Mortality Prediction using Statistical Learning Approaches

by

Yechao Meng

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Actuarial Science

Waterloo, Ontario, Canada, 2022

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Jonathan Ziveyi
Associate Professor
Sch. of Risk and Actuarial Studies, University of New South Wales

Supervisor(s): Chengguo Weng
Professor
Dept. of Statistics and Actuarial Science, University of Waterloo

Liqun Diao
Assistant Professor
Dept. of Statistics and Actuarial Science, University of Waterloo

Internal Member: Mary Hardy
Professor
Dept. of Statistics and Actuarial Science, University of Waterloo

Johnny Li
Professor
Dept. of Statistics and Actuarial Science, University of Waterloo

Internal-External Member: Yaoliang Yu
Associate Professor
Sch. of Computer Science, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Longevity risk, as one of the major risks faced by insurers, has triggered a heated stream of research in mortality modeling among actuaries for effective design/pricing/risk management of insurance products. The idea of borrowing a "proper" amount of information from populations with similar structures, widely acknowledged as a conducive strategy to enhance the accuracy of the mortality prediction for a target population, has been explored and utilized by the actuarial community. However, the problem of determining a "proper" amount of information amounts to a trade-off that one needs to strive well between gains from including relevant signals and adverse impacts from bringing in irrelevant noise. Conventional solutions to determine a "proper" amount of information resort to multiple sources of exogenous data and involve substantial manual work of "feature engineering" without guaranteeing an improvement in prediction accuracy. Therefore, in this thesis, we set sail from the exploration to design fully data-driven frameworks to screen out useful hidden information from different aspects effectively to enhance the predicting accuracy of mortality rates with the assistance of various statistical learning approaches.

First and foremost, Chapter 2 aims to throw light on how to select a "proper" group of populations among a given pool to ensure the success of a multi-population mortality model conducive to improved mortality predicting accuracy. We design a fully data-driven framework, based on a Deletion-Substitution-Addition algorithm, to automatically recommend a group selection for joint modeling through a multi-population model in order to obtain enhanced predicting accuracy. The procedure avoids the excessive involvement of subjective decisions in the group selection. The superiority of the proposed framework in mortality predicting performance is evident by extensive numerical studies when compared with several conventional strategies for population selection problems.

Chapter 3 also focuses on how to effectively borrow information from a given pool of populations to enhance the mortality predicting accuracy in a computationally efficient manner. In this chapter, we propose a bivariate model based ensemble framework to aggregate predictions that use the joint information from each pair of populations in the given pool. In addition, we also introduce a time-shift parameter to the base learner mortality model for extra flexibility. This additional parameter characterizes the time by which one population is ahead of or behind the other in their mortality development stages and allows for borrowing information from populations at disparate mortality development stages. The results of the empirical studies confirm the effectiveness of the proposed framework.

In Chapter 4, we extend the idea of borrowing information by changing the scope of consideration from populations to ages. We provide insights on detecting similarities and

borrowing information that is hidden under the similarities of age-specific mortality patterns among ages. We propose a novel predicting framework where the overall predicting goal is decomposed into multiple individual tasks that search for age-specific age bands to ensure the mortality prediction of each target age can receive the benefit of borrowing information across ages to the largest extent. Extensive empirical studies with the Human Mortality Database confirm noticeable differences for different target ages in their ways of borrowing information from other ages. Those empirical studies also confirm an overall improvement in predicting accuracy of the proposed framework for most ages, especially for adults and retiree groups.

In Chapter 5, information across different ages and different populations is considered simultaneously. We extend the idea of borrowing information among ages to multi-population cases and proposed three different approaches: a distance-based approach, an ensemble-based approach, and an ACF model-based approach. Empirical studies with real mortality data are conducted to compare their predicting performance and significance in improving predicting accuracy compared with some benchmark models. Additionally, several general stylized facts of how ages from multiple populations are borrowed by the distance based-method are provided.

Finally, Chapter 6 briefly outlines some directions worth further exploration for research by the momentum from each chapter and some research ideas that are less relevant to the previous chapters.

Acknowledgements

At this point, I would like to express my sincere gratitude for all people that helped, encouraged, inspired and accompanied me along my journey at the University of Waterloo.

First and foremost, I would like to take this opportunity to express my deepest gratitude to my supervisor Dr. Chengguo Weng and Dr. Liqun Diao, for all their steady support and excellent supervision. Both of them not only set exceptionally good examples as academic researchers for me with their splendid devotions and qualities, but also devote lots of their time in taking care of me in many aspects with exceptional guidance in addressing research problems, continuous support in coping with ups and downs I encountered in the journey and mentally encouragement in overcoming obstacles of all kinds. All their efforts, patience, and guidance are indispensable for me to work on the thesis. I feel exceptionally fortunate to have them as my supervisors and the journey with them would always be my biggest treasure.

I wish to express my special appreciation to Dr. Mary Hardy and Dr. Johnny Li for spending their valuable time reading my thesis and providing insightful comments as thesis proposal and defence committee members. Their expertise and valuable comments helped me deeply in sharpening my thesis and developing insights for the research topic of human mortality modeling. In addition, I would like to thank Dr. Jonathan Ziveyi and Dr. Yaoliang Yu with my sincere gratitude in advance for serving as my thesis committee and also for their time and effort to review my thesis, providing thoughtful and valuable comments.

My journey at the University of Waterloo started as a MMath student in Actuarial Science in 2016 and I have been always grateful for all kinds of supports and guidance received from members in the department. I am grateful to meet Dr. Ruodu Wang, Dr. David Saunders, Dr. Fan Yang, Dr. Tony Wirjanto, Dr. Alexander Schied, Dr. Adam Kolkiewicz, Dr. David Landriault, Dr. Christiane Lemieux and Dr. Phelim Boyle in their thought-provoking lectures for creating such a great academic community. Special thanks to Dr. Tony Wirjanto for sharing his knowledge and expertise in our collaboration. Moreover, I would like to thank Ms. Mary Lou Dufton and Mr. Greg Preston for their continuing help with administrative and technical issues. Thanks to Dr. Keith Freeland and Dr. Mirabelle Huynh for their suggestions and support in helping me get along with the role of an instructor and polishing my teaching skills. Thanks Dr. Peijun Sang for his support and encouragement.

To my sincere friends and peers for the days and nights along the journey: Chi-Kuang Yeh, Wenyuan Li, Fangya Mao, Zhaohan Sun, Ce Yang, Sheng Wang, Takaaki Koike,

Xiaoxue Deng, Zijia Wang, Yuyu Chen, Wenling Zhang, Jie Jian, Trang Bui, Qihuang Zhang, Zhaoran Hou, Jingyue Huang, Kecheng Li, Jianchu Chen, Hongda Hu, Qiuqi Wang, Mingren Yin, Yiping Guo, Yimiao Zhao, Weinan Qi, Feiyu Zhu, Zhenzhen Huang, Qinghua Ren, Zhiqiao Song and Yuling Chen. I also really appreciate my life-long friends Runtong Yang, Ji Qiu, Zihao Zhang, Yanjun Qian, Yifei Song, Mengge Chen, Shuai Yuan, Jing Tu, Jiayi Zheng, Huameng Jia, Zhenni Tan, Shuying Yan, Jingwei Yan, Tian You, Sijia Ma, Rui Jie, Liuyan Ji, Xiaoxuan Zhao, Jiangxue Wu and all friends that are not listed here for their friendship.

Special thanks to my little brother, Xiyue, with whom I went through the bumpy road on the fenseline separating youth and maturity. Those poetic memories about endurance, courage, sacrifice and humanity shine like the last gold of expired stars.

Last but not least, I would like to express my deepest gratitude to my family members, especially my parents Jie Meng and Qiong Li, for their endless love and unconditional support since I was born. Although I haven't meet them for almost three years since the COVID19 pandemic, it would not be possible for me to finish the Ph.D. program without their encouragement.

To my parents

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 An Overview of Human Mortality Modeling

1.1.1 Mortality Models and Mortality Prediction

Longevity risk, attributed to the increase in human life expectancy, has been recognized as one of the major risks faced by insurers, governments, and individuals. The effect of longevity risk is systematic, and it has created substantial financial pressure on the pension fund industry and annuity providers. A reliable mortality forecast is crucial for the pricing and valuation of various life insurance and living benefit products. It is also critical to the establishment of prudent risk management strategies for various insurance institutions. As such, human mortality modeling and prediction have become one of the most popular research topics in past decades in the actuarial community. Being viewed as "a bumpy road to Shangri-La" by demographer [Tuljapurkar, 2005], the task of providing future human mortality prediction of high qualities involves studies on obtaining high-accuracy forecasts for the future from the past and present data based on specific mortality models and has triggered a proliferation of research.

The mortality prediction is typically performed by a two-stage procedure. In the first stage, a specific mortality model is chosen and calibrated using historical mortality data. In the second stage, future death rates are obtained through extrapolation based on some underlying time series models, usually within the class of autoregressive integrated moving average (ARIMA) processes. The choices of both the mortality model and the extrapolative method are crucial to the eventual performance of the resulting mortality forecasts.

