Boosted mortality models with age and spatial shrinkage

Li Li[*], Han Li[†], Anastasios Panagiotelis[‡]

June 18, 2023

Abstract

This paper extends the technique of gradient boosting in mortality forecasting. The two novel contributions are to use stochastic mortality models as weak learners in gradient boosting rather than trees, and to include a penalty that shrinks the forecasts of mortality in adjacent age groups and nearby geographical regions closer together. The proposed method demonstrates superior forecasting performance based on US male mortality data from 1969 to 2019. The boosted model with age-based shrinkage yields the most accurate national-level mortality forecast. For state-level forecasts, spatial shrinkage provides further improvement in accuracy in addition to the benefits achieved by age-based shrinkage. This additional improvement can be attributed to data sharing across states with both large and small populations in adjacent regions, as well as states which share common risk factors.

Keywords: Gradient boosting; Mortality; Forecasting; Shrinkage; Spatial modeling.

[*]School of Economics and Management, Beihang University, China. Email address: by1908007@buaa.edu.cn.

[†]Centre for Actuarial Studies, Department of Economics, The University of Melbourne, Australia. Email address: han.li@unimelb.edu.au.

[‡][Corresponding author] Discipline of Business Analytics, The University of Sydney, Australia. Email address: anastasios.panagiotelis@sydney.edu.au.

1 Introduction

Projections of future mortality levels play a crucial role in the decision-making processes of insurance companies, health service providers, and government agencies. Mortality data exhibit strong patterns across different age groups as well as geographical regions. Models for forecasting mortality are tailored towards exploiting these patterns. However, while such models may work well overall, they can lack accuracy for specific age groups and/or regions, especially those with low population exposure. To address this issue, we extend the mortality forecasting literature in two ways. First, we adapt the machine learning technique of gradient boosting by using popular mortality models as weak learners rather than trees. Second, within this boosting framework, we exploit the structure in mortality data by shrinking forecasts corresponding to similar age groups and neighbouring regions closer together.

The development of mortality models has been ongoing for centuries, and the rich literature encompasses both deterministic mortality models and stochastic mortality models (for a review, see Booth and Tickle, 2008). Stochastic mortality models have become increasingly popular since the introduction of the Lee–Carter model (Lee and Carter, 1992) and other similarly motivated models (see, e.g. Li and Lee, 2005; Cairns *et al.*, 2006; Hyndman and Ullah, 2007; Plat, 2009). These models have been successfully applied in both single-population and multi-population settings. However, most stochastic models impose restrictions on the functional form of the age and time structure of the data, which may only suit the mortality experience of some, but not all, age groups of the population (Cairns *et al.*, 2009; Li *et al.*, 2016, 2017; SriDaran *et al.*, 2022). Another challenge in mortality modeling is to accurately and efficiently forecast mortality rates across a large number of populations. Recent research has investigated this topic and identified common mortality trends across countries (for a review, see Enchev *et al.*, 2017). Furthermore, researchers have utilized information from adjacent regions in multi-population forecasting to improve overall accuracy (see e.g. Cupido *et al.*, 2020; Lin and Tsai, 2022).

As a popular machine learning technique, boosting has been increasingly adopted for forecasting in general (see Januschowski *et al.*, 2022, for the success of gradient boosting in the M5 competition), in actuarial science (see Lee and Lin, 2018) and more specifically in modeling and forecasting mortality (for a review on machine learning techniques in mortality modeling, refer to Richman, 2021). Loosely speaking, gradient boosting fits a so-called "weak learner" to the data, computes (pseudo) residuals, fits the same weak learner to (pseudo) residuals, continuing this process in an iterative fashion. The literature on using boosting together with stochastic mortality models is quite sparse with some notable extensions. Deprez *et al.* (2017) and Levantesi and Pizzorusso (2019) use a tree-based boosting approach for forecasting mortality rates and for backtesting stochastic mortality models. They fit a Poisson regression model on death counts, with the mean being a product of fitted mortality rates from a stochastic model and a term that is trained by tree-based boosting. Bjerre (2022) compare so-called "pure" gradient boosting models with a two-stage approach that applies tree-based gradient boosting to the residuals from a stochastic mortality model. The main finding is that the pure models have superior forecasting performance.

A common feature of these aforementioned studies is that boosting is always carried out using *trees* as weak learners. While trees are a popular choice of weak learners, boosting can be used in conjunction with other models including logistic regression (Friedman *et al.*, 2000), generalized additive

models (Tutz and Binder, 2006), and more recently copulas (Brant and Haff, 2022).[1] The novelty of our approach is in using mortality models - specifically the Lee–Carter model - instead of trees as the weak learner in boosting. A weakness of the Lee–Carter model is that mortality rates for all age groups are proportional to a single time trend. By fitting the Lee–Carter model repeatedly under a boosting framework, we capture mortality trends for age groups and regions that were poorly estimated in previous fits. The final forecast is based on an ensemble of Lee–Carter models. In this way we create a boosting method catered towards the domain of mortality forecasting.

The analysis of spatial data has been extensively studied in statistics (for a review we refer readers to the monograph of Gelfand *et al.*, 2010). Methods in spatial statistics exploit the neighbourhood structure of data. In our setting, this corresponds to geographical neighbours (more specifically US states that share a border) as well as neighbouring age groups (age groups that differ by exactly one year). In the proposed boosting framework, a penalty term that shrinks the forecasts of "neighbouring" mortality rates closer together is added to the objective function. The penalty term depends on the graph Laplacian matrix, to be defined in Section 2. This matrix has been used elsewhere in mortality modelling, by Arató *et al.* (2006) who use it as a prior covariance for random effects in a hierarchical Bayesian model for mortality rates, by Cupido *et al.* (2020) who spatially filter mortality rates of different US counties, and by Huynh and Ludkovski (2021) who model multi-population mortality models with Gaussian processes. Our approach differs from existing work in that it carries out spatial shrinkage in a boosting framework. To the best of our knowledge, our paper is the first to use spatial shrinkage together with boosting, not just in mortality modeling but for any application.

We apply the proposed boosted mortality models to US male mortality data for ages 0–85+, over the period 1969–2019. For both national-level and state-level mortality rates, our boosting approach substantially improves forecast accuracy over benchmark mortality models. For national-level mortality rates, the empirical results demonstrate superior forecasting performance of the boosted model with age-based shrinkage (later on referred to as the "GBLC-age model"). In a multi-population setting, we apply the boosting approach to the US state-level mortality data with both age- and state-based shrinkage incorporated in the model (later on referred to as the "GBLC-age-state model"). Our results show that in addition to the improvement achieved by age-based shrinkage, state-based shrinkage provides an additional enhancement to forecast accuracy, owing to "borrowing" information from neighboring states. The additional benefit from state shrinkage is particularly pronounced in states with sparse population or those subject to common risk factors. It should be noted that the proposed boosting approach with shrinkage is readily applicable to other multi-dimensional forecasting problems where the data structure can be further utilized.

The rest of this paper is organized as follows. Section 2 introduces the new boosting and shrinkage methodologies we propose to forecast mortality rates. Section 3 describes and visualizes the US mortality data used in this research. In Section 4, we present empirical results at both national level and state level based on the data described in Section 3. Section 5 concludes the paper.

2 Methodology

In this section we outline our novel methodology for forecasting using boosting and shrinkage. Section 2.1 describes the Lee–Carter model, which is used as a weak learner in our boosting algorithm.

[1] Also the popular R package `caret` (Kuhn and Max, 2008) implements boosting with GAMs, GLMs, smoothing splines, and neural networks.

Section 2.2 discusses the gradient boosting algorithm including modifications for the mortality forecasting setting. Section 2.3 deals with the case of single-population mortality forecasting, such as national-level mortality for the US. A penalty shrinking the forecasts of "neighbouring" time series together is incorporated into the boosting framework, where neighbouring time series correspond to mortality rates for age groups that differ by one year. In Section 2.4, we describe how the proposed algorithm can be extended to multi-population mortality forecasting, where shrinkage is based on age as well as the geography of the regions.

2.1 The Lee–Carter model

Introduced in the early 1990s, the Lee–Carter model (Lee and Carter, 1992) is regarded as one of the most significant mortality models, inspiring an era of stochastic mortality modeling. The model is formulated as follows

$$\log(m_{x;t}) = a_x + b_x \ _t + \ _{x;t}; \tag{1}$$

for $x = 1;:::;N$ and $t = 1;:::;T$, where $m_{x;t}$ is the mortality rate for age x in year t, a_x and b_x respectively are age-specific intercepts and coefficients, $\ _t$ represents a time-varying factor that is common to all ages, and $\ _{x;t}$ is the error term. To ensure identification, the model is estimated under the following constraints

$$\sum_{x=1}^{N} b_x = 1; \ \sum_{t=1}^{T} \ _t = 0:$$

With these restrictions, it is natural to interpret a_x as average log mortality over time, a measure of baseline mortality for a given age. The coefficient b_x describes the rate of decline of mortality at age x in response to changes in $\ _t$. The bilinear term $b_x \ _t$ captures the mortality improvement over time for age x. However, the Lee–Carter model only includes one time-varying factor $\ _t$, resulting in a trivial correlation structure in mortality improvement across different ages.

There has been a rich body of work on mortality modeling and forecasting over the last few decades, which includes various extensions of the Lee–Carter model or similar approaches (see e.g. Lee, 2000; Brouhns *et al.*, 2002; Cairns *et al.*, 2006; Renshaw and Haberman, 2006; Hyndman and Ullah, 2007; Plat, 2009, amongst others). Despite its limitations, the Lee–Carter model has been widely recognized and applied in the fields of actuarial science and demography and to date, the model is frequently used as a benchmark in mortality research. In the paper, we adopt the Lee–Carter model as a weak learner for forecasting with gradient boosting.

Estimation of the Lee–Carter model is conducted as follows. First let $y_{x;t} := \log(m_{x;t})$ and let $\mathbf{Y}$ denote an $N \quad T$ matrix with $y_{x;t}$ in row x and column t (the same convention is used to denote all matrices throughout the paper). Each age-specific intercept is estimated by the sample mean over time of log mortality rates for that age group. The age-specific intercepts are then subtracted from the log mortality rates and stacked in a matrix $\tilde{\mathbf{Y}}$. The singular value decomposition of $\tilde{\mathbf{Y}}$ yields the estimates of the age-specific coefficients b_x and the time trend $\hat{} = (\hat{}_1;:::;\hat{}_T)^0$. This is summarized in Algorithm 1. In addition to the model parameters, the fitted values $\hat{y}_{x;t}$ are also required for boosting. Note that although the Lee–Carter is a linear model, the forecasts depend on the singular value decomposition of $\tilde{\mathbf{Y}}$, which is a highly non-linear transformation of the input matrix. In our empirical work, Algorithm 1 is implemented using the R package `demography` (Hyndman, 2023).

