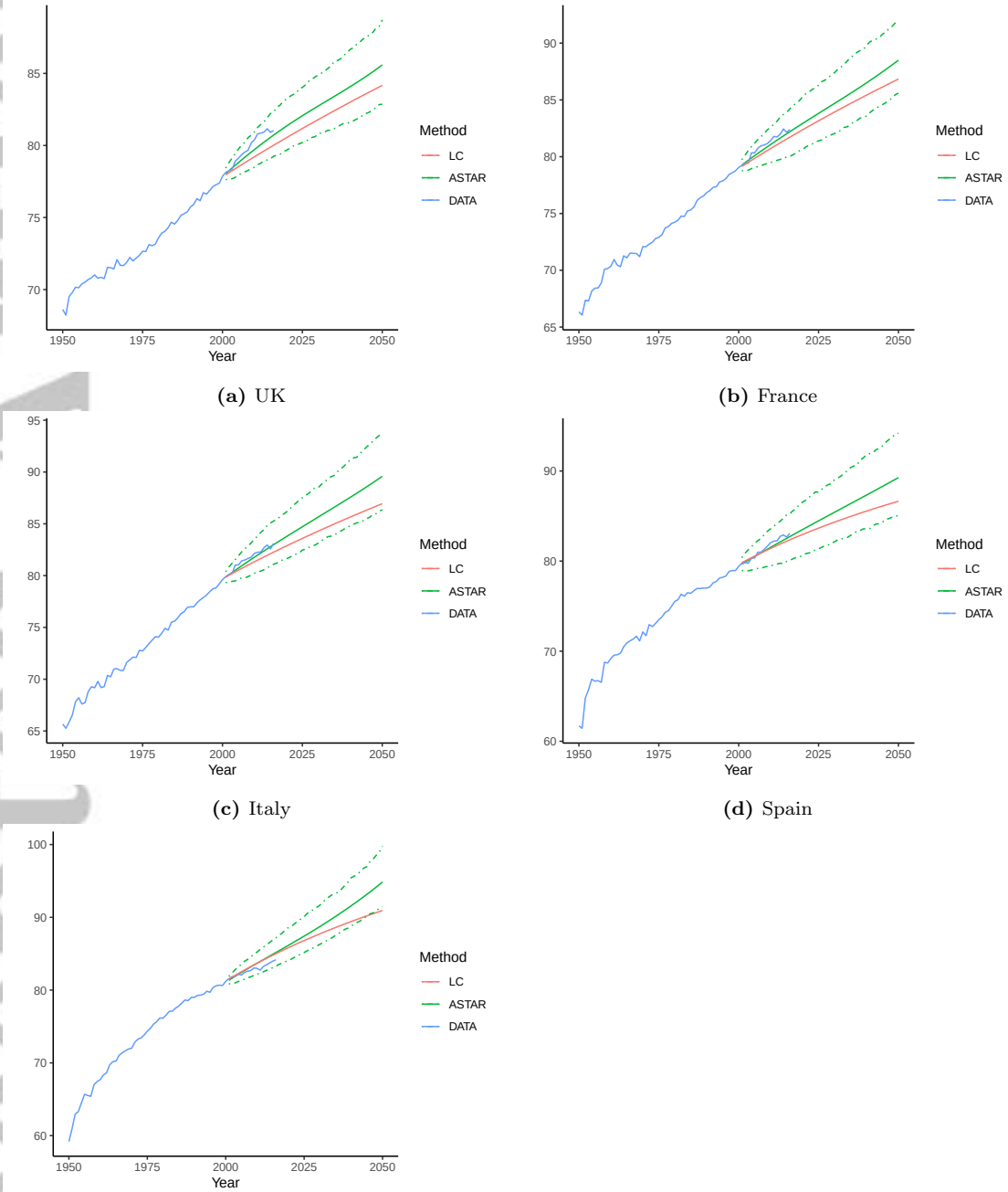
Figure 6: Density plots of RMSEs of simulated results

5.3. A long-term analysis of the life expectancies

We now consider a long-term analysis of the life expectancies at new born (e_0) for each country. To produce e_0, a comprehensive set of all forecast mortality rates needs to be employed. Hence, an analysis of e_0 will provide useful information, especially for a long-term investigation.