Mortality models, as the inevitable tool to comprehensively understand and quantify human mortality rates, usually focus on how mortality rates are influenced by different

ages of individuals, the medical and social progress with respect to time, and the lifelong effects that follow individuals from birth. The study of mortality models has a very long history. Numerous mortality models have been developed since Gompertz published his law of mortality in 1825 [Gompertz, 1825]. Among all the mortality models developed thus far, the Lee-Carter model [Lee and Carter, 1992] has become the benchmark model since its seminal publication. The Lee-Carter model decomposed the age-specific mortality rates over a certain time period into the mean age-specific mortality rates, the mortality trend, the amount of mortality change at a given age, and an error term. Since then, various extensions of the Lee-Carter model have been proposed with alternative estimation procedures or adjustments on assumptions for improvements in modeling or forecasting of mortality rates. Below are some examples. Wilmoth [1993] introduced the weighted least squares estimation for the model in the presence of zero-mortality rates and non-constant variance in mortality data. Lee and Miller [2001] and Booth et al. [2002] addressed the issue with the jump-off bias and put forward different approaches to adjust the overall time trend. Brouhns et al. [2002a] substituted the Singular Value Decomposition (SVD) approach with a log-bilinear Poisson regression procedure for the calibration of the Lee-Carter model. Renshaw and Haberman [2003a] reinterpreted the Lee-Carter model and introduced a methodology based on a generalized linear model. De Jong and Tickle [2006] generalized the model under the state-space framework to encompass a wide range of flexible multivariate time series models, among which the Lee-Carter model is a special case.

In addition to the adjustments made to the Lee-Carter model, many new methods and models have also been proposed in the literature. Currie et al. [2004] used P-splines for the smoothing and forecasting of two-dimensional mortality tables. Hyndman and Ullah [2007] promoted ideas from functional data analysis, nonparametric smoothing, and robust statistics to form a mortality model that could also be viewed as a generalization of the Lee-Carter model. Renshaw and Haberman [2006] extended the Lee-Carter model through a cohort parameter to describe the non-linear mortality trend in order to improve its forecasting performance. Hatzopoulos and Haberman [2009] and Hatzopoulos and Haberman [2011] employed a sparse age-period model structure for mortality experience under a generalized linear model framework and utilized multiple principal components to extract mortality dynamics. To better capture the mortality trend in older ages, Cairns et al. [2006] introduced the Cairns-Blake-Dowd (CBD) model which employs linear or quadratic functions of age to model the logit of the death probabilities for senior ages. Other worth-mentioned mortality models includes Plat [2009], and O'Hare and Li [2012]. As a holistic analysis for the general APC-type mortality model, Hunt and Blake [2021] examined the similarities and differences among a number of mortality models and provided a classification scheme

for these models. Furthermore, Cairns et al. [2009] summarized the most famous M1-M8 and provides a quantitative comparison between them.

All aforementioned mortality models fit into the concept of the single population model, which, as the name suggests, is intended to model the mortality rates of a single population. When it comes to the forecast of multiple populations, these models treat each population separately and provide a forecast for each population independently. In recognition of the common features in mortality development patterns across populations, coherent multi-population mortality methods have been developed and their capacity to enhance forecasting accuracy has been well attested. Coherent multi-population mortality models have been designed to avoid unrealistic crossovers or divergence in anticipated future mortality across countries or between genders, which could arise from applying a single-population model to each population separately. The main idea behind coherent forecasting is that mortality forecasts for populations with similar mortality developments are not expected to diverge substantially even though structural differences remain across populations. As a prototype of various coherent multi-population mortality forecasting models, the Augmented Common Factor (ACF) model proposed by Li and Lee [2005] extended the Lee-Carter model to multiple populations. The model jointly models the logarithmic age-specific death rates of a group of different populations with a common factor, and at the same time, it includes population-specific components to allow disparate development of mortality for individual populations. After the introduction of the ACF model, many other multiple-population mortality models have been proposed. In particular, Cairns et al. [2011b] and Dowd et al. [2011] discussed the incorporation of the age-period-cohort structure into a two-population modeling framework. Kleinow [2015] proposed the common age effect model in which age has the same effect on the centralized logarithmic mortality rates for all the populations in the joint model. The functional-data-based approach in Hyndman and Ullah [2007] was also adopted for a multi-population mortality modeling in Hyndman et al. [2013], Shang [2016] and Shang and Hyndman [2017]. Russolillo et al. [2011] incorporated a third component that represents differences in populations into the decomposition of the logarithmic mortality rates, thereby extending the Lee-Carter model to a three-way structure. Hatzopoulos and Haberman [2013] extended the GLM framework in Hatzopoulos and Haberman [2009] and Hatzopoulos and Haberman [2011] into multi-population cases with the help of clustering method to construct the common age/period effect terms. Moreover, Li et al. [2015b] proposed bivariate-population generalizations for each of the seven single-population models M1-M3 and M5-M8 from Cairns et al. [2009]. Villegas et al. [2017] offered a comprehensive review of multiple-population models and a systematical assessment that appraises the suitability of available two-population mortality models for the assessment of basis risks

resulting from using longevity swaps for pension scheme de-risking. Enchev et al. [2017] compared between variations of the ACF model [Li and Lee, 2005] and the common age effect (CAE) model [Kleinow, 2015]. Other papers that provide insights on model comparison and model selection include Danesi et al. [2015], Atance et al. [2020]

As mentioned earlier in this section, the task of mortality prediction is typically performed in a two-stage procedure, where the employed mortality model is fitted to historical mortality data for the estimates of the model parameters in the first stage and a proper time series model, usually an autoregressive integrated moving average (ARIMA) process, is chosen to extrapolate the time effect sequences and obtain a forecast of future death rates. With the aim of improving eventual mortality forecasts, researchers proposed multivariate time series models for relevant components in a joint mortality model. For example, Zhou et al. [2014] considered modeling the dynamics of mortality rates of two related populations simultaneously with a two-dimensional vector autoregressive model (VAR) and a vector error correction model (VECM). Other examples of this type of model include VAR models with higher dimensions and different sparsity constraints [Li and Lu, 2017, Guibert et al., 2019, Shi, 2021, Li and Shi, 2021, Chang and Shi, 2022] and VECM-based models [Yang and Wang, 2013, Zhou et al., 2019]. Furthermore, semiparametric models have also been applied to mortality modeling to provide a different way for mortality predictions [Li et al., 2015a, 2016, Li and O'Hare, 2017]. Actuarial researchers also seek to explore and develop alternative statistical representations for mortality models that avoid the traditional two-stage procedure by allowing modeling, estimation, and prediction of mortality under a unified framework. As Fung et al. [2017] and Fung et al. [2019] have demonstrated, popular mortality models such as the Lee–Carter class of models can be written in a general state-space modeling methodology.

Researchers also explored other aspects of human mortality modeling. Firstly, as mentioned in Booth and Tickle [2008], the correct estimation of forecast uncertainty has become an important goal. The impact of parameter uncertainty has been studied by Brouhns et al. [2002b, 2005], Koissi et al. [2006] with a bootstrapping methodology, and Czado et al. [2005], Cairns et al. [2006] with a Bayesian framework. Secondly, to better account for multiple sources of uncertainties and deal with missing data, a Bayesian framework has been incorporated with mortality and demographic modeling [Pedroza, 2006, Kogure and Kurachi, 2010, Cairns et al., 2011b, Li, 2014, Wiśniowski et al., 2015, Wong et al., 2018, Lin and Tsai, 2022]. Moreover, consideration has also been given to the impact of potential structural changes in mortality trends with representative examples including the broken-trend stationary model in Li et al. [2011], the regime switching models to mortality modeling in Milidonis et al. [2011], and the incorporation of some testing methods for structural break

detection in Coelho and Nunes [2011], O'Hare and Li [2015], and Van Berkum et al. [2016]. Other topics include the non-Gaussian error terms of mortality models [e.g., Wang et al., 2011, 2013] and age coherence in Li and Lu [2017], Li [2013], and Gao and Shi [2021].

1.1.2 Mortality Data

The Human Mortality Database (HMD), launched online in 2002, is regarded as an important data resource to provide detailed, consistent, reliable, and accurate human mortality data for longevity-related research. As described in Barbieri et al. [2015], the database contains mortality data for 46 populations (including sub-national groups) in total, with data classified by age (from age 0 up to age 110+), sex, year of death, and year of birth. The database consists of both original raw data collected from official sources such as birth counts, death counts, and calculated data, including population size, exposure-to-risk, death rates, and life tables. Although the database has the earliest data stemming back to 1751 in Sweden and contains observations over 100 years for one-third of the included countries or regions, we aim at studying a recent period of data from 1970 to 2010 to avoid the potential impact of rare events such as the Second World War throughout the thesis. The mortality data are accessible through software `R` as the common programming language. Typical `R` packages related to demographic analysis and mortality modeling includes `demography` by Hyndman et al. [2019] and `StMoMo` by Villegas et al. [2015].

1.2 Statistical Learning in Human Mortality Modeling

A revolution in statistics occurred between 1960 and 1980 when the statistical learning theory was introduced and developed [Vapnik, 1999a,b]. As vast amounts of data are being generated in many fields, the evolving statistical learning approaches have played an increasingly important role in coping with the challenge of extracting important information hidden in the data.