Algorithm 1 Estimating the Lee–Carter model

Input An $N \quad T$ **matrix** $\mathbf{Y}$

Output $\hat{\mathbf{y}}_t$ **for** $t = 1; 2; :::; T$, $\hat{} = (\ _1; :::; \ _T)^{\prime}$, $\hat{\mathbf{a}} = (\hat{a}_1; :::; \hat{a}_N)^{\prime}$ **and** $\hat{\mathbf{b}} = (\hat{b}_1; :::; \hat{b}_N)^{\prime}$

1: **procedure** LC($\mathbf{Y}$)

2: $\hat{a}_x \quad \frac{1}{T}\sum_{t=1}^{T} y_{x;t}; 8x.$

3: $\tilde{y}_{x;t} \quad y_{x;t} \quad \hat{a}_x; 8x; t$ and stack into a matrix $\tilde{\mathbf{Y}} = (\tilde{\mathbf{y}}_1; :::; \tilde{\mathbf{y}}_T)$.

4: Carry out the singular value decomposition of $\tilde{\mathbf{Y}}$.

5: $\hat{} \quad s\mathbf{u}\sum_x v_x$, where $\hat{}$ is the estimated time trend, s, $\mathbf{u}$ and $\mathbf{v}$ are the first singular value, left singular vector and the right singular vector of $\tilde{\mathbf{Y}}$ respectively.

6: $\hat{\mathbf{b}} \quad \frac{1}{\sum_x v_x}\mathbf{v}; 8x.$

7: $\hat{y}_{x;t} \quad \hat{a}_x + \hat{b}_x\hat{}_t$, for $t = 1; 2; :::; T; 8x.$

2.2 Gradient boosting

Boosting (Breiman, 1997; Friedman, 2001) is a popular method in machine learning for regression and classification, with modern variants having been seen considerable success in forecasting applications. Boosting requires the use of a non-linear prediction, most typically from a tree, as a weak learner. Boosting aims to find an ensemble of weak learners $\mathbf{f}_t = \sum_{l=0}^{j} \ _l\hat{\mathbf{y}}_t^{(l)}$, where $\hat{\mathbf{y}}_t^{(l)}$ is the prediction from the l^{th} weak learner and $\ _l$ is the corresponding coefficient. We consider a quadratic loss function

$$L(\mathbf{y}_t; \mathbf{f}_t) = \sum_{t=1}^{T}(\mathbf{y}_t \quad \mathbf{f}_t)^{\prime}(\mathbf{y}_t \quad \mathbf{f}_t)\,; \tag{2}$$

Other loss functions can be considered, for example, if the objective is quantile prediction then pinball loss can be used.

Our algorithm for boosting follows the process outlined in Friedman (2001), adapted so that the Lee–Carter model is used as a weak learner and to account for a vector-valued response as in Equation 2. Boosting fits the weak learner to the gradient of the loss function taken with respect to $\mathbf{f}_t$ in an iterative fashion. For the case of a quadratic loss, this implies fitting the weak learner to the residuals, finding the coefficient for each weak learner in the ensemble using one-dimensional optimisation[2], and computing new residuals and iterating this procedure. We set the maximum number of iterations to $J = 50$, while the algorithm also stops if the change in the value of the loss function is less than a small positive number $" = 10^{\ 8}$. This procedure is summarized in Algorithm 2, and throughout the remainder of the paper it is referred to as the Gradient Boosted Lee–Carter (GBLC) model.

To forecast with a Lee–Carter model, it is common to fit a time series model to the estimated $\hat{}$. Then forecasts $\hat{}_{T+hjT}$ can be found where the subscript denotes a h-step ahead forecast made using information up to time T. Forecasts of log mortality are computed as $\hat{y}_{x;T+hjT} = \hat{a}_x + \hat{b}_x\hat{}_{T+hjT}$ for $x = 1; :::; N$. When an ensemble from a boosted Lee–Carter model is available, forecasts of $\hat{}^{(l)}_{T+hjT}$

[2]This is implemented using the `optim` package in R.

Algorithm 2 Gradient Boosted Lee–Carter forecasting

Input $\mathbf{Y}$

Output $\hat{}^{(l)}$, $\hat{\mathbf{a}}^{(l)}$,$\hat{\mathbf{b}}^{(l)}$ **and** $_l$ **for** $l = 0;:::;j$

procedure GBLC($\mathbf{Y}$)

  $\hat{\mathbf{y}}_t^{(0)};\hat{\mathbf{a}}^{(0)};\hat{\mathbf{b}}^{(0)}$ $LC(\mathbf{Y})$ for $t = 1;:::;T$

  $_0 = \arg\min L(\mathbf{y}_t;\hat{\mathbf{y}}_t^{(0)})$

  $\mathbf{z}_t^{(0)}$ $\mathbf{y}_t$ $_0\hat{\mathbf{y}}_t^{(0)}$ for $t = 1;:::;T$

  $\mathbf{Z}^{(0)}$ $(\mathbf{z}_1^{(0)};\mathbf{z}_2^{(0)};:::;\mathbf{z}_T^{(0)})$

  $L^{(0)}$ $\sum_{t=1}^{T}(\mathbf{y}_t \quad {}_0\hat{\mathbf{y}}_t^{(0)})'(\mathbf{y}_t \quad {}_0\hat{\mathbf{y}}_t^{(0)})$

  j 0

  while j J or $\left|L^{(j)} \quad L^{(j+1)}\right| >$ " **do**

    $\hat{\mathbf{y}}_t^{(j+1)};\hat{\mathbf{a}}^{(j+1)};\hat{\mathbf{b}}^{(j+1)}$ $LC(\mathbf{Z}^{(j)})$ for $t = 1;:::;T$

    $_{j+1} = \arg\min L(\mathbf{z}_t^{(j)};\hat{\mathbf{y}}_t^{(j+1)})$

    $\mathbf{z}_t^{(j+1)}$ $\mathbf{z}_t^{(j)}$ $_{j+1}\hat{\mathbf{y}}_t^{(j+1)}$

    $\mathbf{Z}^{(j+1)}$ $(\mathbf{z}_1^{(j+1)};\mathbf{z}_2^{(j+1)};:::;\mathbf{z}_T^{(j+1)})$

    $L^{(j+1)}$ $\sum_{t=1}^{T}(\mathbf{z}_t^{(j)} \quad {}_{j+1}\hat{\mathbf{y}}_t^{(j+1)})'(\mathbf{z}_t^{(j)} \quad {}_{j+1}\hat{\mathbf{y}}_t^{(j+1)})$

    j $j+1$

can be produced using a different time series model for each l and the ensemble forecast is given by

$$\sum_{l=0}^{j} {}_l\left(\hat{a}_x^{(l)} + \hat{b}_x^{(l)}\hat{}_{T+hjT}^{(l)}\right); \text{ for } x = 1;:::;N:$$

In this paper, the forecasting model used for $\hat{}$ is a random walk with drift implemented with the `rwf` in the R package `forecast` (Hyndman *et al.*, 2020), although other time series models, such as an ARIMA, could also be used.

2.3 Boosted Lee–Carter with age shrinkage

A distinct characteristic of mortality data is the age structure of the data. This structure should be taken into account in forecasting, for example, the mortality forecast for the 50 year age group should tend to be similar to those for the 49 and 51 year age groups. In order to shrink these forecasts closer together we replace the loss function in Equation 2 with the following objective function

$$L(\mathbf{y}_t;\mathbf{f}_t) = \sum_{t=1}^{T}(\mathbf{y}_t \quad \mathbf{f}_t)'(\mathbf{y}_t \quad \mathbf{f}_t) + \quad \mathbf{f}_t'\mathbf{W}\mathbf{f}_t; \tag{3}$$

where is a shrinkage parameter found by cross validation and $\mathbf{W}$ is a shrinkage matrix. Algorithm 2 remains mostly the same, however, rather than fit a Lee–Carter model to the residuals, line 4 is replaced by

$$\mathbf{z}_t^{(0)} \quad \mathbf{y}_t \quad {}_0\mathbf{f}_t^{(0)} \quad 2 \ \mathbf{W}\mathbf{f}_t^{(0)};$$

and Algorithm 2 line 11 is similarly changed as

$$\mathbf{z}_t^{(j+1)} \quad \mathbf{z}_t^{(j)} \quad {}_{j+1}\mathbf{f}_t^{(j+1)} \quad 2 \ \mathbf{W}\mathbf{f}_t^{(j+1)}.$$

Age level shrinkage is achieved by setting $\mathbf{W}$ to the following matrix

$$\mathbf{W}^{(\text{age})} = \begin{pmatrix} 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 1 & 0 & & & \vdots \\ 0 & 1 & 2 & 1 & 0 & & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & 1 & 2 & 1 & 0 \\ \vdots & & & 0 & 1 & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & 0 \end{pmatrix};$$

where the superscript (age) denotes that the matrix shrinks across neighbouring age groups. In this case we have

$$\mathbf{W}^{(\text{age})}\mathbf{f}_t^{(j)} = \begin{pmatrix} 0 & & \\ & (f_{2;t}^{(j)} \quad f_{3;t}^{(j)}) & \\ & (f_{3;t}^{(j)} \quad f_{2;t}^{(j)} + f_{3;t}^{(j)} \quad f_{4;t}^{(j)}) & \\ \vdots & & \\ & (f_{N\ 1;t}^{(j)} \quad f_{N\ 2;t}^{(j)}) & \\ 0 & & \end{pmatrix}$$

At each step of the boosting algorithm, not only is the Lee–Carter model fitting residuals unexplained by previous Lee–Carter fits, but is also fitting any discrepancy that may exist between the prediction for an age group and the predictions for adjacent age groups. Note that age 0 (infant mortality) and the composite 85+ age are highly idiosyncratic therefore predictions for these groups are not shrunk at all, and the corresponding rows and columns of $\mathbf{W}$ are entirely made up of zeros.[3]

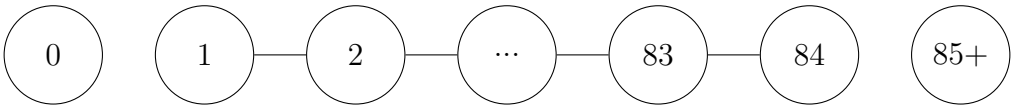


Figure 2.1: Graph representing neighbourhood structure of ages.