Using the same estimates obtained over the training sample of 1950–2000, we forecast e_0 up to 2050 with both the LC and ASTAR models. Both the point and interval forecasts are presented in Figure 7. In addition to the forecast rates, the true data over 1950–2016 are included for comparison. Following Li and Lu (2017), we produce the 95% prediction intervals (PIs) of the ASTAR model, based on the 1,000 simulated replicates, assuming that residuals follow a multi-Gaussian distribution for each country. It can be seen that the mean forecasts of e_0 produced by ASTAR are uniformly longer than those by LC. This is consistent with our statement such that the forecasts of ASTAR are age-coherent. In other words, mortality improvements at old ages will not diverge from those at the young ages in the long run, which eventually lead to longer life expectancies than non-coherent forecasts across ages (Li and Lu, 2017). In 2050, the forecast e_0 generated by the ASTAR (LC) model are 85.6, 88.5, 89.6, 89.3 and 94.9 (84.2, 86.8, 86.9, 86.6 and 90.9) for UK, France, Italy, Spain and Japan, respectively. As for the comparison with the true value over 2001–2016, e_0 of the ASTAR model are almost consistently closer to the true values than those of LC. Also, all true values uniformly fall in the corresponding 95% PIs of ASTAR for the five populations.

To sum up, via analyzing out-of-sample forecasts of the mortality data of UK, France, Italy, Spain and Japan over 2001–2016, our proposed ASTAR model consistently outperforms the LC and STAR counterparties. Its superiority is demonstrated via the RMSE over age groups and time horizons, and the overall $RMSE_{all,16}$. The sensitivity tests and simulation results also provide robust conclusions. The long-term analysis of life expectancies is consistent with the argued age-coherence property of the ASTAR model. Using the true values over 2001–2016, we also observe more accurate out-of-sample forecasts of ASTAR than those of LC. Thus, our proposed inclusion of data-driven adaptive weights to model the effects of younger cohorts is powerful in improving the forecasting accuracy of mortality rates with the VAR framework.



(a) UK **(b)** France

(c) Italy **(d)** Spain

(e) Japan

Figure 7: Forecast life expectancies at new born: 2001–2050

6. Concluding remarks

In this paper, we propose an effective adaptive spatial temporal VAR model (ASTAR) to investigate and forecast log mortality rates. Two key results can be drawn from our study. First, the proposed ASTAR model is effective in mortality rates and life expectancies forecasting. As measured by RMSE, our ASTAR model outperforms the famous Lee-Cater model (Lee and Carter, 1992), and the recently developed spatial temporal autoregressive model (Li and Lu, 2017). This conclusion consistently holds for mortality data including aged 0–100 of the United Kingdom, France, Italy, Spain and Japan, when the sample of 1950–2000 is fitted and forecasts of 2001–2016 are produced. Second, the ASTAR model has a more flexible framework that retains all advantages of the original spatial temporal autoregressive (STAR) model investigated in Li and Lu (2017). Using a data-driven adaptive structure, our ASTAR model allows for spatially weighted influences of younger cohorts. Compared with STAR, the new model requires less computational cost, whereas all desirable properties of STAR, including stationarity (co-integration) and closed-form solution, still hold. More importantly, with its more flexible spatial structure, the forecasting accuracy of ASTAR is improved in all cases over that of the STAR model. Robust results are further seen when the age and training period are shrunk individually, and a simulation study is performed.

There are also some pathways for future research. First, effects of the older cohorts (e.g. $x+1$ on x) may be allowed for more flexible structure of the coefficient matrix $\boldsymbol{B}$. Second, a selection criterion of autoregressive lags to be included in the STAR and ASTAR models could be investigated. Third, instead of including a full range of younger cohorts, it may be desirable to find an optimal threshold, such that the cohorts with weights smaller than this threshold are excluded from the model. Finally, the extension to multi-population mortality modeling using the proposed ASTAR model may be investigated. It is expected that a similar straightforward parametric structure as studied in Li and Lu (2017) could be applicable in our case. For instance, assuming a two-population case (1 and 2), an ASTAR model may be extended to

$$y_{1,i,t} = m_{1,i} + (1-\beta_{1,i}-\alpha_{1,i})y_{1,i,t-1} + \sum_{l=1}^{i-1} \beta_{1,i} w_{1,i,i-l} y_{1,l,t-1} + \alpha_{1,i} y_{2,i,t-1} + \varepsilon_{1,i,t}$$

where $y_{1,i,t}$ and $y_{2,i,t}$ are the logged mortality rates of age i at time t of the 1st and 2nd population, respectively. Systematically analyzing such an extension remains for future works.