The statistical learning approaches can be broadly categorized into two classes: supervised learning and unsupervised learning. Generally, supervised learning refers to predicting or estimating an output based on one or more inputs. Unsupervised learning, on the other hand, describes the associations and patterns within the given data without a supervised output. There are already conventional statistical models which have long been applied in actuarial science. While a detailed introduction of statistical learning approaches can be found in Vapnik [1999a] and Hastie et al. [2009], we provide a brief overview of how

other more advanced complementary statistical learning methods have been incorporated into the research area of human mortality modeling and forecasting. Tsai and Cheng [2021] has incorporated statistical clustering methods into mortality models to improve prediction performance. Other examples of applying clustering methods include Schnürch et al. [2021] and Levantesi et al. [2022]. Under the topic of model combination, Shang and Haberman [2018] utilized the model confidence set approach proposed in Hansen et al. [2011] to combine multiple mortality models for improved mortality predictions, Kontis et al. [2017] applied a probabilistic Bayesian model averaging (BMA) approach to mortality and life expectancy projection, and Kessy et al. [2021] developed a stacked regression ensemble method to combine predictions from different mortality models. Neural network, as one of the prevailing machine learning tools, has also been incorporated into mortality prediction procedures recently. Perla et al. [2021] generalized the Lee-Carter model using convolutional network models for predictions. Richman and Wüthrich [2021] extended the model to multiple populations by using neural networks for automatic selection of optimal model structure, and Nigri et al. [2019] applied a recurrent neural network with a long short-term memory architecture to the Lee-Carter model for improved predictive capacity. Wang et al. [2021] has considered capturing the "neighboring" effect with the help of neural networks to enhance predictive power. Furthermore, the gaussian process regression approach has also been adopted for mortality prediction in Huynh and Ludkovski [2021] and Lam and Wang [2021].

1.3 Review of Mortality Models

This subsection serves as a quick review of several mortality models that will be constantly involved in the thesis.

1.3.1 Lee-Carter Model

Let $m_{x,t}$ denote the central death rate, also known as the age-specific death rate (ASDR), for age $x \in \{0, 1, \ldots, \omega\}$ and year $t \in \{0, 1, \ldots, T\}$ of a population, where ω represents the limit age of the population. The Lee-Carter model decomposes the age-period surface of logarithmic ASDRs in the following form:

$$\log m_{x,t} = a_x + b_x k_t + \epsilon_{x,t}, \tag{1.1}$$

with normalization constraints $\sum_x b_x = 1$ and $\sum_t k_t = 0$, where $\epsilon_{x,t}$ represents the white noise term.

The right-hand side of Equation (1.1) contains a static age function a_x that captures a general shape of mortality across ages and features of the mortality curve that do not change with time, a period function k_t that determines the evolution of mortality rates over time, and a non-parametric age function b_x that captures the relative speed of change in mortality at each age x.

The estimation methods of the parameters in the Lee-Carter model can be categorized into non-likelihood-based and likelihood-based. No probability distributions are assumed when the non-likelihood-based methods are adopted while a probability distribution for the death counts is specified when the likelihood-based methods are adopted. As a typical example of the non-likelihood-based methods, the Singular Value Decomposition (SVD) method is proposed in Lee and Carter [1992], where the static age function a_x is estimated as the average of logarithmic ASDRs over the modeling time-period:

$$\hat{a}_x = \frac{1}{T+1} \sum_{t=0}^{T} \log m_{x,t}, \tag{1.2}$$

and b_x and k_t are respectively identified as the first left and the first right singular vectors of the matrix $\log m_{x,t} - \hat{a}_x$. Another example that belongs to the category of the non-likelihood-based methods is the weighted least squares (WLS) estimation method in Wilmoth [1993]. As for the likelihood-based methods, classic examples like Wilmoth [1993], Brouhns et al. [2002a] and Renshaw and Haberman [2006] all assume a Poisson distribution for the observed number of deaths with the mean equal to the expected number of deaths under the Lee-Carter model; that is

$$D_{x,t} \sim \text{Poisson}\left(E_{x,t}, \exp(a_x + b_x k_t)\right), \tag{1.3}$$

with $D_{x,t}$ and $E_{x,t}$ denoting the number of deaths and the exposures-to-risk at age x and time t, respectively. The estimates of the model parameters are consequently obtained by maximizing the corresponding log-likelihood function. Other distributions like the negative binomial distribution have also been proposed as the candidate of the distribution for the observed number of deaths; see, for example, Li et al. [2009].

An appropriate time series model for the mortality index k_t is vital to the mortality predictions through the Lee-Carter model. Lee and Carter [1992] proposed a random walk with drift (RWD):

$$k_{t+1} = k_t + d + e_t, \tag{1.4}$$

where d is a constant drift and $\{e_t, t \geq 0\}$ are the independent and identically distributed (i.i.d.) error terms. Both the drift d and the variance of the error terms e_t can be estimated by standard time series estimation techniques. k_t is predicted stochastically and then used

to forecast ASDRs through Equation (1.1). Apart from the RWD, the Lee-Carter model has also been proposed for implementation with the more general ARIMA model [e.g., Li et al., 2009]. The ARIMA model can be estimated by the Box-Jenkins method [Box et al., 2015] and implemented using the `auto.arima` function from the `R` package `forecast` [Hyndman et al., 2007].

1.3.2 CBD Model

The CBD model, introduced in Cairns et al. [2006], is regarded as one of the most popular competitor models to the Lee-Carter model for the prediction of senior ages. Let $q(x,t)$ denote the probability that an individual in population j aged x will die between t and $t+1$ given the individual is alive at time t, for $t = 0, 1, \ldots, T$ and $x = x_1, \ldots, x_n$. The CBD model depicts the mortality development of a population in the following form:

$$\operatorname{logit} q(x,t) = \log\left[\frac{q(x,t)}{1-q(x,t)}\right] = K_t + (x-\bar{x})k_t + \epsilon_{x,t}, \tag{1.5}$$

with $\bar{x}$ denotes the average age in the data range being used and $\epsilon_{x,t}$'s are the i.i.d. error terms.

The absence of the static age function and the adoption of the parametric age function in Equation (1.5) reduce the number of free parameters and make the model more parsimonious. If we denote $D_{x,t}$ as the number of deaths at age x in year t, the corresponding exposure number $E_{x,t}$ and one-year death probability $q(x,t)$, the estimation of model parameters can be obtained via maximizing the model log-likelihood where $D_{x,t}$ is assumed to follow a binomial distribution with parameters $E_{x,t}$ and $q(x,t)$.

1.3.3 Several Extensions of Lee-Carter and CBD Models

As we previously mentioned, there are many extensions of the Lee-Carter models in the literature. Below we recall the specific contents of a few extensions that are relevant to our discussion in subsequent chapters of the thesis. First of all, the following models with multiple age/period terms are studied by Booth et al. [2002], Renshaw and Haberman [2003b] and Hyndman and Ullah [2007]:

$$\log m_{x,t} = a_x + \sum_{i=1}^{N} b_x^{(i)} k_t^{(i)} + \epsilon_{x,t}, \tag{1.6}$$

with $b_x^{(i)} k_t^{(i)}$ representing the multiple pairs of age/period effect terms. Furthermore, Renshaw and Haberman [2006] and Haberman and Renshaw [2009] considered the classic APC

model structure with a simplified form of

$$\log m_{x,t} = a_x + b_x k_t + \gamma_{t-x} + \epsilon_{x,t}, \tag{1.7}$$

where the extra cohort effect terms γ_{t-x} is introduced to address the mortality differences between people with different birth years $t - x$.

For CBD-type models, it is natural to consider extending the model by adding age/period terms with higher-order polynomial age functions or cohort terms. For instance, the following extensions of CBD models can be found in Cairns et al. [2009]:

$$\textbf{M6} : \text{logit}\, q(x,t) = K_t + (x - \bar{x})k_t + \gamma_{t-x} + \epsilon_{x,t}, \tag{1.8}$$

$$\textbf{M7} : \text{logit}\, q(x,t) = K_t + (x - \bar{x})k_t^{(1)} + \left((x - \bar{x})^2 - \hat{\sigma}_x^2\right) k_t^{(2)} + \gamma_{t-x} + \epsilon_{x,t}, \tag{1.9}$$

$$\textbf{M8} : \text{logit}\, q(x,t) = K_t + (x - \bar{x})k_t + (x_c - x)\gamma_{t-x} + \epsilon_{x,t}. \tag{1.10}$$

In the above, k_t and $k_t^{(i)}$ represent the period effect terms, γ_{t-x} represents the cohort effect term, the constant $\hat{\sigma}_x^2$ is calculated as the mean of $(x - \bar{x})^2$ over all the involved ages, and x_c is a constant parameter that needs to be estimated. In the above, we also used the same model labels M6-M8 as used in Cairns et al. [2009].

1.3.4 Augmented Common Factor (ACF) Model

Li and Lee [2005] pointed out that mortality patterns and trajectories in closely related populations are likely to be similar in some respects. For this reason, they argued that mortality forecasts for individual populations can be improved by taking into account the common patterns in a group. They proposed the augmented common factor (ACF) model which incorporates a common factor shared by all the individual populations in the group with the population-specific factors which further accommodate the remaining disparate development patterns in mortality for each population.

Let $m_{x,t,i}$ denote the ASDR of the ith population in a group G of populations for age x and year t. The ACF model depicts the mortality development of each individual population in the following form:

$$\log m_{x,t,i} = a_{x,i} + B_x K_t + b_{x,i} k_{t,i} + \epsilon_{x,t,i}, \tag{1.11}$$

with normalization constraints $\sum_x B_x = 1$, $\sum_t K_t = 0$, $\sum_x b_{x,i} = 1$, and $\sum_t k_{t,i} = 0$, where $\epsilon_{x,t,i}$'s are the i.i.d. error terms. The static age functions $a_{x,i}$ are still estimated as the average of logarithmic ASDRs over the modeling time period for each population:

$$\hat{a}_{x,i} = \frac{1}{T+1} \sum_{t=0}^{T} \log m_{x,t,i}, i \in G. \tag{1.12}$$

For the estimation of the other parameters, the SVD procedure is applied to the aggregate data of the group to obtain B_x and K_t, and then an extra round of SVD is further applied to the residual data, $(\log m_{x,t,i} - \hat{a}_{x,i} - \hat{B}_x \hat{K}_t)$, to get $b_{x,i}$ and $k_{t,i}$ for each individual population.