The structure of $\mathbf{W}$ can be understood through the graph in Figure 2.1. We can see that the "neighbors" are defined as ages that differ by exactly one year. Since we do not shrink the age 0 and 85+ groups at all, these nodes are disconnected from the rest of the graph. Ignoring the disconnected nodes, the graph has the simple structure of a chain. The matrix $\mathbf{W}^{(\text{age})}$ is the graph

[3]Shrinking the age 0 group to the age 1 group, and the age 85+ group towards the age 84 group was implemented, but led to a deterioration in forecast accuracy for these groups.

Laplacian, defined as the matrix with off-diagonal elements $w_{ik}^{(\text{age})} = \quad 1$ when node i and k are neighbours and $w_{ik}^{(\text{age})} = 0$ otherwise, and diagonal elements equal to the degree (or the number of neighbours) of each node. Each age group has two neighbours, hence the diagonal elements of $\mathbf{W}$ are all 2, with the exception of the ages of 0, 1, 84, 85+. While the connection to the graph Laplacian may seem superfluous for a simple structure like a chain, its usefulness becomes apparent for more complex graphs, such as a neighbourhood graph of US states illustrated in the next section.

2.4 Boosted Lee–Carter with age and state shrinkage

We now propose a multi-population version of the boosted Lee–Carter with age and state shrinkage. We first discuss how the neighbourhood graph is constructed for states and used to obtain a Laplacian $\mathbf{W}^{(\text{state})}$. Since our application is based on state-level mortality in the US, we use the term "*state*" rather than "*region*", although our methods could be generalized to any geographical units, for example, counties within a single state, or countries within the European Union.

Figure 2.2: Graph representing neighbourhood structure of selected US states.

Similar to the age-based shrinkage, the structure of $\mathbf{W}$ can be illustrated by a graph. The graph is constructed with each node corresponding to a state, and two nodes sharing an edge if the corresponding states share a border. States sharing a border at a single point (as occurs, for example, between Arizona and Colorado) also share an edge in the graph and are treated as neighbours. In our study, all 50 states as well as the District of Columbia (DC) are included in the shrinkage. The inclusion of Alaska and Hawaii makes the graph a disconnected graph, implying no shrinkage for these states. The graph Laplacian for 50 states and DC is too large to present here. Therefore, a graph for a subset of seven states is shown in Figure 2.2 for illustrative purposes. The Laplacian of these states is shown as follows, where state labels[4] are given in the rows and columns and the data and forecasts for the states would need to be stacked in this order for shrinkage to have the intended effect.

$$\mathbf{W}^{(\text{state})} = \begin{array}{c} \\ OH \\ PA \\ NY \\ WV \\ NJ \\ MD \\ DE \end{array} \begin{array}{c} \begin{array}{ccccccc} OH & PA & NY & WV & NJ & MD & DE \end{array} \\ \begin{pmatrix} 2 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 6 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 3 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 3 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 3 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 3 \end{pmatrix} \end{array}$$

[4]For a full list of US state abbreviations and names, please refer to Table A.1 of the Appendix.

With the construction of $\mathbf{W}^{(\text{state})}$ established, the model in Section 2.3 can now be extended to the multi-population case. Let $y_{x;t;i} := \log(m_{x;t;i})$ be the log mortality rate for age group x, time period t, and state i, for $x = 1;:::;N$, $t = 1;:::;T$, and $i = 1;:::;S$. Let $\mathbf{y}_{t;i} := (y_{1;t;i};:::;y_{N;t;i})^{0}$ denote an N-vector of log mortalities for state i at time t. We stack the observations into an NS-vector as follows

$$\mathbf{y}_t = \begin{pmatrix} \mathbf{y}_{t;1} \\ \mathbf{y}_{t;2} \\ \vdots \\ \mathbf{y}_{t;S} \end{pmatrix} :$$

Other quantities in Algorithm 1 and Algorithm 2, such as $\mathbf{f}_t$ are stacked in a similar fashion as $\mathbf{y}_t$ above, and for both algorithms a single Lee–Carter fit is replaced with separate Lee–Carter fits for each state.

To employ shrinkage, we must define a neighbourhood structure between nodes that correspond to age-state pairs. Let $N_{x;i}$ denote the node corresponding to age group x and state i. Two nodes $N_{x;i}$ $N_{x';i'}$ are connected by an edge if either $x = x^{0}$ and $i; i^{0}$ are neighbours, or if $x; x^{0}$ are neighbours and $i = i^{0}$. For example, age 50 mortality in New York and age 51 mortality in New York are neighbours. Also, age 50 mortality in New York and age 50 mortality in Pennsylvania are neighbours. However, age 50 mortality in New York and age 51 mortality in Pennsylvania are not neighbours.

A graph with the structure described above can be obtained by taking the *Cartesian product* of two graphs, which are the neighbourhood graphs for the age and state structure. The graph Laplacian for combining age and state shrinkage is given by

$$\mathbf{W}^{(\text{age-state})} = \mathbf{W}^{(\text{age})} \quad \mathbf{I}_{S \;\; S} + \mathbf{I}_{N \;\; N} \quad \mathbf{W}^{(\text{state})};$$

where $\mathbf{I}$ is an identity matrix and is the Kronecker product. While $\mathbf{W}^{(\text{age-state})}$ could be substituted directly into the stacked version of Equation 3, this would lead to a single shrinkage parameter controlling both age and state shrinkage, which we found to perform poorly in practice. To allow for different age and state shrinkage parameters, we consider the following objective function

$$L(\mathbf{y}_t; \mathbf{f}_t) = \sum_{t=1}^{T} (\mathbf{y}_t \quad \mathbf{f}_t)^{0} (\mathbf{y}_t \quad \mathbf{f}_t) + \mathbf{f}_t^{0} (\;\; {}_a\mathbf{W}^{(\text{age})} \quad \mathbf{I}_{S \;\; S} + \mathbf{I}_{N \;\; N} \quad\;\; {}_s\mathbf{W}^{(\text{state})}) \mathbf{f}_t \,; \tag{4}$$

where $_a$ and $_s$ are shrinkage terms for age and state respectively. These terms are set using cross-validation over a 2-dimensional grid. In our empirical analysis in Section 4, we consider four special cases of the objective function in Equation 4 as follows

- **GBLC:** No shrinkage, set $_a = \;\; _s = 0$;
- **GBLC-age:** Age shrinkage only, set $_s = 0$;
- **GBLC-state:** State shrinkage only, set $_a = 0$;
- **GBLC-age-state:** Age and state shrinkage, set $_a \neq 0$; $_s \neq 0$.

Note setting $_a = \;\; _s = 0$ is equivalent to fitting the GBLC model described in Section 2.2 to each state independently. The same holds for setting $_s = 0$ and the GBLC-age approach from Section 2.3.

3 Data

In this study, we consider age-specific male mortality rates in the US over the investigation period 1969–2019 for ages 0–85+, at both the national and state levels. Mortality data up to 2019 was used to avoid the potential impact of the COVID-19 pandemic on the results. For the state-level analysis, we collect data from all 50 states as well as DC.

For death count numbers, we collect the data from the National Center for Health Statistics (NCHS) for the period 1969–2004, and from the Centers for Disease Control and Prevention (CDC) WONDER online database for the period 2005–2019.[5] For the corresponding exposure data, information on the annual population is obtained from the Survey of Epidemiology and End Results (SEER), for ages from 0 to 85+.

3.1 National-level mortality

In Figure 3.1, we plot the national-level mortality rates for the US male population during 1969–2019. The various colors in the plot present different years of observation and illustrate an overall decreasing trend in the mortality levels, across all age groups. We can clearly observe an "accident hump" in the data which is made up of elevated mortality rates around at ages in the early 20s. Another interesting observation is that there seems to be a notable increase in the mortality level for ages 25–45, over the last 5 to 10 years of the investigation period (see discussions in Couillard *et al.*, 2021). Studies find that this rise in mortality among young and middle-aged adults is partly due to drug overdoses, alcohol, suicides, and cardio-metabolic conditions (Harris *et al.*, 2021). Overall, the total mortality rate series exhibit low variances in the data particularly at older ages, and show clear and smooth patterns in the age and time dimensions.

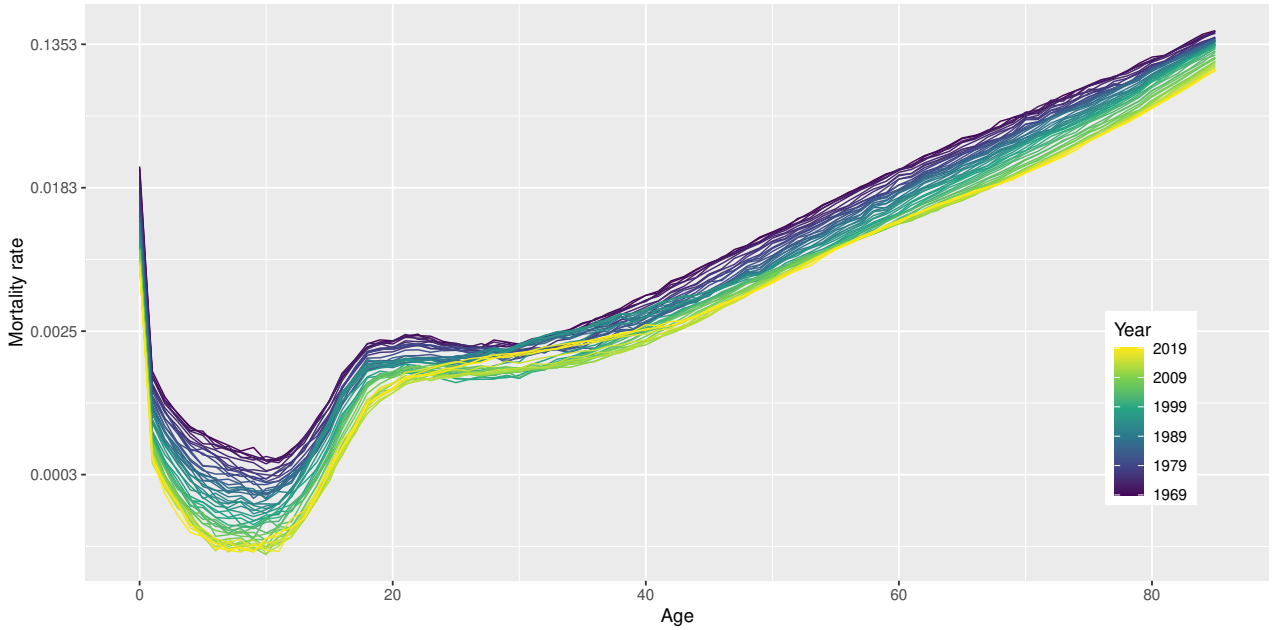


Figure 3.1: National-level male mortality for US: 1969–2019.