Acknowledgment

The author would like to thank the Macquarie University for research support. We particularly thank the Editor (David Stoffer) and two anonymous referees for providing valuable and insightful comments on earlier drafts. The usual disclaimer applies.

Appendix

A. Proof of Proposition 1

For the smallest age 1, $y_{1,t}$ is clearly a random work with drift m_1.

For age 2, using the expression of $y_{2,t}$ to subtract that of $y_{1,t}$, we have that

$$y_{2,t} - y_{1,t} = m_2 - m_1 + (1 - b_{21})(y_{2,t-1} - y_{1,t-1}) + \varepsilon_{2,t} - \varepsilon_{1,t}.$$

Hence $y_{2,t} - y_{1,t}$ is stationary.

For age 3, in a similar fashion, we have that

$$y_{3,t} - y_{2,t} = m_3 - m_2 + (1 - b_{32} - b_{31})(y_{3,t-1} - y_{2,t-1}) - (b_{31} - b_{21})(y_{2,t-1} - y_{1,t-1}) + \varepsilon_{3,t} - \varepsilon_{2,t}.$$

Hence $y_{3,t} - y_{2,t}$ is stationary.

Following the same induction, for age i ($3 < i \leq N$), we have that

$$\begin{aligned} y_{i,t} - y_{i-1,t} =& m_i - m_{i-1} + (1 - \sum_{l=1}^{i-1} b_{il})(y_{i,t-1} - y_{i-1,t-1}) \\ & - \sum_{j=2}^{i-1} \left[\sum_{k=1}^{j-1} (b_{i,k} - b_{i-1,k}) \right] (y_{j,t-1} - y_{j-1,t-1}) + \varepsilon_{i,t} - \varepsilon_{i-1,t}. \end{aligned}$$

and all those $y_{j,t} - y_{j-1,t}$ where $1 < j < i$ are stationary. Hence, $y_{i,t} - y_{i-1,t}$ is also stationary, which completes the proof.

B. Proof of existence of the closed-form solution

From (10), we have that for $1 < i \leq N$,

$$\begin{aligned} y_{i,t} &= m_i + (1 - \beta_i) y_{i,t-1} + \sum_{l=1}^{i-1} \beta_i w_{i,i-l} y_{l,t-1} + \varepsilon_{i,t} \\ y_{i,t} &= m_i + (1 - \beta_i) y_{i,t-1} + \beta_i y^w_{i,t-1} + \varepsilon_{i,t} \\ y_{i,t} - y_{i,t-1} &= m_i + \beta_i (y^w_{i,t-1} - y_{i,t-1}) + \varepsilon_{i,t} \end{aligned}$$

where $y^w_{i,t-1} = \sum_{l=1}^{i-1} w_{i,i-l} y_{l,t-1}$. Hence, the constrained PLS problem of $y_{i,t}$ can be transformed to an equivalent non-constrained PLS problem of $y^w_{i,t-1} - y_{i,t-1}$ regressed against $y_{i,t} - y_{i,t-1}$.

It is then straightforward to rewrite (10) in the following matrix form.

$$LF_1 = (\triangle \boldsymbol{y} - \boldsymbol{X\theta})'(\triangle \boldsymbol{y} - \boldsymbol{X\theta}) + \lambda_m \boldsymbol{\theta' S_M \theta} + \lambda_\beta \boldsymbol{\theta' S_\beta \theta}$$

where $\triangle \boldsymbol{y} = (\triangle \boldsymbol{y}_{1,t}, \triangle \boldsymbol{y}_{2,t}, ..., \triangle \boldsymbol{y}_{N,t})'_{N(T-1)\times 1}$, $\triangle \boldsymbol{y}_{i,t} = (y_{i,2} - y_{i,1}, y_{i,3} - y_{i,2}, ..., y_{i,T} - y_{i,T-1})'_{(T-1)\times 1}$, $\boldsymbol{\theta} = (M, \boldsymbol{\beta})'_{(2N-1)\times 1}$,