To obtain the forecast of mortality, Li and Lee [2005] proposed to model K_t with an RWD and $k_{t,i}$ with either a random walk (RW) without drift or a first-order autoregression (AR(1)) model to guarantee a nondivergent forecast in the long run. After both K_t and $k_{t,i}$ have been modeled, mortality forecasts are obtained by extrapolating these time series into years $t > T$ and by Equation (1.11).

Li and Lee [2005] argued that members in a properly chosen group should have similar mortality trends which are sufficiently captured by the common factor so that the remaining population-specific terms would be stable in their scales. They also explained that the model would fail if $k_{t,i}$ has a long-term trend because the long-term trend might be an indication of a systematically and significantly different trend between population i and the rest of the group.

Extensions of a multi-population model based on the ACF framework are not hard. For example, the common age effect (CAE) model proposed by Kleinow [2015] sets the same effect on the centralized logarithmic mortality rates for all the populations in the joint model as follows:

$$\log m_{x,t,i} = a_{x,i} + \sum_j b_x^{(j)} k_{t,i}^{(j)} + \epsilon_{x,t,i}, \tag{1.13}$$

where $b_x^{(j)} k_{t,i}^{(j)}$ representing the multiple pairs of age/period effect terms. The key spirit of the CAE model lies in the assumption that the age effect terms $b_x^{(j)}$ are the same shared by different populations in the system while the period effect terms $k_{t,i}^{(j)}$ remain population-specific. Yang et al. [2016] considered adding the cohort effect to the ACF model to model the centralized logarithmic mortality rates as follows:

$$\log m_{x,t,i} = a_{x,i} + B_x K_t + \sum_j b_{x,i}^{(j)} k_{t,i}^{(j)} + \gamma_{t-x} + \epsilon_{x,t,i}, \tag{1.14}$$

where γ_{t-x} are the cohort effect terms added to the system for all populations to depict the mortality differences between cohorts born in different years.

1.4 Research Questions

The idea of borrowing information from populations with similar structures has been well recognized as a useful strategy to enhance the accuracy of the mortality prediction for

a target population. A supportive example is the superiority of the ACF model to the Lee-Carter model in terms of prediction errors when the former is applied to a group of populations with some common mortality development patterns. As the discussion of human mortality modeling proceeds, the scope of borrowing information has been broadened to deeply dig into the useful hidden information in the mortality data itself from all aspects instead of being restricted to utilizing exogenous information, such as socioeconomic or geographic variables.

Figure 1.1 provides a preliminary illustration of the mortality trajectories at some different scopes. The left plot displays the age-aggregated (over ages 0 to 100) logarithmic mortality data sequences of 30 male populations (see Table A.1 for a detailed list) from 1970 to 2002, with two different colors indicating two different types of mortality development patterns based on some clustering results. The middle figure shows the age-aggregated (over ages 55 to 90) logarithmic mortality data sequences of the Canadian male and the U.S.A. male from 1970 to 2002. The dashed curve is a horizontal shift of the Canadan male curve by six years. The right panel demonstrates the age-specific logarithmic mortality data sequences of Canadan male from 1970 to 2002.

Figure 1.1 reveals the existence of both similarity and difference in mortality data across different dimensions of age, period, and populations. First of all, the age-aggregated logarithmic mortality rate sequences in the left panel indicate that the overall mortality development pattern can be similar within a certain group of populations and also can be very different across different groups of populations. Second, a contrast in the age-aggregated mortality level between the senior Canadian male and the senior U.S.A. male in the middle panel of the figure indicates the existence of a time lag in the development stage of mortality levels across populations. Finally, the right panel of the figure discloses that the similarity of mortality development is shared not only across populations but also across different ages within the same population. Certainly, the figure also clearly illustrates the difference in mortality development patterns when two ages are away from each other enough.

When it comes to borrowing information for mortality prediction, there exists a trade-off between the gains from including valuable signals and the adverse impact of bringing in irrelevant noise. This thesis aims to develop frameworks to dig out useful information from mortality data in different aspects by prudently designing and deploying sensible statistical learning approaches. The major focus of the thesis lies in considering borrowing information across populations and among ages. A brief overview of the topics we approached is given as follows whereas a more detailed discussion can be found in relevant chapters.

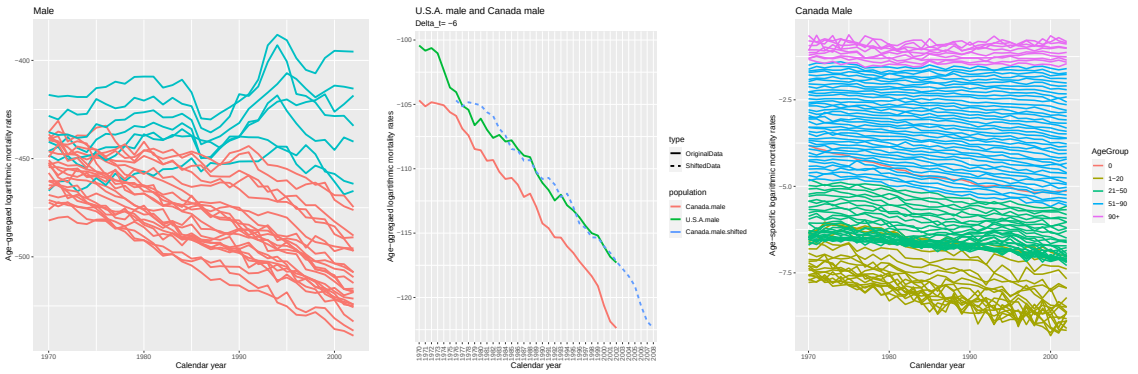


Figure 1.1: An illustration of similarity and difference in mortality trajectories. Left: age-aggregated logarithmic mortality data sequences of 30 male populations for ages 0 to 100. Middle: Age-aggregated logarithmic mortality rate sequences for the US male population and the Canadian male population for ages 55 to 90. Right: Age-specific logarithmic mortality rate sequences for the Canadian male population.

1.4.1 Borrowing Information across Populations

When the information across different populations is to be considered, one crucial step is to conduct multi-population mortality modeling. Chapter 2 aims to throw light on the problem of the selection of a "proper" group of populations, which is believed to be the key to the success in multi-population mortality modeling. When the group of populations are appropriately chosen, the hidden information would have a conducive impact on the study of the multi-population system and become particularly useful for insurance companies with overseas business to decide how to improve their domestic mortality prediction with the help of their overseas data. However, most of the existing joint mortality modeling procedures in current literature assume that the grouping information is either known a priori or pre-specified according to certain exogenous information, which involves substantial manual "feature engineering" and significant subjective judgment but is unable to insure an improvement in future prediction accuracy for the multi-population mortality modeling.

Moreover, further issues would arise in the current multi-population mortality modeling framework when a high-dimensional multi-population model is adopted as the base learner. First of all, an explosion in computational demand may result in intractability of the model or convergence issues for the parameter calibration procedure. Moreover, the "curse of dimensions" would lead to a lack of flexibility for the base learner mortality models and prevent the base learner to have a rich structure to depict some desired or interesting characteristics found in the mortality data of different populations. As an illustrative example, the kind of "parallelity" of trajectories and patterns of mortality development among pop-

ulations, shown in Figure 1.1, has been widely observed as a developing population may demonstrate a similar mortality improvement pattern in the recent decade to what a developed population has experienced over the past decades. Since it is reasonable to believe that even at different development stages of mortality level, populations could still be pulled together for joint modeling of mortality rates and that the similarity in the trajectories and patterns of their mortality rates can still bring in beneficial information for an enhancement of mortality prediction. However, the potential is not fully studied in current literature. Chapter 3 focuses on providing solutions to avoid possible computational hurdles and allow for the "parallelity".

1.4.2 Borrowing Information among Ages

The evolution in the discussion of multi-population mortality modeling has made the idea of borrowing information quite popular. On top of that, literature has confirmed that the concept of borrowing information can be further extended to other aspects like ages, for example, Shang and Haberman [2020] and Tsai and Cheng [2021]. Chapter 4 is inspired by the belief that detecting similarities among different ages and borrowing useful information among ages can potentially benefit mortality predictions in the form of enhancement in predicting accuracy. However, there are no clear views on how to determine a "proper" amount of information to be borrowed from the similarities among different ages and how differently the information among ages is utilized for different target ages. Let alone consider a more comprehensive task where the information across different ages and different populations is considered simultaneously.

1.5 Overview of the Thesis

The remainder of the thesis is organized as follows.

In chapter 2 we fill in the gap between multi-population mortality modeling and the grouping strategy that heavily depends on manually selection with exogenous explanatory information such as those from geographic or socioeconomic aspects. We develop a flexible framework for the selection of populations from a given candidate pool to assist a target population in mortality forecasting. The defining feature of the framework is the deletion-substitution-addition (DSA) algorithm, which is entirely data-driven and versatile to work with any multiple-population model for mortality prediction. As an illustration, the framework with an extended augmented common factor model is applied to the Human Mortality

Database (HMD) and the superiority of the proposed framework is evident in mortality predicting performance. Meanwhile, the automatic selecting outcome of the DSA-based model has insightfully provided recommendations on the membership of the optimal group for different target populations to circumvent difficulties in manual selections.