[5]Starting from 2005, the geographic identifier has been removed from the NCHS multiple causes of death database due to restrictions on the release of sub-national mortality data.

3.2 State-level mortality

On top of the national-level mortality, it is also important to investigate state-level mortality as it provides a more granular picture of mortality patterns and trends. This also helps to identify mortality disparities across different regions, pointing towards causal factors that contribute to the differences in mortality. To gain a better understanding of how mortality experiences have changed over the recent decades across the US, we examine snapshots of mortality rates in both 1999 and 2019. In Figure 3.2, we visualize the geographical variations in the US mortality rates for three representative age groups, namely, 20, 50, and 80. The left panel of Figure 3.2 illustrates mortality rates in 1999, while the right panel shows rates in 2019.

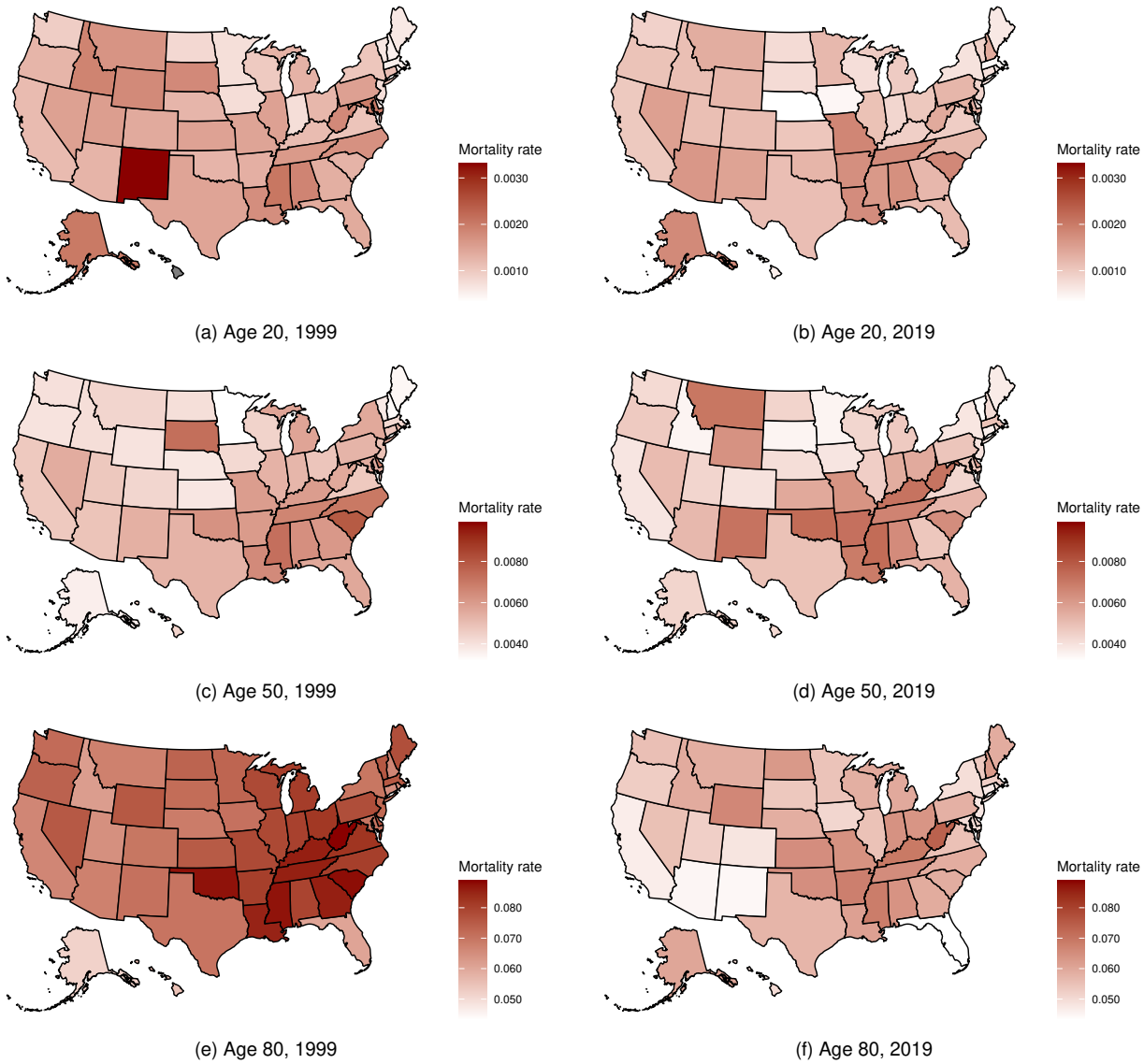


Figure 3.2: Male mortality rates across the US in 1999 (left panel) and 2019 (right panel).

From Figure 3.2, we can see an overall improvement in mortality rates across all states for the three age groups between 1999 and 2019. The improvement is particularly pronounced for age 80, as indicated by the much lightened color in the right panel compared to the left panel. From both panels, we can see that geographical variations in mortality rates are very prominent. These variations can be attributed to several factors, encompassing demographics, socio-economic status, and lifestyle-related behaviors. For example, a rich body of research has established the relationship

between mortality experience and socioeconomic determinants including income, education level, and unemployment rate (see *e.g.* Jemal *et al.*, 2008; Chapman *et al.*, 2010; Dwyer-Lindgren *et al.*, 2017; Woolf and Schoomaker, 2019; Lourés and Cairns, 2020). These socioeconomic factors can be used to explain the geographical variations in state-level mortality in the US. Studies have also found that the mortality disparities across the US have become more apparent over time (see e.g. Vierboom *et al.*, 2019; Woolf and Schoomaker, 2019; Couillard *et al.*, 2021). It can be argued that the mortality experience at the state-level demonstrates comparable geographical patterns across all three age groups. In general, East and West Coast states have lower mortality rates compared to Southern states and certain "rust belt"[6] states. More specifically, we have identified heavier mortality rates from Alabama, Arkansas, Georgia, Indiana, Kentucky, Louisiana, Mississippi, South Carolina, and Tennessee. These states are all part of the "stroke belt"[7] in the US. This observation is unsurprising as cardiovascular disease is the leading cause of death in the US.

To further investigate mortality patterns across various regions, we plot the mortality rates for selected states including California, Florida, Ohio, and Pennsylvania over the full period of 1969–2019 in Figure 3.3. All four states exhibit relatively smooth patterns in mortality rates, although the mortality profiles are slightly more jagged for the lower population states of Ohio and Pennsylvania. The latter part of the 2010s, saw an increase in mortality noticeable among younger to middle-aged individuals in Ohio and Pennsylvania; on the plot the yellow lines rise above the darker colored lines in the middle of the two lower panels. This may be attributed to the well documented "deaths of despair" (Case and Deaton, 2015) that was particularly pronounced in the rust belt. The presence of this common factor across states motivates the use of spatial shrinkage as described in Section 2.4.

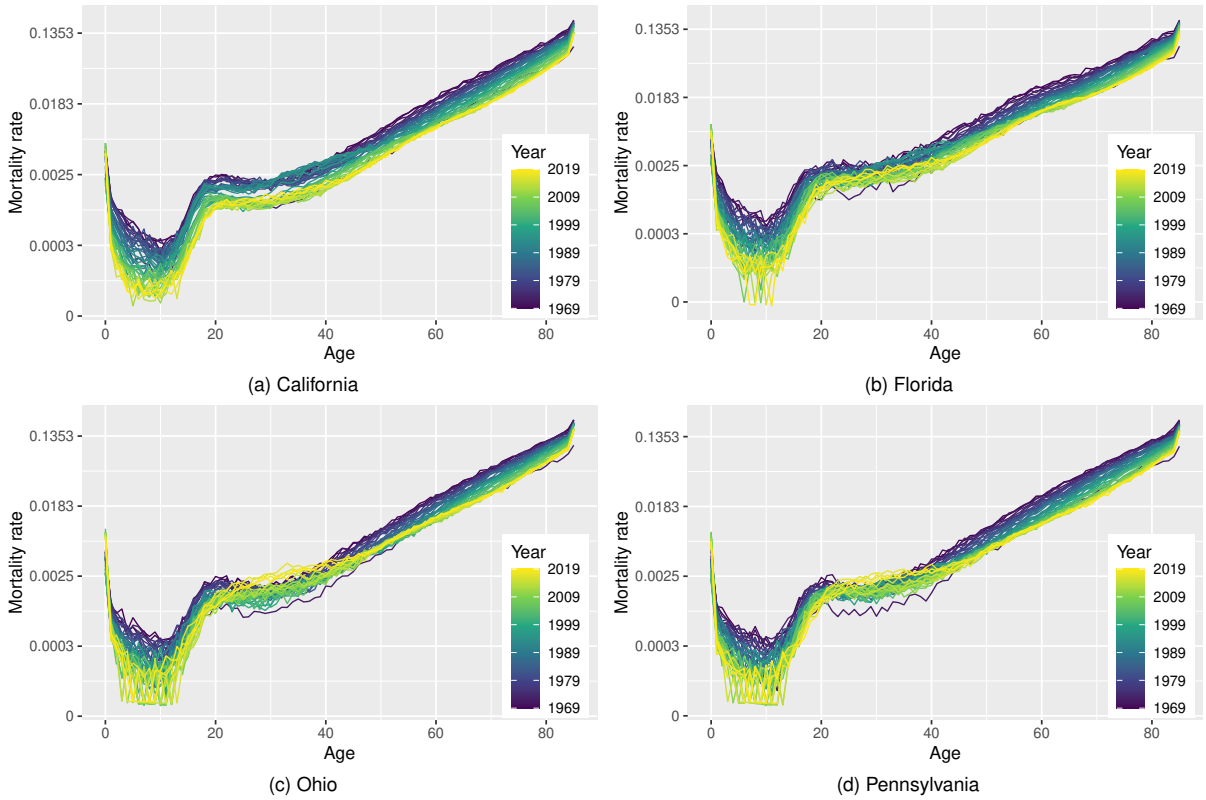


Figure 3.3: State-level male mortality for California, Florida, Ohio, and Pennsylvania: 1969–2019.