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 & \cdots & \cdots & \cdots \\ 1 & 0 & \cdots & \cdots & 0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & \cdots & \cdots & y^w_{2,1} - y_{2,1} & 0 & \cdots \\ 0 & 1 & 0 & \cdots & \cdots & y^w_{2,2} - y_{2,2} & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 1 & 0 & \cdots & \cdots & y^w_{N,1} - y_{N,1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 1 & 0 & \cdots & \cdots & y^w_{N,T} - y_{N,T} \end{bmatrix}_{N(T-1)\times(2N-1)},$$

$$\boldsymbol{S_M} = \begin{bmatrix} 1 & -1 & 0 & \cdots & \cdots & \cdots & \cdots \\ -1 & 2 & -1 & 0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & -1 & 2 & -1 & 0 & \cdots \\ 0 & \cdots & \cdots & -1 & 1 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}_{(2N-1)\times(2N-1)}$$

and

$$\boldsymbol{S_\beta} = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 1 & -1 & 0 & \cdots \\ 0 & \cdots & \cdots & -1 & 2 & -1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & \cdots & \cdots & -1 & 1 \end{bmatrix}_{(2N-1)\times(2N-1)}$$

Since both $\boldsymbol{S_M}$ and $\boldsymbol{S_\beta}$ are symmetric and $\boldsymbol{\theta}$ is the only unknown value, we have that

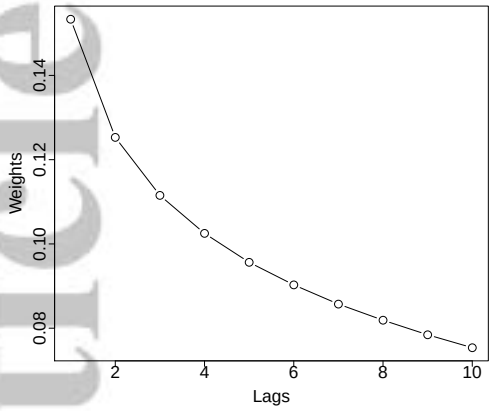
$$\frac{dLF_1}{d\boldsymbol{\theta}} = -2\boldsymbol{X}'\triangle\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\theta} + 2\lambda_m\boldsymbol{S_M}\boldsymbol{\theta} + 2\lambda_\beta\boldsymbol{S_\beta}\boldsymbol{\theta}.$$

Set it to 0, we have the estimated parameter vector

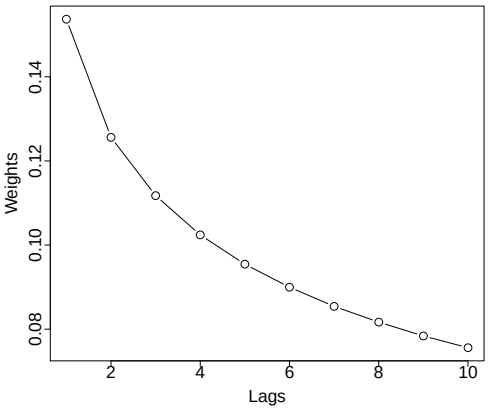
$$\hat{\boldsymbol{\theta}} = [\boldsymbol{X}'\boldsymbol{X} + \lambda_m\boldsymbol{S_M} + \lambda_\beta\boldsymbol{S_\beta}]^{-1}\boldsymbol{X}'\triangle\boldsymbol{y},$$

which proves the existence of the closed-form solution.

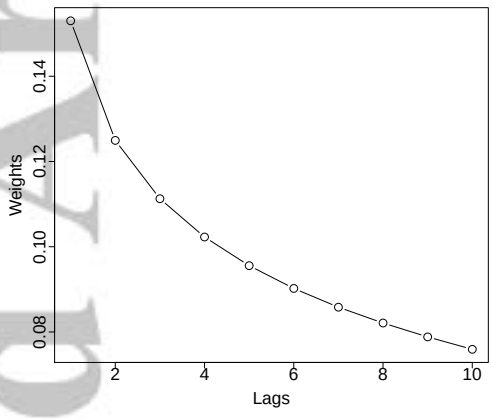
C. An example of w_{ik}: Weights assigned for all younger cohorts of the age 10