In chapter 3, We provide potential remedies for the drawbacks in high-dimensional multi-population mortality modeling by proposing a model averaging predicting framework that allows for borrowing information from the mortality data of a given pool of auxiliary populations to enhance the accuracy of the mortality forecast for a target population. Unlike directly fitting a high-dimensional multi-population model for future prediction, the model averaging idea is novelly used to aggregate information from different auxiliary populations by jointly linking each of them to the target population from a cascade of base learner models of a lower dimension. The usage of a bivariate-population base learner circumvents the potential computational difficulties and intractability of the high-dimensional multi-population model and makes it possible to capture vivid characteristics in mortality data with a more flexible base learner model. A time-shifting parameter Δt is introduced in the bivariate-population base learner model to capture the "parallelity" by characterizing the time by which one population is ahead of or behind the other in their mortality development stages. We consider various model averaging strategies, including a simple average, an average inside geographical groups, an average within k-means clusters, and a fully data-driven "Rank and average" strategy. The "Rank and average" ranks auxiliary populations according to their capability of assisting the prediction of the target population and average over those on the top, which utilizes the information from multiple populations effectively in a computationally friendly way. In addition, we add a time-shifting term to the base learner for extra flexibility to allow for borrowing information from populations at disparate mortality development stages. We conduct empirical studies with the Human Mortality Database to investigate the performance of the proposed model averaging method and study the value of including a time-shifting term in the base learner. It is interesting that these empirical studies reveal that the time-shifting parameter in the model is capable to characterize the development stage of one population relative to the other.

In chapter 4, we aim to provide insights on detecting similarities and borrowing information that is hidden under the trajectories of age-specific mortality among ages by proposing a novel predicting framework where the overall predicting goal is decomposed into multiple individual tasks of searching for individual age set to ensure the mortality prediction of each target age can receive the benefit of borrowing information across ages to the largest extent. Extensive numerical studies with the Human Mortality Database (HMD) have demonstrated noticeable differences for different target ages in their ways to borrow information among

other ages and confirm an overall improvement in predicting the accuracy of the proposed framework for the majority of ages, especially for adults and retiree groups.

In chapter 5, we aim to extend the idea of borrowing information among ages to multi-population scenarios with a more comprehensive framework that can take the information among different ages and across different populations into consideration simultaneously. Three different approaches are proposed. One of them extends the age-specific age set framework to a more general case with a proposed "distance" measure to quantify the similarities among different age-specific mortality sequences in their development patterns. Two additional approaches, respectively based on an ensemble paradigm and the ACF model are then introduced to address potential drawbacks. Numerical studies with the real mortality data have confirmed all the three approaches' potential capability to consider borrowing information among different ages from different populations with the desired improvement in predicting accuracy. Additionally, several general stylized facts of how ages from multiple populations are borrowed are provided based on the results of the proposed multi-population distance-based method. In general, there exists a noticeable difference that young and old ages generally borrow information from more ages with a wider range and from external ages while the adult/retiree ages choose less amount of reference ages with a more concentrated range within the population.

We finally summarize the thesis and outline some relevant topics which we are interested in for further research in Chapter 6.

Chapter 2

A DSA based Framework for Mortality Forecast

2.1 Introduction

It has been widely accepted as an advantage of using multi-population models for mortality predictions because borrowing information from populations with similar structural mortality patterns and trajectories can be helpful to the mortality forecasting of a target population. As mentioned in the preceding chapter, one crucial step in gaining the benefit of multi-population models is the selection of a "proper" group of populations in the joint modeling. Most existing joint mortality modeling procedures assume that the grouping information is either known a priori or pre-specified according to certain exogenous information, such as geographic proximity or socioeconomic variables. However, such a heuristic specification fails to guarantee that every selected population is indeed conducive to an enhancement in mortality prediction. Cluster analysis, insofar as we can tell, is the only structured method adopted in the literature for the purpose. Hatzopoulos and Haberman [2013] adopted the fuzzy C-means (FCM) cluster analysis to select populations with similar mortality characteristics for joint modeling. The FCM yields a list of cluster centers and a matrix to indicate the level of association each data element is with each center.

In this chapter, we propose a novel, flexible and effective method for the selection of populations. Instead of relying on geographic, socioeconomic, or any other exogenous information for grouping, our method is built on a well-designed deletion-substitution-addition (DSA) algorithm, which is entirely data-driven and directly "learns" for the best group of populations from a given candidate pool without resorting to any explanatory data input. Our DSA algorithm is inspired by the partDSA [see Molinaro et al., 2010, Sinisi and van der

Laan, 2004], a deletion-substitution-addition algorithm originally designed for prediction by recursively partitioning the covariate space in a regression context. The DSA algorithm simultaneously derives joint modeling results for the target population and the selected group of populations. We then extrapolate the established model into future periods to obtain mortality predictions for the target population.

Just as a typical clustering method (e.g. Hatzopoulos and Haberman [2013]), our DSA algorithm can also associate a target population with a particular group of populations for joint modeling, but with a substantial difference. At the outset of the implementation, one needs to designate a target population for the DSA algorithm to work towards the selection of populations for joint modeling. The best mortality prediction accuracy which the DSA algorithm aims to achieve pertains exclusively to the target population. If the target population changes to another in the pool, a rerun of the DSA algorithm is necessary. In general, the resulting grouping does not remain unchanged. In contrast, a typical clustering algorithm does not have a target population, and it aims at obtaining a grouping of objects that are more homogeneous within each group (also called a cluster) than those across groups.

We apply our DSA based prediction model to the Human Mortality Database (HMD) for sixty gender-specific populations from thirty countries or regions in the world. We compare the performance of the DSA based model with that of the Lee-Carter model and the ACF models which make use of different grouping strategies. A simple all-in-one strategy, a grouping strategy based on geographic proximity information, as well as grouping obtained by the k-means and the k-median cluster methods are respectively adopted in the implementation of the ACF models. Using data from 1970 to 2002, we train the models and extrapolate them into the period of 2003-2010 for prediction. The performance is measured by four metrics, that is, the mean squared error, the mean absolute error, the mean squared percentage error, and the mean absolute percentage error. Compared with the benchmark models mentioned earlier, our DSA based model proves to have provided results with higher prediction accuracy. Summary statistics as well as a formal hypothesis test, the Diebold-Mariano test, on the prediction results for the sixty populations corroborate the superiority of our DSA based model over the benchmarks.

The rest of this chapter is organized as follows. Section 2.2 gives a detailed account of the DSA algorithm and the DSA based framework for mortality forecasting. Section 2.3 presents the numerical studies. Section 2.4 offers some further discussion and remarks.

2.2 DSA based Framework for Mortality Forecasting

This section introduces our DSA algorithm-based framework for mortality forecasting. The framework provides a fully data-driven method and is flexible to be applied with any multi-population joint models, while we will take the extended ACF model for illustration purposes in the chapter. The DSA algorithm is designed to generate a sequence of groups of populations in increasing size, and a validation procedure is applied to select the optimal group for a given target population. The mortality forecasting is conducted based on the selected optimal group.

2.2.1 Extended ACF model

The implementation of our proposed prediction framework entails the specification of a multi-population mortality model and the application of the designed DSA algorithm. For illustrative purposes, we adopt the Poisson common factor model by Li [2013] in our studies. Viewed as an extension of the ACF model of Li and Lee [2005], the Poisson common factor model allows more than one population-specific terms. For simplicity, we will refer this model as the extended ACF model in the following discussion of the chapter.

Given a group of populations G of size M, the extended ACF model describes logarithmic ASDRs of each individual population $i \in G$ by the following form:

$$\log m_{x,t,i} = a_{x,i} + B_x K_t + \sum_{j=0}^{N_i} b_{x,j,i} k_{t,j,i} + \epsilon_{x,t,i}, \tag{2.1}$$

where $a_{x,i}$, B_x, K_t and $\epsilon_{x,t,i}$ carry the same meanings as in the ACF model. The difference lies in that we allow more complexity in population-specific effect with $(b_{x,j,i}, k_{t,j,i})$ as pairs of population-specific components with N_i as the number of the population-specific components for population i. Constraints $\sum_x b_{x,j,i} = 1$, $\sum_t k_{t,j,i} = 0$, $\sum_x B_x = 1$ and $\sum_t K_t = 0$ are also imposed to avoid the unidentifiability issue for the extended model.

As an integrated part of the above extended ACF model, $a_{x,i}$ is calibrated as the average of logarithmic ASDRs over the modeling time-period for each population:

$$a_{x,i} = \frac{1}{T+1} \sum_{t=0}^{T} \log m_{x,t,i}, \tag{2.2}$$

but B_x and K_t are calibrated in a different way. The extended model applies the SVD procedure to $(\log m_{x,t,i} - a_{x,i})$ for each individual population i to obtain $B_{x,i}$ and $K_{t,i}$ as

the resulting first left and right singular vectors, and we compute B_x and K_t as the simple average of those $B_{x,i}$'s and $K_{t,i}$'s, respectively, i.e.,

$$B_x = \frac{1}{M}\sum_{i\in G} B_{x,i} \text{ and } K_t = \frac{1}{M}\sum_{i\in G} K_{t,i}. \tag{2.3}$$

There followed another round of SVD procedures applied to the residuals, $(\log m_{x,t,i} - a_{x,i} - B_x K_t)$, to calibrate the population-specific components. The first N_i left and right vectors are used to calibrate $b_{x,i,j}$ and $k_{t,i,j}$, $j = 1, \ldots, N_i$, respectively.