[6]Although definitions vary, the rust belt states are generally understood to include Illinois, Indiana, Michigan, Ohio, Pennsylvania, West Virginia, and Wisconsin.

[7]For a full list of stroke belt states, please refer to Parcha *et al.* (2021).

4 Empirical results

4.1 Setup

In this section, we present the empirical results based on the US male mortality data described in Section 3. Due to the time series nature of the data, we implement 13 expanding windows to generate forecasts for horizons of $h = 1; 2; \ldots; 10$. More specifically, our training sample for each expanding window is from 1969 to $1996; 1997; \ldots; 2009$. The subsequent ten years of mortality data after each training sample are used for testing purposes. We compare the model performance of the proposed GBLC, GBLC-age, GBLC-state, GBLC-age-state models against several benchmark models, including the Lee–Carter (L–C) model, the Hyndman–Ullah (H–U) model, and the Hyndman–Booth–Yasmeen (H–B–Y) model. As the L–C model has already been introduced in Section 2.1, in the following sections, we provide a brief overview of the H–U and H–B–Y models.

4.1.1 The Hyndman–Ullah model

Hyndman and Ullah (2007) proposed a generalized version of the Lee–Carter model which accommodates smoothness in the age dimension by employing nonparametric smoothing techniques. The model is formulated as follows

$$\log(m_{x;t}) = \quad (x) + \sum_{k=1}^{K} \quad _{t;k} \quad _{k}(x) + \quad _{x;t}; \tag{5}$$

where $\quad (x)$ represents the level of log mortality rates at age x. $\quad _{k}(x)$ refers to a set of orthonormal basis functions, $\quad _{t;k}$ are the corresponding coefficients for $k = 1; 2; \ldots; K$, and $\quad _{x;t}$ is the error term. To predict future mortality levels, the coefficients $\quad _{t;k}$ are fitted into ARIMA time series models. The H–U model is implemented by using `demography::fdm` in R for the following experiments. The value of K is set to be 6 as suggested by the package.

4.1.2 The Hyndman–Booth–Yasmeen model

In a multi-population mortality modeling setting, it is important to avoid long-run divergence in mortality forecasts using individually fitted models. Li and Lee (2005) proposed a coherent multi-population extension of the Lee–Carter model which includes common factors for all populations within a group of countries. In line with the research by Li and Lee (2005), Hyndman *et al.* (2013) proposed a product-ratio functional method to coherently forecast mortality rates across different populations. This model can be viewed as a generalization of the Li and Lee (2005) model, and a multi-population extension of the Hyndman and Ullah (2007) model. The Hyndman–Booth–Yasmeen model is presented as follows

$$\log(m_{x;t;i}) = \quad _{i}(x) + \sum_{k=1}^{K} \quad _{t;k} \quad _{k}(x) + \sum_{m=1}^{M} \quad _{t;m;i} \Psi_{m;i}(x) + \quad _{x;t;i}; \tag{6}$$

where $m_{x;t;i}$ is the mortality rate for age x in year t for population i. $\quad _{i}(x)$ represents the level of log mortality rates at age x for population i. $\quad _{k}(x)$ and $\Psi_{m;i}(x)$ are orthonormal basis functions, $\quad _{t;k}$ represents the common trend for all populations in the group, and $\quad _{t;m;i}$ denotes the population-specific time trend. $\quad _{x;t;i}$ is the error term. To ensure the coherence of mortality forecasts, we restrict $\quad _{t;m;i}$ to be stationary processes. The H–B–Y model can be implemented by using `demography::coherentfdm` in R. The values of K and M are both set to be 6.

4.2 National-level results

To evaluate forecast accuracy we choose the mean absolute scaled error (MASE) as the error measure. The MASE for h-step-ahead forecasts across all ages and expanding windows is defined as follows

$$\text{MASE} = \frac{1}{86 \quad h \quad r} \sum_{x=0}^{85+} \sum_{h=1}^{10} \sum_{r=1}^{13} \frac{j\hat{m}_{x;28+r+h} \quad m_{x;28+r+h}j}{\frac{1}{28+r \quad 1} \sum_{t=2}^{28+r} jm_{x;t} \quad m_{x;t \quad 1}j};$$

where 28 is the number of years in the first training sample 1969–1996. Our first study applies the proposed GBLC and GLBC-age models to mortality rates of a single population (*i.e.* the national-level male mortality), and we compare our results with those from the Lee–Carter model and the Hyndman–Ullah model. Table 4.1 presents the MASE of out-of-sample forecasts of national-level male mortality over the forecast horizon h from 1 to 10. The MASE values with the smallest values among the four models (representing the best-performing model) are highlighted in bold for each forecast horizon.

The results in Table 4.1 state that GBLC-age model has the best performance across all four models and all forecast horizons. In contrast, the L–C model provides the worst forecast accuracy, which is not surprising given its simple structure, with only one age-time interaction term included in the model. It is worth noting that by adding multiple orthonormal basis functions, the H–U model shows a considerable improvement in forecast accuracy over the L–C model, especially for shorter forecast horizons, although these improvements are modest compared to those achieved via boosting.

Table 4.1: MASE of out-of-sample forecasts for national-level male mortality

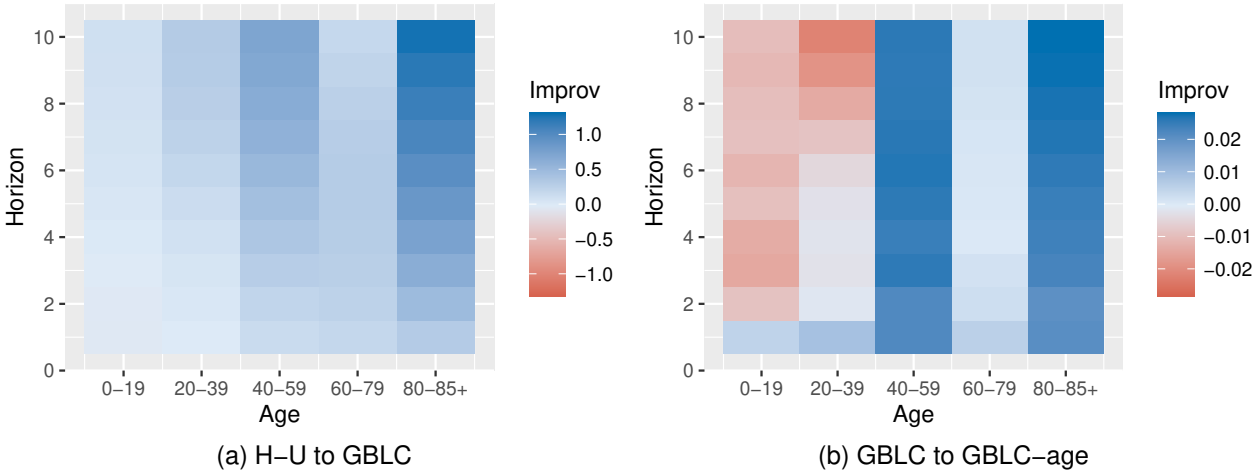
h	L–C	H–U	GBLC	GBLC-age
1	1.468	0.695	0.591	**0.580**
2	1.539	0.835	0.683	**0.678**
3	1.606	0.981	0.785	**0.781**
4	1.672	1.121	0.881	**0.877**
5	1.741	1.262	0.973	**0.968**
6	1.814	1.402	1.072	**1.068**
7	1.895	1.545	1.181	**1.177**
8	1.980	1.681	1.294	**1.290**
9	2.066	1.806	1.404	**1.402**
10	2.151	1.919	1.512	**1.511**

To assess whether the differences in MASE are statistically significant, we compute the model confidence sets introduced by Hansen *et al.* (2011). The model confidence set is analogous to a confidence interval, in the sense that over repeated samples, that the best forecasting method is included in the model confidence set with a predetermined probability. We consider a 95% confidence level for the model confidence set using the R package `MCS` (Bernardi, 2017). The results are shown in Table 4.2, where the tick indicates that the forecasts lie inside the model confidence set. It can be seen that GBLC and GBLC-age perform much better than the L–C and H–U models. In particular, forecasts from the GBLC-age model lie in the model confidence sets for all forecast horizons. On the other hand, the L–C and H–U models are always outside of the model confidence sets. Clearly, the GBLC-age model gives the best performance across all models.

Table 4.2: MCS results for national-level male mortality

h	L–C	H–U	GBLC	GBLC-age
1				X
2			X	X
3			X	X
4			X	X
5				X
6			X	X
7			X	X
8			X	X
9			X	X
10			X	X

To gain a better understanding of how our proposed models enhance forecast accuracy across different age groups, in Figure 4.1 we present improvement in MASE by 5 broad age groups, namely, 0–19, 20–39, 40–59, 60–79, and 80–85+. The H–U model is used as the benchmark model. Panel (a) plots the MASE improvement from the H–U model to the GBLC model. We observe an overall improvement across all age groups, over almost all forecast horizons. In particular, the improvement is more pronounced for the older age group 80–85+ over longer forecast horizons. Panel (b) presents the improvement from the GBLC model to the GBLC-age model. Some mixed results are observed over age groups: while an improvement in forecast accuracy is achieved in middle to older age groups, especially in ages 40–59 and 80–85+, MASE values slightly worsen in younger age groups below 40. However, these changes in MASE yield an overall improvement from the GBLC model to the GBLC-age model, as stated in Table 4.1. It should be noted that the magnitude of these changes is relatively small compared to panel (a). We then plot the improvement in MASE from the H–U model to GBLC-age in Figure 4.2. The improvement patterns across age groups can be shown to be very similar to those in panel (a) of Figure 4.1.



(a) H–U to GBLC (b) GBLC to GBLC–age

Figure 4.1: Improvement in MASE for national-level male mortality, a) from H–U to GBLC, b) from GBLC to GBLC-age.