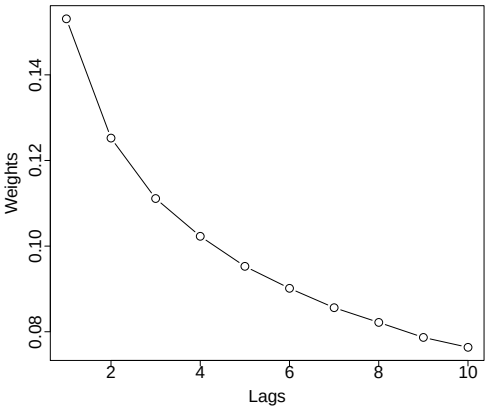
(a) UK



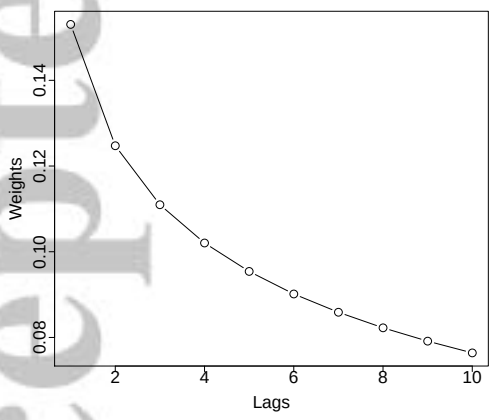
(b) France



(c) Italy

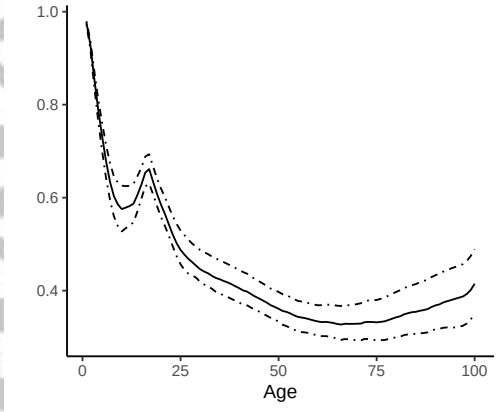


(d) Spain

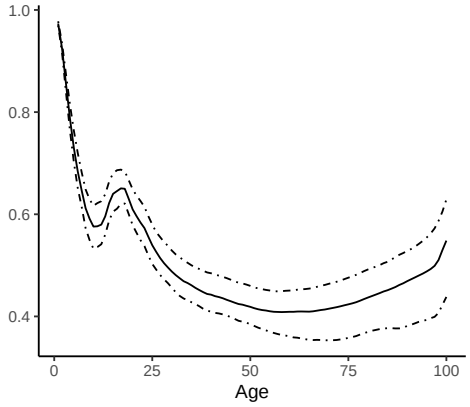


(e) Japan

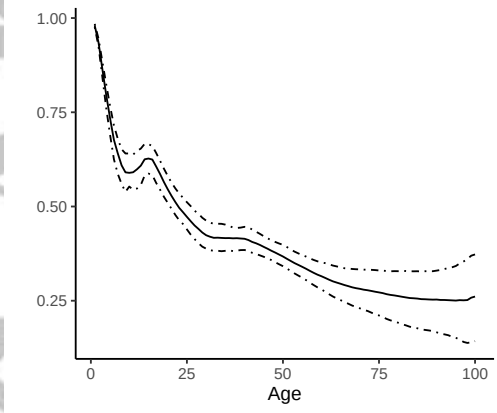
D. Estimates of $1-\beta_i$



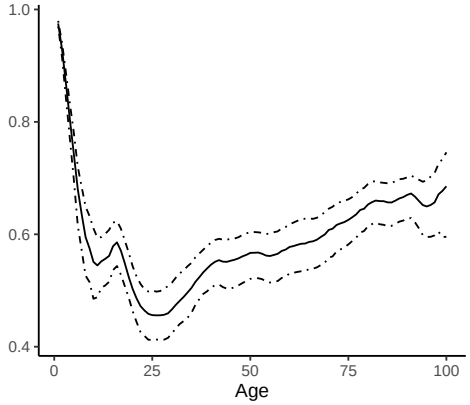
(a) UK



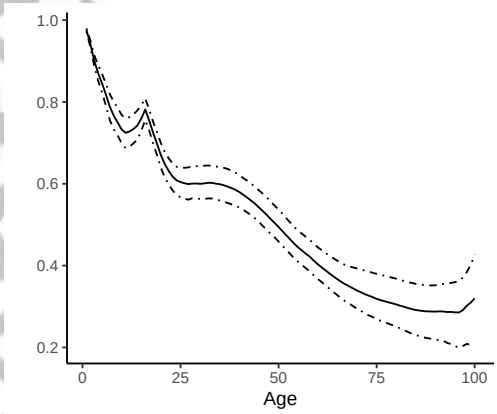
(b) France



(c) Italy



(d) Spain



(e) Japan

E. Computational packages

We use the software **R** to perform the computations of all models. LC models are estimated using the **demography** package. STAR and ASTAR are computed with codes written by the authors.

Data Availability Statement

The data used in this research are publicly available and sourced from Human Mortality Database (2019).

References

Baillie, R.T., Bollerslev, T., Mikkelsen, H.O., 1996. Fractionally integrated generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 74, 3 – 30.

Barrieu, P., Bensusan, H., El Karoui, N., Hillairet, C., Loisel, S., Ravanelli, C., Salhi, Y., 2012. Understanding, modelling and managing longevity risk: key issues and main challenges. Scandinavian actuarial journal 2012, 203–231.

Bollerslev, T., Mikkelsen, H.O., 1996. Modeling and pricing long memory in stock market volatility. Journal of Econometrics 75, 151–184.

Booth, H., Hyndman, R., Tickle, L., De Jong, P., 2006. Lee-Carter mortality forecasting: A multi-country comparison of variants and extensions. Demographic Research 15, 289–310.

Booth, H., Maindonald, J., Smith, L., 2002. Applying lee-carter under conditions of variable mortality decline. Population studies 56, 325–336.

Chang, L., Shi, Y., 2020. Dynamic modelling and coherent forecasting of mortality rates: A time-varying coefficient spatial-temporal autoregressive approach. Scandinavian Actuarial Journal , 1–21.

Davidson, J., 2004. Moment and memory properties of linear conditional heteroscedasticity models, and a new model. Journal of Business & Economic Statistics 22, 16–29.