The above calibration procedure for B_x and K_t is in the same spirit of the model averaging idea in statistics. That is, averaging estimation of the same objective from different sources can potentially reduce the estimation uncertainty. So, we work with the assumption that the individual populations in the group G share the common component pair (B_x, K_t), and the estimation formed by an average of the calibrated value from each individual population can potentially lead to a more efficient estimate of (B_x, K_t). As an alternative to the simple average in Equation (2.3), a weighted average using population size can potentially further reduce the estimation uncertainty for the common factor (B_x, K_t), because there is a reason to believe that an estimation from a larger population tends to bring in less uncertainty. However, it is challenging to quantify the precise impact of the population size on the estimation uncertainty of the common factor; meanwhile, an enhanced performance of the weighted average highly relies upon this piece of information to be precisely incorporated into the calculation. For this reason, we adopt the simple average for the calibration of the common factor in this chapter.

For mortality forecasting, we fit the sequence K_t with a random walk with drift model to forecast the common trend in future mortality changes as in Li and Lee [2005]. In the meanwhile, we allow the population-specific components $k_{t,i,j}$ to embrace the potential benefits of generality by fitting them with the "best" ARIMA model, which is obtained by using the `auto.arima` function from the `R` package `forecast`. A detailed description of the ARIMA fitting procedure used by the `R` function can be found in Hyndman et al. [2007]. After obtaining the time series models for these components, we extrapolate them into years $t > T$ to form mortality forecasts for the target population i.

Some clarifications about the above extended ACF model are in order. The parameters M and N_i allow for different levels of model complexity. Tuning M enables the model to have the desired size of the group, and tuning N_i allows for the desired complexity of population-specific effects. As we will explain later, the rational choice for the parameters is no longer an issue since the values of M and N_i can be learned through a validation step during the modeling procedure. Moreover, this extended model includes the Lee-Carter and

the ACF models as special cases. When $N_i = 0$ and $M = 1$ (i.e., the group G only contains the target population i), the extended ACF model degenerates into the Lee-Carter model. When $N_i = 1$, the model reduces to the ACF model.

2.2.2 DSA Algorithm: Risk Functions and the Three Moves

In our framework, the data period (years 0 to T) is divided into a modeling period (years 0 to S) and a validation period (years $(S+1)$ to T), and the dataset is accordingly divided into a modeling set and a validation set. The DSA algorithm is applied with the modeling set only, and the validation procedure utilizes the validation set for selecting the optimal group.

The DSA algorithm starts with a group of size one containing the target population only. The algorithm generally utilizes three specific moves or step functions (i.e., deletion, substitution, and addition) to generate a sequence of group selections. For each group size, the algorithm aims to find a group of populations to minimize a pre-specified risk function over groups of the same size. Choices of risk function include the mean squared error (MSE), mean absolute error (MAE), mean squared percentage error (MSPE) and mean absolute percentage error (MAPE).

Given a group G containing the target population i, we calibrate the extended ACF model (2.1) with the modeling dataset. Denote

$$\log \widehat{m}_{x,t,i} = a_{x,i} + B_x K_t + \sum_{j=0}^{N_i} b_{x,j,i} k_{t,j,i},$$

where $N_i \in \{1, 2, \ldots, N_{max}\}$ is fixed, and with a slight abuse of notation, all the items on the right-hand side mean their calibrated values. Note that N_{max} denotes the maximum number of population-specific components that we consider in the extended ACF model and we take $N_{max} = 5$ in our numerical implementation in the sequel.

The risk function on the group G for target population i can be calculated as follows:

- MSE:
$$f(G) = \frac{1}{(\omega+1) \times (S+1)} \sum_{x,t} \left(\log m_{x,t,i} - \log \widehat{m}_{x,t,i}\right)^2, \tag{2.4}$$
- MAE:
$$f(G) = \frac{1}{(\omega+1) \times (S+1)} \sum_{x,t} \left|\log m_{x,t,i} - \log \widehat{m}_{x,t,i}\right|, \tag{2.5}$$

- MSPE:
$$f(G) = \frac{1}{(\omega+1)\times(S+1)} \sum_{x,t} \left(\frac{\log m_{x,t,i} - \log \widehat{m}_{x,t,i}}{\log m_{x,t,i}} \right)^2, \tag{2.6}$$

- MAPE:
$$f(G) = \frac{1}{(\omega+1)\times(S+1)} \sum_{x,t} \left| \frac{\log m_{x,t,i} - \log \widehat{m}_{x,t,i}}{\log m_{x,t,i}} \right|, \tag{2.7}$$

  where the summation is for x ranging over the considered set of ages and for t over the modeling period (i.e., years 0 to S). The right-hand side of the above formula depends on the group G via B_x and K_t.

With the parameter N_i fixed over the set $\{1, 2, \ldots, N_{max}\}$, the proposed algorithm searches to minimize the in-sample value of the adopted risk function (hereafter "risk value" for short) over groups of the same size. For each group size s, the optimal group $\mathbb{G}_s$ attains the minimum in-sample risk value among $\mathcal{G}_s$ that denotes the collection of all the subsets of populations in size s from the candidate pool, that is,

$$\mathbb{G}_s := \text{argmin}_{G\in\mathcal{G}_s} \ f(G).$$

It is computationally inefficient to search for the optimal $\mathbb{G}_s$ among the whole set $\mathcal{G}_s$ for every s. The DSA algorithm proceeds in a stepwise manner to complete the search. The result, though suboptimal, is more computationally efficient. It follows the same spirit as in many machine learning methods, such as regression tree methods and stepwise (forward or backward) regression procedures, to obtain a good balance between optimality and computational demands.

The DSA algorithm searches through specifically designed iterations and stops with the control of a specific stopping criterion. At each iterative step, the algorithm first maps the current group G of size s into groups of size $s-1$, s, and $s+1$, respectively, through moves or step functions, Deletion, Substitution, and Addition. Then, the algorithm solves an optimization problem to secure the optimal outcome throughout all possible results. The three moves are defined as follows:

- **Deletion**: A deletion move allows the removal of one population from the current group. Formally, given the current group G of size s, this move first returns set $\text{DEL}(G)$ that contains s different groups of size $s-1$ by deleting one member from the current group G. Then, the algorithm searches for the optimal element within $\text{DEL}(G)$, denoted as G^-, which has the smallest risk value among all possible outcomes, i.e.,
$$G^- := \text{argmin}_{G'\in\text{DEL}(G)} \ f(G').$$

- **Substitution**: A substitution move allows the replacement of one population in the current group with another from the remaining set in the candidate pool. Formally, given the current group G of size s, this move first returns set $\mathrm{SUB}(G)$ that contains a number of different groups of size s by making substitutions. Then the algorithm searches for the optimal element within $\mathrm{SUB}(G)$, denoted as $G^{=}$, which has the smallest risk value among all possible outcomes, i.e.,

$$G^{=} := \mathrm{argmin}_{G' \in \mathrm{SUB}_{(G')}} \ f(G').$$

- **Addition**: An addition move allows the introduction of one more candidate population into the current group. Formally, given the current group G of size s, this move first returns set $\mathrm{ADD}(G)$ that contains different groups of size $s+1$ by adding one member into the current group from the rest of the candidates. Then the algorithm searches for the optimal element within $\mathrm{ADD}(G)$, denoted as G^{+}, which has the smallest risk value among all possible outcomes, i.e.,

$$G^{+} := \mathrm{argmin}_{G' \in \mathrm{ADD}_{(G)}} \ f(G').$$

2.2.3 DSA algorithm: Ordering of Moves and Pseudo Codes

After having given an account of the risk function and the three allowed moves (i.e., deletion, substitution, and addition), we now move on to the ordering of the moves, the initiation, and the stopping criterion of the DSA algorithm. We need the following notations to record information:

- A list $\mathbb{G}^*$ to record the best groups of different group sizes. $\mathbb{G}_s$, as an element of $\mathbb{G}^*$, represents the best group of size s, $s = 1, 2, \ldots$.
- A vector BEST to record the in-sample risk values corresponding to each element of the list $\mathbb{G}^*$:

$$\mathrm{BEST}(s) = f(\mathbb{G}_s), \ \ s = 1, 2, \ldots.$$

Throughout the iterations of the algorithm, information contained in vectors G^* and BEST keeps updating untill the algorithm comes to a stop. The details of the algorithm are described as follows:

1. Initialization:

- The algorithm starts with a group of size one which contains the target population only, and retains the rest of candidate populations in the pool.
- $\mathbb{G}_1 = G_1$, since the best group of size one must be the group containing only the target population.
- The risk value $f(G_1)$ is calculated and BEST(1) is assigned to equal $f(G_1)$.
- A stopping value, called cut-off-growth (COG), is assigned to indicate the maximum number of candidate populations considered for joint mortality modeling. A small COG results in an early stop for the algorithm, whereas a large COG results in a computationally demanding algorithm. COG can be set as large as the size of candidate pool. With consideration having been given to computational feasibility, COG should be set as large as possible for the reason that its value corresponds with the class size of candidate groups being considered in the searching of the best group.

2. Move through step functions:

   - †Let G be the current working group and denote its size by s.
   - **Deletion** If $s > 3$, search for the optimal updated G^- of size $s-1$ among all allowed deletion moves, where $G^- = \text{argmin}_{G' \in \text{DEL}(G)} \ f(G')$. If $f(G^-) < \text{BEST}(s-1)$, put $\text{BEST}(s-1) = f(G^-)$, set $\mathbb{G}_{s-1} = G^-$ to extract the grouping information of G^- to update the optimal group of size $s-1$, update the current working group by $G = G^-$, and return to †.
   - **Substitution** If $s > 2$, find the optimal updated $G^=$ of size s among all allowed substitution moves, where $G^= = \text{argmin}_{G' \in \text{SUB}(G')} \ f(G')$. If $f(G^=) < \text{BEST}(s)$, put $\text{BEST}(s) = f(G^=)$, set $\mathbb{G}_s = G^=$ to extract the grouping information of $G^=$ to update the optimal group of size s, update the current working group by $G = G^=$, and return to †.
   - **Addition** Find an optimal updated G^+ of size $s+1$ among all allowed addition moves, where $G^+ = \text{argmin}_{G' \in \text{ADD}(G)} \ f(G')$. If $f(G^+) < \text{BEST}(s+1)$, put $\text{BEST}(s+1) = f(G^+)$, set $\mathbb{G}_{s+1} = G^+$ to extract the grouping information of G^+ to update the optimal group of size $s+1$, update the current working group $G = G^+$, and return to †.