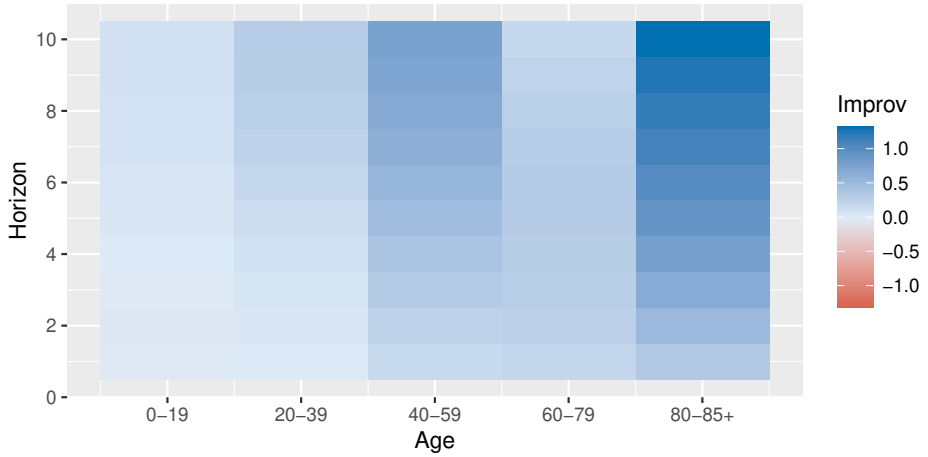


Figure 4.2: Improvement in MASE for national-level male mortality, from H–U to GBLC-age.

4.3 State-level results

In the second study, we apply the proposed GBLC, GLBC-age, GLBC-state, and GLBC-age-state models to forecast multi-population mortality rates (*i.e.* the state-level mortality). We compare our results with those from the Lee–Carter model, the Hyndman–Ullah model, and the Hyndman–Booth–Yasmeen model. Since both age and state shrinkages are considered in this exercise, we aim to disentangle the two effects by visualizing the MASE improvement over different models. To gain more insights on the spatial relationship of mortality across states, we also produce choropleth maps to visualize how improvement is achieved via state shrinkage.

Table 4.3: MASE of out-of-sample forecasts for state-level male mortality

h	L–C	H–U	H–B–Y	GBLC	GBLC-age	GBLC-state	GBLC-age-state
1	0.743	0.657	0.609	0.614	0.592	0.611	**0.590**
2	0.761	0.677	0.636	0.632	0.610	0.630	**0.608**
3	0.778	0.698	0.666	0.652	0.628	0.649	**0.626**
4	0.796	0.719	0.694	0.675	0.650	0.672	**0.647**
5	0.813	0.741	0.716	0.697	0.671	0.694	**0.668**
6	0.830	0.762	0.738	0.722	0.695	0.718	**0.691**
7	0.848	0.783	0.761	0.748	0.720	0.744	**0.717**
8	0.867	0.805	0.782	0.776	0.747	0.772	**0.743**
9	0.885	0.825	0.801	0.803	0.773	0.799	**0.769**
10	0.902	0.844	0.818	0.829	0.798	0.824	**0.793**

Table 4.3 presents the MASE values of out-of-sample forecasts of state-level male mortality across h from 1 to 10. Once again, we have highlighted in bold the smallest MASE values among the four models, representing the best-performing model, for each forecast horizon. The results show that the GBLC-age-state model has the best forecasting performance over all forecast horizons. First, it should be noted that we observe consistent improvements from the L–C model to the H–U model, and from the H–U model to the H–B–Y model. This observation is unsurprising, as the H–U model improves upon the L–C model via additional age-time factors, and the H–B–Y model ensures coherence in multi-population forecasts. On the other hand, the GBLC model provides comparable but slightly better results compared to the H–B–Y. It can be argued that improvement

achieved from age shrinkage is more substantial than the stage shrinkage. However, once both age and state shrinkages are incorporated in the model, the model performance is enhanced even further.

In Table 4.4, we present the MCS results for all 7 models, across different forecast horizons. It can be seen that the GBLC-age-state model is the only model that lies within the model confidence sets for $h = 1$ to 7. In the case of $h = 8$, the H–B–Y model, the GBLC-age model, and the GBLC-age-state model all provide forecasts within the model confidence sets. Notably, for longer forecast horizons of $h = 9$ and 10, the H–B–Y model provides the best results across all models, demonstrating its strong forecast performance in longer terms.

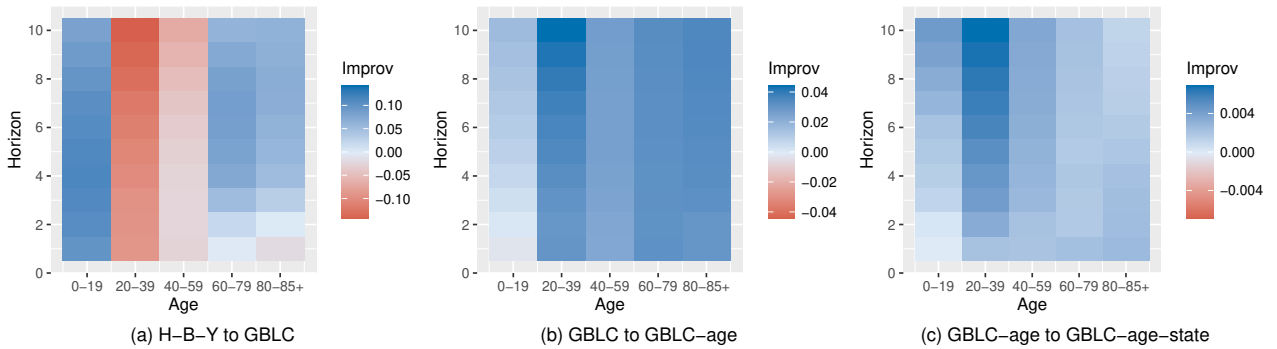
Table 4.4: MCS results for state-level male mortality

h	L–C	H–U	H–B–Y	GBLC	GBLC-age	GBLC-state	GBLC-age-state
1							X
2							X
3							X
4							X
5							X
6							X
7							X
8			X		X		X
9			X				
10			X				

Similar to the national-level exercise, we present improvement in MASE by 5 broad age groups in Figure 4.3. The results are based on average across all states. We use H–B–Y as the benchmark model. Panel (a) shows MASE improvement from the H–B–Y model to the GBLC model. We can see that the forecast accuracy for younger ages below 20, and for older ages above 60 has been improved by the gradient boosting method. For age groups 40–59 and particularly 20–39, there has been some deterioration in forecast accuracy. Panel (b) presents the improvement achieved by the GBLC-age model over the GBLC model. Across the five age groups, we observe improvements in MASE, with the exception of ages 0–19, over shorter forecast horizons. The improvement seems to be particularly strong among ages 20–29 over longer forecast horizons. Panel (c) demonstrates additional improvement achieved by adding the state shrinkage. Again, the biggest improvement lies in age group 20–29. Figure 4.4 plots the aggregated MASE improvement from the H–B–Y model to the GBLC-age-state model. While a slight decrease in accuracy still persists for ages 20–59, the magnitude of the decrease has been significantly lower, as indicated by the shade of red color used in Figure 4.3 Panel (a) and Figure 4.4. These visualizations are consistent with the findings and conclusions made earlier based on Table 4.3, where the overall forecast accuracy across age groups is improved via gradient boosting together with age and stage shrinkages.

As mentioned before, another aim of this research is to uncover the spatial relationship of mortality rates across different regions. Understanding how state shrinkage improves forecast accuracy is an important step towards understanding how the mortality experience of neighbouring states are related to one another. In Table 4.5, we present the MASE results for selected states for $h = 10$. The left side of the table ranks the results according to the improvement in MASE from GBLC to GBLC-state while the right side does the same for GBLC-age to GBLC-age-state[8].

[8]For brevity only the ten "most improved" and ten "least improved" states are shown in Table 4.5. A full list of results is available in Table A.2 of the Appendix.



(a) H–B–Y to GBLC (b) GBLC to GBLC–age (c) GBLC–age to GBLC–age–state

Figure 4.3: Improvement in MASE for state-level male mortality, a) from H–U to GBLC, b) from GBLC to GBLC-age, c) from GBLC-age to GBLC-age-state.

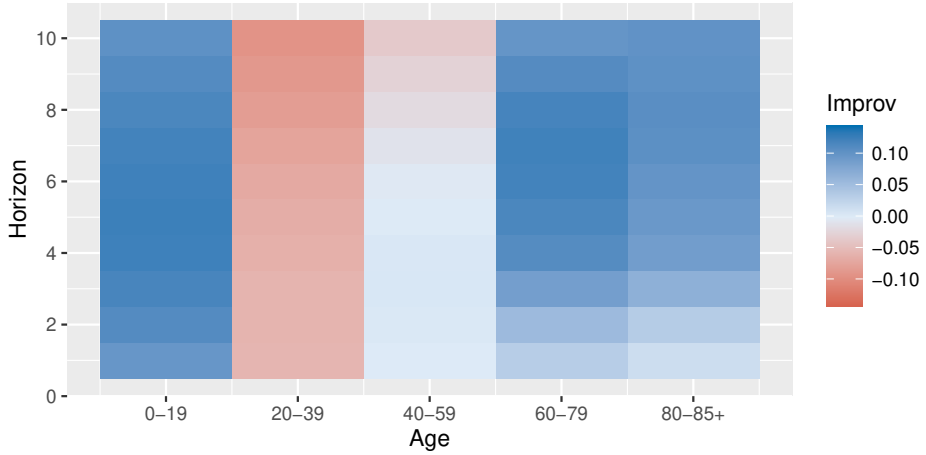


Figure 4.4: Improvement in MASE for state-level male mortality from H–B–Y to GBLC-age-state.