Feng, L., Shi, Y., 2017. Fractionally integrated garch model with tempered stable distribution: A simulation study. Journal of Applied Statistics 44, 2837–2857.

Feng, L., Shi, Y., 2018. Forecasting mortality rates: Multivariate or univariate models? Journal of Population Research 35, 289–318.

Feng, L., Shi, Y., Chang, L., 2020. Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting .

Giacometti, R., Bertocchi, M., Rachev, S.T., Fabozzi, F.J., 2012. A comparison of the Lee-Carter model and AR-ARCH model for forecasting mortality rates. Insurance: Mathematics and Economics 50, 85–93.

Girosi, F., King, G., 2007. Understanding the Lee-Carter Mortality Forecasting Method1. Technical Report. Rand Corporation, Santa Monica, CA.

Granger, C.W., Hyung, N., 2004. Occasional structural breaks and long memory with an application to the s&p 500 absolute stock returns. Journal of Empirical Finance 11, 399–421.

Guibert, Q., Lopez, O., Piette, P., 2019. Forecasting mortality rate improvements with a high-dimensional var. Insurance: Mathematics and Economics 88, 255–272.

Ho, K.Y., 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.J., Athanasopoulos, G., 2018. Forecasting: Principles and practice. OTexts.

Lee, R.D., Carter, L.R., 1992. Modeling and forecasting US mortality. Journal of the American statistical association 87, 659–671.

Li, H., Lu, Y., 2017. Coherent forecasting of mortality rates: A sparse vector-autoregression approach. ASTIN Bulletin: The Journal of the IAA 47, 563–600.

Li, M., Li, W.K., Li, G., 2015. A new hyperbolic garch model. Journal of econometrics 189, 428–436.

Renshaw, A.E., Haberman, S., 2003. Lee–carter mortality forecasting with age-specific enhancement. Insurance: Mathematics and Economics 33, 255–272.

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

Trefethen, L.N., Bau, D., 1997. Numerical linear algebra. volume 50. Siam.

Wood, S.N., 1994. Monotonic smoothing splines fitted by cross validation. SIAM Journal on Scientific Computing 15, 1126–1133.