3. Stop criterion: If $s = \text{COG}$, stop the algorithm.

The implementation of the above algorithm will result in a sequence of groups $\mathbb{G}^* \equiv \{\mathbb{G}_s, s = 1, \ldots, \text{COG}\}$, which gives the optimal choice of grouping for each size $s = 1, \cdots, \text{COG}$ in

the sense to attain the minimum in-sample risk value. Each element in the vector BEST records the in-sample risk value for the corresponding element in the vector $\mathbb{G}^*$. As a concise summary, the DSA algorithm is described by the pseudo-codes in Algorithm 1 and the flow chat in Figure 2.1.

2.2.4 The DSA based Prediction Model

For the target population i, the parameters M_i and N_i in the extended ACF model (2.1) are the group size and the number of population-specific components, respectively. With a fixed $N_i \in \{1, 2, \ldots, N_{max}\}$, the implementation of the DSA algorithm yields the "optimal" group of populations, denoted by G^*_{i,M_i,N_i}, for each group size $M_i \in \{1, \ldots, \text{COG}\}$. For each G^*_{i,M_i,N_i}, we can obtain a mortality model, so implementing the DSA algorithm for $N_i \in \{1, 2, \ldots, N_{max}\}$ yields $N_{max} \times$ COG mortality models in total.

To determine the best model for prediction, we resort to a validation procedure, which entails the selection of the best values of parameters M_i and N_i. The extended ACF model can be calibrated with the identified G^*_{i,M_i,N_i} from the DSA algorithm for each combination of M_i and N_i. We project the calibrated models into the validation period (i.e., years $S+1$ to T) and calculate the validation risk values using formulae (2.4)-(2.7) with the summation changed to be over the validation period. The optimal M_i and N_i are selected as those that achieve the lowest validation risk value (measured by one of the four metrics calculated as in Equations (2.4)-(2.7) using validation set). In so doing, we obtain the calibrated model corresponding to the optimal parameters M_i and N_i. Then, the obtained model is projected into the future periods for mortality forecasting.

The DSA-based prediction model can be summarized into a procedure of the following five steps:

1. **Initialization.** Specify the target population i, the corresponding candidate pool (all possible choices of other populations) and COG. Divide the training data of age-specific mortality rates into two periods: modeling period, and validating period.

2. **Implementation of the DSA algorithm.** For each fixed $N_i \in \{1, 2, \ldots, N_{max}\}$, apply the DSA algorithm with the modeling dataset to obtain a sequence of optimal groups G_{i,M_i,N_i}, $M_i \in \{1, 2, \ldots, \text{COG}\}$, which gives the lowest in-sample risk value for each group size M_i.

3. **Validation.** The optimal group G^*_{i,M_i,N_i} emerges by choosing the smallest validation risk among all the COG $\times N_{max}$ models identified from the DSA algorithm.

Algorithm 1: DSA algorithm for optimal grouping selection

Input: Target population, Candidate Pool, COG

Output: $\mathbb{G}^*$ (a sequence of groups in increasing size), and BEST (a vector of risks values)

Initiation: $s = 1$, $G = G_1$, $\mathbb{G}_1 = G_1$, BEST(1) $= f(G_1)$, BEST(s) $= \infty$ for $s = 2, \ldots,$ COG;

† **Load information:** Current group G of size s;

while $s < COG$ **do**

  if $s > 3$ **then**

    Deletion: Find the optimal G^- of size $s-1$;

    if $f(G^-) < BEST(s-1) = f(\mathbb{G}^*{}_{s-1})$ **then**

      BEST$(s-1) = f(G^-)$;

      Set $\mathbb{G}^*{}_{s-1} = G^-$, $G = G^-$, and return to †;

    end

  end

  if $s > 2$ **then**

    Substitution: Find the optimal $G^=$ of size s;

    if $f(G^=) < BEST(s) = f(\mathbb{G}_s)$ **then**

      BEST$(s) = f(G^=)$;

      Set $\mathbb{G}^*{}_s = G^=$, $G = G^=$, and return to †;

    end

  end

  Addition: Find the optimal G^+ of size $s+1$;

  if $f(G^+) < BEST(s+1) = f(\mathbb{G}_{s+1})$ **then**

    BEST$(s+1) = f(G^+)$;

    Set $\mathbb{G}_{s+1} = G^+$, $G = G^+$, and return to †;

  end

end

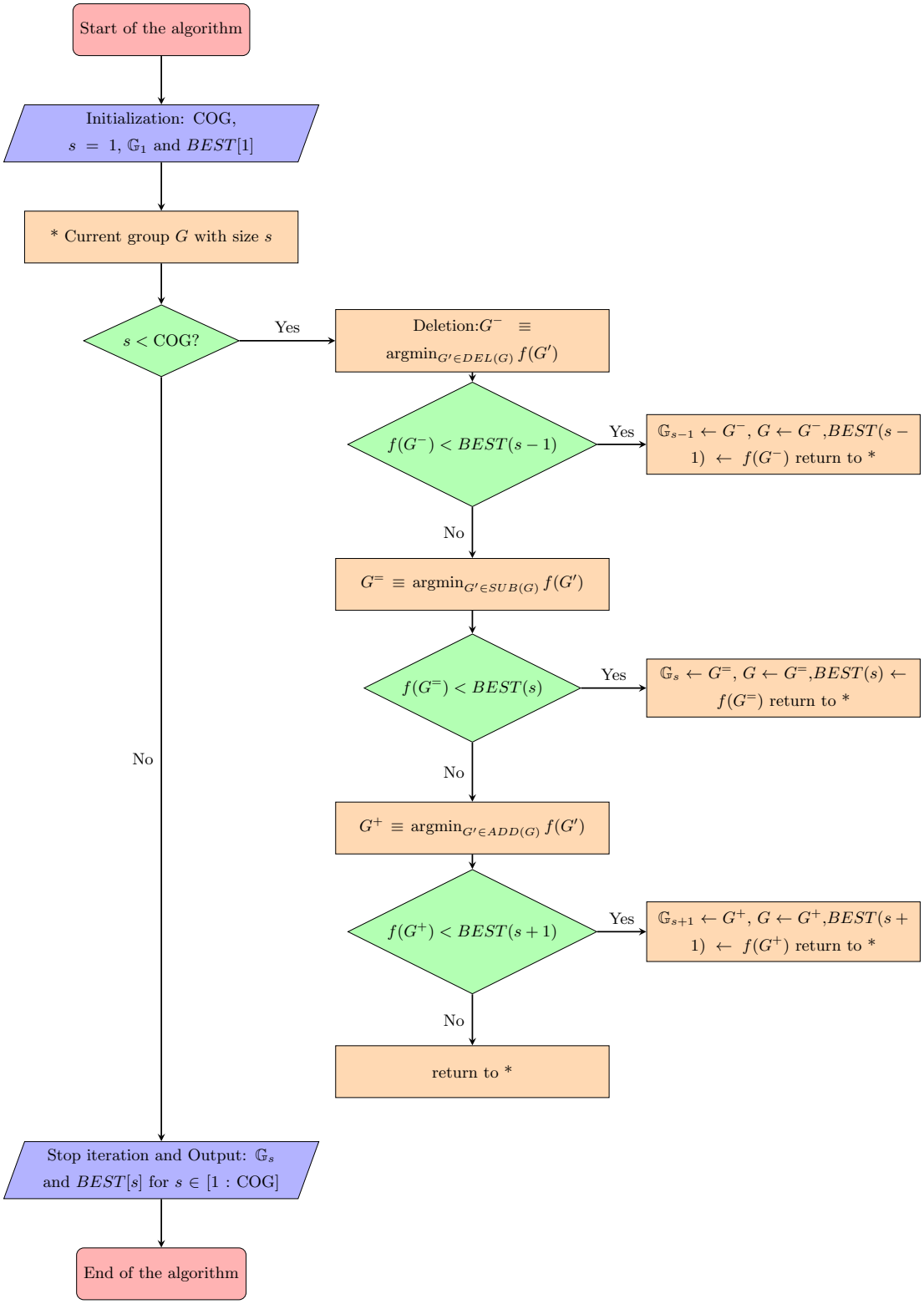


Figure 2.1: Illustration of DSA: the flow chart

4. **Parameter Estimation.** With the determined values of M_i and N_i as well as the selected group G^*_{i,M_i,N_i} from the preceding step, we re-calibrate the selected model using the whole training dataset (including both the modeling and validating datasets).

5. **Forecasting.** We project the re-calibrated model into the future periods for mortality forecasting.

2.3 Numerical analysis

2.3.1 Data and Benchmark Models

We apply our DSA-based procedure to a set of populations from the HMD for mortality prediction. The HMD is available for download on the website: https://www.mortality.org/. The HMD contains data from 46 countries or regions, with different scopes and lengths of time. We selected 30 out of the 46; see Appendix A.1 for the detailed list.