Table 4.5: MASE of out-of-sample forecasts for selected states, h=10, males

Rank	State	GBLC	GBLC-state	Imprv	State	GBLC-age	GBLC-age-state	Imprv
1	OH	1.015	1.005	0.011	OH	1.002	0.990	0.012
2	NE	0.824	0.815	0.010	ID	0.724	0.712	0.012
3	UT	0.734	0.725	0.009	CO	0.696	0.684	0.012
4	MA	0.799	0.790	0.009	DE	0.882	0.870	0.012
5	WY	0.962	0.953	0.009	IA	0.779	0.768	0.011
6	CO	0.714	0.706	0.009	NE	0.782	0.772	0.010
7	GA	0.720	0.712	0.008	GA	0.693	0.684	0.009
8	MO	0.840	0.832	0.008	MN	0.695	0.685	0.009
9	IA	0.808	0.801	0.007	WI	0.771	0.762	0.009
10	ID	0.755	0.748	0.007	TN	0.833	0.824	0.009
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
41	WV	0.994	0.992	0.002	AK	0.769	0.769	0.000
42	MI	0.813	0.811	0.001	HI	0.910	0.910	0.000
43	RI	0.945	0.944	0.001	MS	0.776	0.777	-0.001
44	ME	0.960	0.960	0.001	FL	0.741	0.742	-0.001
45	SC	0.742	0.742	0.000	TX	0.667	0.669	-0.002
46	AK	0.824	0.824	0.000	RI	0.928	0.930	-0.002
47	HI	0.957	0.957	0.000	CA	0.642	0.644	-0.002
48	VA	0.728	0.729	0.000	MI	0.784	0.786	-0.002
49	FL	0.747	0.748	-0.001	AZ	0.690	0.693	-0.003
50	MD	0.828	0.831	-0.002	VA	0.698	0.702	-0.004

We observe that Ohio (OH), Colorado (CO), Georgia (GA), Iowa (IA), and Idaho (ID) are ranked in the top 10 on both sides of the table. On the other hand, Michigan (MI), Rhode Island (RI), and Virginia (VA) are ranked in the bottom 10 on both sides of the table.[9]

Figures 4.5 and 4.6 show the same information as Table 4.5 (improvement due to state-based shrinkage) for all states using a choropleth for the case where $h = 10$. In Figure 4.5, we can see that besides an overall elevation of forecast performance across all states, the state shrinkage works particularly well in the states of the Northern Great Plains such as Wyoming, Nebraska, Idaho, and Colorado, and the states of the Rust Belt such as Ohio and Wisconsin. The Northern Great Plains is a region with relatively sparse population in the US, which potentially explains the stronger improvement of state shrinkage by "borrowing" information from nearby states with a larger population. As discussed in Section 3, over the evaluation period, the Rust Belt states suffered high rates of "deaths of despair" and this common local factor may explain the gains from state-based shrinkage being particularly strong in these states. To sum up, improvement in MASE is observed in almost every single state, with the only exceptions being Maryland, Florida, and Virginia. Figure 4.6 provides another way to assess the improvement resulting from the state shrinkage by looking at the change in MASE from GBLC-age to GBLC-age-state. The magnitude of improvements shows a similar geographical pattern. It can be seen that adding state shrinkage has led to improvement in most states, especially in Ohio, Idaho, Colorado, and Delaware. While there has been some deterioration in forecast accuracy in a few states, such as Virginia, Arizona, and Michigan, it is relatively small in magnitude. On the whole, Figures 4.5 and 4.6 are consistent with each other. It can be argued that, when age shrinkage is included in gradient boosting, the improvement from adding state shrinkage may not be as substantial as when state shrinkage is added to the GBLC model without age shrinkage. On average, the GBLC-age-state model still has superior performance over the GBLC-age model across all states, as shown in Table 4.3.

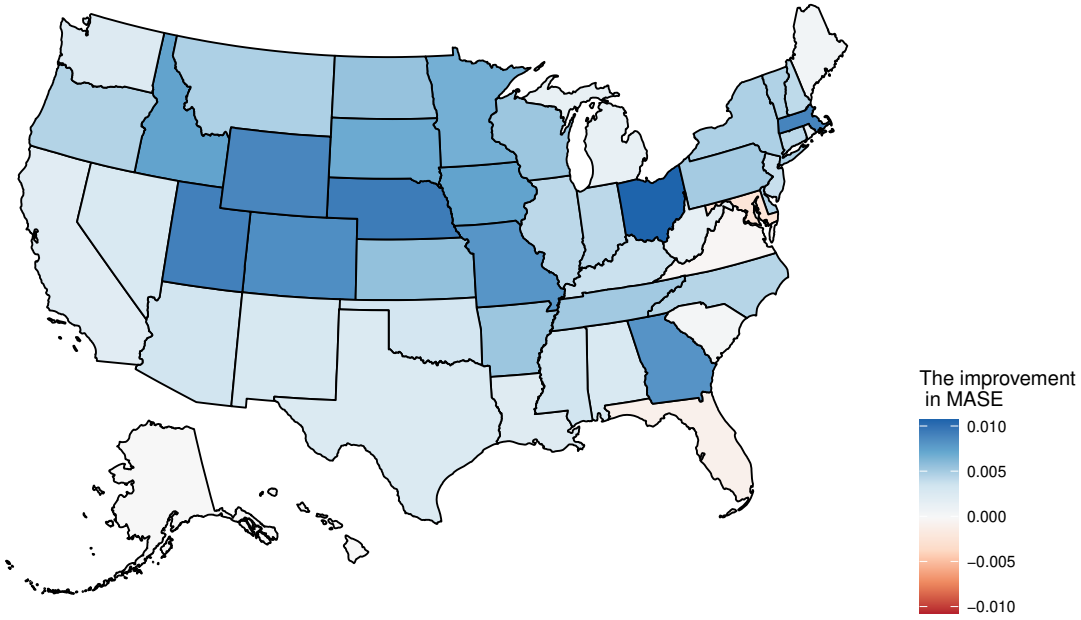


Figure 4.5: Improvement in MASE for state-level male mortality from GBLC to GBLC-state, h=10.

[9]Note that for Alaska (AK) and Hawaii (HI), state shrinkage is not applicable. Therefore, the change in MASE is simply 0.

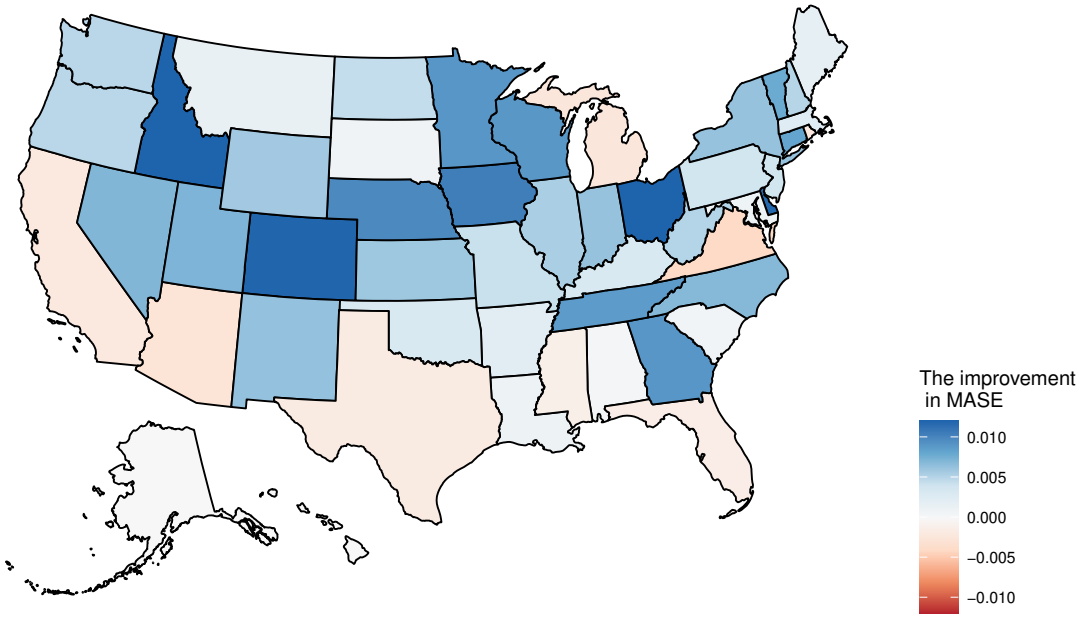


Figure 4.6: Improvement in MASE for state-level male mortality from GBLC-age to GBLC-age-state, h=10.

5 Conclusion

In this paper, we propose single- and multi-population mortality models that utilize gradient boosting methods with shrinkage in age and spatial dimensions. We then apply these models to US male mortality in our empirical studies. For national-level mortality forecasting, we demonstrate that the gradient boosted Lee–Carter model provides significant improvement over the benchmark models. Further improvement is achieved when age shrinkage is added to the model, especially for older age groups. Overall, the gradient boosted Lee–Carter model with age shrinkage achieves the highest forecast accuracy for national-level mortality projections. For state-level mortality forecasts, we demonstrate that the gradient boosted Lee–Carter model with age and state shrinkage exhibits the best performance in terms of out-of-sample forecast accuracy. The state shrinkage works particularly well for those states with small population and/or common risk factors.

Our proposed methodology opens up several avenues for future research. It should be noted that the proposed gradient boosting model with shrinkage is readily applicable to other stochastic mortality models such as the Cairns-Blake-Dowd model and its variations. Another interesting extension to our work would be the application of the proposed shrinkage methods at a finer geographical granularity (e.g. US counties rather than US states), if such data become available. At this scale, regional effects may play a greater role, for example higher mortality due to environmental hazards may only be observed at a very local level. Therefore, we anticipate that when the proposed models are applied to mortality rates of finer geographical areas, the spatial shrinkage will have an even more significant impact on forecast accuracy. Finally, our methods for age-based and spatial shrinkage only rely on a concept of neighbourhood which could be broadened to include other notions of similarity between time series. For instance, mortality rates of people in "neighbouring" income deciles could also be shrunk together using our proposed methodology.

References

Arató, N. M., Dryden, I. L., and Taylor, C. C. (2006). Hierarchical bayesian modelling of spatial age-dependent mortality. *Computational Statistics & Data Analysis*, **51**(2), 1347–1363.

Bernardi, L. C. . M. (2017). *MCS: Model Con dence Set Procedure*. R package version 0.1.3.

Bjerre, D. S. (2022). Tree-based machine learning methods for modeling and forecasting mortality. *ASTIN Bulletin: The Journal of the IAA*, **52**(3), 765–787.

Booth, H. and Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science*, **3**(1-2), 3–43.

Brant, S. B. and Haff, I. H. (2022). Copulaboost: additive modeling with copula-based model components.

Breiman, L. (1997). Arcing the edge. Technical report, Citeseer. Tech. Report 486, Statistics Department, University of California, Berkeley.

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

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., 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.

Case, A. and Deaton, A. (2015). Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. *Proceedings of the National Academy of Sciences*, **112**(49), 15078–15083.

Chapman, B. P., Fiscella, K., Kawachi, I., and Duberstein, P. R. (2010). Personality, socioeconomic status, and all-cause mortality in the United States. *American Journal of Epidemiology*, **171**(1), 83–92.

Couillard, B. K., Foote, C. L., Gandhi, K., Meara, E., and Skinner, J. (2021). Rising geographic disparities in US mortality. *Journal of Economic Perspectives*, **35**(4), 123–46.

Cupido, K., Jevtić, P., and Paez, A. (2020). Spatial patterns of mortality in the United States: A spatial filtering approach. *Insurance: Mathematics and Economics*, **95**, 28–38.

Deprez, P., Shevchenko, P. V., and Wüthrich, M. V. (2017). Machine learning techniques for mortality modeling. *European Actuarial Journal*, **7**, 337–352.

Dwyer-Lindgren, L., Bertozzi-Villa, A., Stubbs, R. W., Morozoff, C., Mackenbach, J. P., van Lenthe, F. J., Mokdad, A. H., and Murray, C. J. (2017). Inequalities in life expectancy among US counties, 1980 to 2014: temporal trends and key drivers. *JAMA Internal Medicine*, **177**(7), 1003–1011.

Enchev, V., Kleinow, T., and Cairns, A. J. (2017). Multi-population mortality models: fitting, forecasting and comparisons. *Scandinavian Actuarial Journal*, **2017**(4), 319–342.

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics*, **28**(2), 337 – 407.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, pages 1189–1232.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of spatial statistics*. CRC press.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, **79**(2), 453–497.

Harris, K. M., Majmundar, M. K., Becker, T., *et al.* (2021). *High and rising mortality rates among working-age adults*. National Academies Press.

Huynh, N. and Ludkovski, M. (2021). Multi-output Gaussian processes for multi-population longevity modelling. *Annals of Actuarial Science*, **15**(2), 318–345.

Hyndman, R. (2023). *demography: Forecasting Mortality, Fertility, Migration and Population Data*. R package version 2.0.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F. (2020). *forecast: Forecasting Functions for Time Series and Linear Models*. R package version 8.12.

Hyndman, R. J. and Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, **51**(10), 4942–4956.

Hyndman, R. J., Booth, H., and Yasmeen, F. (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography*, **50**(1), 261–283.

Januschowski, T., Wang, Y., Torkkola, K., Erkkilä, T., Hasson, H., and Gasthaus, J. (2022). Forecasting with trees. *International Journal of Forecasting*, **38**(4), 1473–1481.

Jemal, A., Ward, E., Anderson, R. N., Murray, T., and Thun, M. J. (2008). Widening of socioeconomic inequalities in US death rates, 1993–2001. *PLOS One*, **3**(5), e2181.

Kuhn and Max (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, **28**(5), 1–26.

Lee, R. (2000). The Lee-Carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal*, **4**(1), 80–91.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association*, **87**(419), 659–671.

Lee, S. C. and Lin, S. (2018). Delta boosting machine with application to general insurance. *North American Actuarial Journal*, **22**(3), 405–425.

Levantesi, S. and Pizzorusso, V. (2019). Application of machine learning to mortality modeling and forecasting. *Risks*, **7**(1), 26.

Li, H., O'Hare, C., and Vahid, F. (2016). Two-dimensional kernel smoothing of mortality surface: An evaluation of cohort strength. *Journal of Forecasting*, **35**(6), 553–563.

Li, H., O'hare, C., and Vahid, F. (2017). A flexible functional form approach to mortality modeling: Do we need additional cohort dummies? *Journal of Forecasting*, **36**(4), 357–367.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, **42**, 575–594.

Lin, T. and Tsai, C. C.-L. (2022). Hierarchical Bayesian modeling of multi-country mortality rates. *Scandinavian Actuarial Journal*, **2022**(5), 375–398.

Lourés, C. R. and Cairns, A. J. (2020). Mortality in the US by education level. *Annals of Actuarial Science*, **14**(2), 384–419.

Parcha, V., Kalra, R., Best, A. F., Patel, N., Suri, S. S., Wang, T. J., Arora, G., and Arora, P. (2021). Geographic inequalities in cardiovascular mortality in the United States: 1999 to 2018. In *Mayo Clinic Proceedings*, volume 96, pages 1218–1228. Elsevier.

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

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

Richman, R. (2021). AI in actuarial science–a review of recent advances–part 2. *Annals of Actuarial Science*, **15**(2), 230–258.

SriDaran, D., Sherris, M., Villegas, A. M., and Ziveyi, J. (2022). A group regularisation approach for constructing generalised age-period-cohort mortality projection models. *ASTIN Bulletin: The Journal of the IAA*, **52**(1), 247–289.

Tutz, G. and Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, **62**(4), 961–971.

Vierboom, Y. C., Preston, S. H., and Hendi, A. S. (2019). Rising geographic inequality in mortality in the United States. *SSM-Population Health*, **9**, 100478.

Woolf, S. H. and Schoomaker, H. (2019). Life expectancy and mortality rates in the United States, 1959–2017. *JAMA*, **322**(20), 1996–2016.

Appendix

Table A.1: US state abbreviations and names (including the District of Columbia)

Abbreviation	State name	Abbreviation	State name
AL	Alabama	MT	Montana
AK	Alaska	NE	Nebraska
AZ	Arizona	NV	Nevada
AR	Arkansas	NH	New Hampshire
CA	California	NJ	New Jersey
CO	Colorado	NM	New Mexico
CT	Connecticut	NY	New York
DE	Delaware	NC	North Carolina
DC	District of Columbia	ND	North Dakota
FL	Florida	OH	Ohio
GA	Georgia	OK	Oklahoma
HI	Hawaii	OR	Oregon
ID	Idaho	PA	Pennsylvania
IL	Illinois	RI	Rhode Island
IN	Indiana	SC	South Carolina
IA	Iowa	SD	South Dakota
KS	Kansas	TN	Tennessee
KY	Kentucky	TX	Texas
LA	Louisiana	UT	Utah
ME	Maine	VT	Vermont
MD	Maryland	VA	Virginia
MA	Massachusetts	WA	Washington
MI	Michigan	WV	West Virginia
MN	Minnesota	WI	Wisconsin
MS	Mississippi	WY	Wyoming
MO	Missouri		

Table A.2: MASE of out-of-sample forecasts for all states and District of Columbia, h=10, males

State	L–C	H–U	H–B–Y	GBLC	GBLC-age	GBLC-state	GBLC-age-state
AL	0.846	0.799	0.777	0.843	0.810	0.841	0.809
AK	0.901	0.755	0.720	0.824	0.769	0.824	0.769
AZ	0.625	0.614	0.640	0.708	0.690	0.704	0.693
AR	0.730	0.755	0.756	0.831	0.791	0.826	0.788
CA	0.780	0.711	0.890	0.656	0.642	0.654	0.644
CO	0.699	0.669	0.656	0.714	0.696	0.706	0.684
CT	0.842	0.794	0.867	0.809	0.803	0.804	0.795
DE	1.267	0.971	0.961	0.943	0.882	0.939	0.870
DC	1.798	1.646	1.345	1.237	1.205	1.241	1.210
FL	0.755	0.727	0.825	0.747	0.741	0.748	0.742
GA	0.698	0.670	0.706	0.720	0.693	0.712	0.684
HI	1.099	0.970	0.928	0.957	0.910	0.957	0.910
ID	0.697	0.712	0.605	0.755	0.724	0.748	0.712
IL	0.673	0.748	0.843	0.720	0.696	0.716	0.691
IN	0.934	0.907	0.861	0.860	0.826	0.856	0.820
IA	0.726	0.711	0.680	0.808	0.779	0.801	0.768
KS	0.740	0.795	0.720	0.828	0.790	0.823	0.784
KY	0.969	0.943	0.813	0.905	0.867	0.901	0.864
LA	0.769	0.847	0.734	0.884	0.850	0.882	0.849
ME	1.394	1.058	0.973	0.960	0.936	0.960	0.934
MD	0.747	0.776	0.872	0.828	0.793	0.831	0.792
MA	0.863	0.810	0.874	0.799	0.774	0.790	0.772
MI	0.893	0.895	0.868	0.813	0.784	0.811	0.786
MN	0.677	0.716	0.685	0.730	0.695	0.724	0.685
MS	0.710	0.724	0.698	0.831	0.776	0.827	0.777
MO	0.847	0.850	0.792	0.840	0.810	0.832	0.805
MT	1.187	0.918	0.852	0.927	0.898	0.922	0.897
NE	0.919	0.819	0.673	0.824	0.782	0.815	0.772
NV	0.685	0.775	0.507	0.621	0.609	0.618	0.602
NH	1.281	1.037	0.977	0.929	0.903	0.925	0.898
NJ	0.781	0.746	0.898	0.764	0.742	0.760	0.738
NM	0.792	0.761	0.716	0.757	0.723	0.754	0.717
NY	0.717	0.664	0.924	0.660	0.651	0.655	0.645
NC	0.795	0.817	0.758	0.779	0.752	0.775	0.745
ND	1.183	1.003	0.935	0.999	0.955	0.993	0.951
OH	1.183	1.129	1.078	1.015	1.002	1.005	0.990
OK	0.875	0.834	0.876	0.864	0.820	0.860	0.817
OR	0.695	0.664	0.629	0.705	0.684	0.700	0.679
PA	0.900	0.870	0.976	0.868	0.847	0.863	0.843
RI	1.333	1.207	0.993	0.945	0.928	0.944	0.930
SC	0.700	0.680	0.664	0.742	0.698	0.742	0.698
SD	1.138	1.010	0.815	0.900	0.862	0.893	0.861
TN	0.895	0.898	0.811	0.865	0.833	0.860	0.824
TX	0.702	0.685	0.744	0.681	0.667	0.678	0.669
UT	0.637	0.630	0.601	0.734	0.691	0.725	0.684
VT	1.323	1.093	1.090	0.985	0.941	0.980	0.933
VA	0.770	0.757	0.840	0.728	0.698	0.729	0.702
WA	0.615	0.595	0.615	0.646	0.634	0.643	0.629
WV	1.261	1.151	1.004	0.994	0.958	0.992	0.953
WI	0.786	0.771	0.764	0.808	0.771	0.803	0.762
WY	1.145	0.974	0.894	0.962	0.897	0.953	0.891