As mentioned in Section 1.1.2, we will focus on mortality data from 1970 to 2010. We applied the following principles to the choices of populations. The mortality data of 34 countries/regions from 1970 to 2010 are chosen in our study. Our assumption is that the more recent the data, the more relevance it bears on current and future mortality. Therefore, our study does not take into consideration the time period before the Second World War. This provision may also help to avoid the potential impact of rare events or pattern changes. In addition, a moderately sufficient amount of data are preferred to learn a prediction model. That being said, we do our best to maintain a balance between the size of the candidate pool and the length of the time period for the data. For countries and regions that have mortality data with multiple scopes, only one specific scope is chosen to avoid double counting. Four countries/regions (i.e., Estonia, Iceland, Luxembourg, and Northern Ireland) are left out because their datasets contain many missing data and zero death rates. We use interpolation to fix missing data, and replace zero mortality rates with the average value of the same age group from the years preceding and succeeding the given year.

For each of the 30 populations, we consider mortality forecasts for both genders. Each gender of the same population is treated separately, thus we actually have 60 populations in our study. When one of the 60 populations is the target for prediction, the rest of the 59 populations form the candidate pool, to which we apply the DSA algorithm. We compare our model with the following benchmark models:

- **Lee-Carter**: The Lee-Carter model will be fit for each of the 60 populations independently, using mortality rates in the whole training set. The sequence of k_t obtained from the SVD procedure is fitted using the `auto.arima` function from the `R` package `forecast` to search for a suitable ARIMA model.

- **ACF·AIO**: Each gender consists of 30 populations, and the ACF model is fitted with all the 30 populations jointly.

- **ACF·GeoInfo**: The 30 populations of each gender is divided into 8 groups based on geographic proximity (for grouping information, see Appendix A.1). The grouping basically follows Richman and Wüthrich [2021] with a few necessary adjustments, because the populations in our study are not entirely identical to theirs. For each gender, the ACF model is fitted to each geographic group.

- **ACF·kmeans**: For each gender, groups are obtained by k-means cluster method applied to logarithmic mortality rates data. Then the ACF model is fitted to each group. We set parameter $k = 8$ (the same number of groups as in ACF·GeoInfo) in the clustering algorithm and apply the best clustering result from 1,000 independent random initializations in the clustering algorithm.

- **ACF·kmedian**: The same as ACF·kmeans except that groups are obtained by k-median cluster method.

Regarding the above benchmark models, two further points are worth mentioning:

(a) Since our DSA based procedure fits all the time trend sequences with ARIMA models, we fit an ARIMA model to the sequence k_t obtained from the Lee-Carter model so that the results can be comparable.

(b) The last two benchmark models in the above list are resulted from two prevailing clustering algorithms, the k-means and the k-median. It should be noted that k (or K) in these two clustering algorithms is a standard notation in the machine/statistical learning literature, and it differs from the time trend sequence k_t from the Lee-Carter model.

2.3.2 Numerical Procedure

The numerical procedure is carried out throughout our study in the following steps:

1. We split the mortality dataset (consists of data from 1970 to 2010) into a training set (1970-2002) and a test set (2003-2010). This gives roughly 80% of the data for training and 20% for testing. The training set is further split into a modeling set (1970-1993) and a validation set (1994-2002).

2. We set each of the 60 populations as the target for prediction and take the rest 59 gender-specific populations as the candidate pool. The DSA algorithm proceeds with the following specifications:

   (a) We use each of the four metrics defined in Equations (2.4)-(2.7) as the risk function in the DSA algorithm. The resulting prediction models are, respectively, labeled as DSA·MSE, DSA·MAE, DSA·MSPE, DSA·MAPE.

   (b) COG = 30. The COG is set in the DSA algorithm to maintain a balance between model generality and computational demand.

   (c) $N_{max} = 5$. This parameter controls the complexity of population-specific effects which are contained in the residuals after extracting the common factor in the extended ACF model. The choice of $N_{max} = 5$ is consistent with the setup of $p = 5$ considered by Booth et al. [2002] in the rank-p SVD approximation for mortality modeling.

3. The model validation step is implemented in accordance with the procedure described in Section 3.3 to determine the optimal group for each target population. Then, the whole training dataset (including both the modeling and validating sets) is used to recalibrate the parameters in the determined model.

4. For each target population, by using the re-calibrated model from the preceding step, we form forecasts of logarithmic ASDRs over the testing period (i.e., 2003-2010). We compute the test risk value for each of the four metrics (i.e., MSE, MAE, MSPE, and MAPE) in Equations (2.4)-(2.7) with the summation taken over the testing period.

5. For each benchmark model (i.e., Lee-Carter, ACF·AIO, ACF·GeoInfo, ACF·kmeans, and ACF·kmedian), we calibrate with the training data (including both the modeling and the validating sets), extrapolate the calibrated model into the testing period to obtain mortality forecasts, and then calculate risk values for each risk metric as we do with the DSA-based model.

2.3.3 Prediction Performance

This subsection compares the prediction performance of our DSA based models with that of the benchmarks through summary statistics and a formal hypothesis test. The comparison is based on the test risk values (MSE, MAE, MSPE, or MAPE) from the 60 populations described in Section 2.3.1.

2.3.3.1 Comparison via Summary Statistics

The numerical procedure in Section 2.3.2 with each prediction model yields a test risk value for each population. A smaller test risk value is more desirable, as it ensures a better performance for a prediction model. Table 2.1 reports the 1st-quartile, 3rd-quartile, mean, and median of the test risk values of the 60 populations obtained under different combinations of risk metrics and prediction models. These summary statistics show that our DSA-based prediction models generally yield smaller risk values, that is, they perform better than all benchmarks. In particular, DSA·MSE and DSA·MAE consistently outperform other prediction models in terms of MSE, MAE, and MAPE. Even under the risk metric MSPE, the performance of these two DSA models is comparable to all benchmark models if not better. The DSA·MSPE model, though having less satisfactory performance than other DSA models, is nevertheless competitive when compared with the benchmark models.

Figure 2.2 gives boxplots of the test risk values of 60 populations under various prediction models evaluated using each risk metric. The test MSPE values are rather dispersive, and a zoom-in version with all outliers excluded is given in Figure 2.3. These figures indicate that the DSA-based models generally yield smaller mean and median, fewer outliers, and smaller variations of the resulting risk values. Therefore, the performance of our DSA model is superior to that of other models, regardless of the risk function that is applied in the DSA step and the metric that is used in calculating the test risk values.

2.3.3.2 Comparison via Diebold-Mariano Test

For a formal comparison of prediction performance, we conduct the one-sided Diebold-Mariano (DM) test to determine if the prediction error of a given model is statistically significantly smaller than that of another. The DM test, introduced by Diebold and Mariano [1995] and further improved by Harvey et al. [1997], is a formal statistical hypothesis testing method in the field of forecast comparison. The test is based on two sequences of forecast values from two predictive models. To implement the DM test, we made some modifications

Table 2.1: Summary statistics of test risk values of the 60 populations.

MSE	1st quartile	Median	Mean	3rd quartile
Lee-Carter	0.0397	0.0680	0.1409	0.0886
ACF·AIO	0.0291	0.0538	0.0579	0.0794
ACF·GeoInfo	0.0324	0.0629	0.0589	0.0810
ACF·kmeans	0.0318	0.0576	0.0611	0.0773
ACF·kmedian	0.0304	0.0520	0.0583	0.0851
DSA·MSE	0.0233	0.0491	0.0518	0.0751
DSA·MAE	0.0238	0.0494	0.0512	0.0756
DSA·MSPE	0.0276	0.0519	0.0580	0.0811
DSA·MAPE	0.0277	0.0491	0.0548	0.0758
MAE	1st quartile	Median	Mean	3rd quartile
Lee-Carter	0.1406	0.1644	0.1917	0.1971
ACF·AIO	0.1227	0.1575	0.1567	0.1889
ACF·GeoInfo	0.1237	0.1660	0.1567	0.1838
ACF·kmeans	0.1254	0.1560	0.1554	0.1793
ACF·kmedian	0.1248	0.1558	0.1573	0.1860
DSA·MSE	0.1105	0.1445	0.1465	0.1792
DSA·MAE	0.1104	0.1480	0.1455	0.1751
DSA·MSPE	0.1102	0.1520	0.1558	0.1921
DSA·MAPE	0.1126	0.1444	0.1495	0.1749
MSPE	1st quartile	Median	Mean	3rd quartile
Lee-Carter	0.0013	0.0026	0.0087	0.0053
ACF·AIO	0.0012	0.0027	0.0048	0.0045
ACF·GeoInfo	0.0011	0.0026	0.0057	0.0052
ACF·kmeans	0.0012	0.0024	0.0054	0.0049
ACF·kmedian	0.0012	0.0026	0.0049	0.0048
DSA·MSE	0.0012	0.0022	0.0055	0.0043
DSA·MAE	0.0012	0.0021	0.0052	0.0047
DSA·MSPE	0.0012	0.0022	0.0056	0.0044
DSA·MAPE	0.0014	0.0023	0.0049	0.0039
MAPE	1st quartile	Median	Mean	3rd quartile
Lee-Carter	0.0262	0.0353	0.0414	0.0436
ACF·AIO	0.0252	0.0324	0.0343	0.0415
ACF·GeoInfo	0.0247	0.0343	0.0346	0.0433
ACF·kmeans	0.0240	0.0311	0.0347	0.0399
ACF·kmedian	0.0254	0.0312	0.0346	0.0420
DSA·MSE	0.0242	0.0309	0.0330	0.0386
DSA·MAE	0.0235	0.0295	0.0324	0.0394
DSA·MSPE	0.0256	0.0311	0.0344	0.0422
DSA·MAPE	0.0251	0.0312	0.0328	0.0383

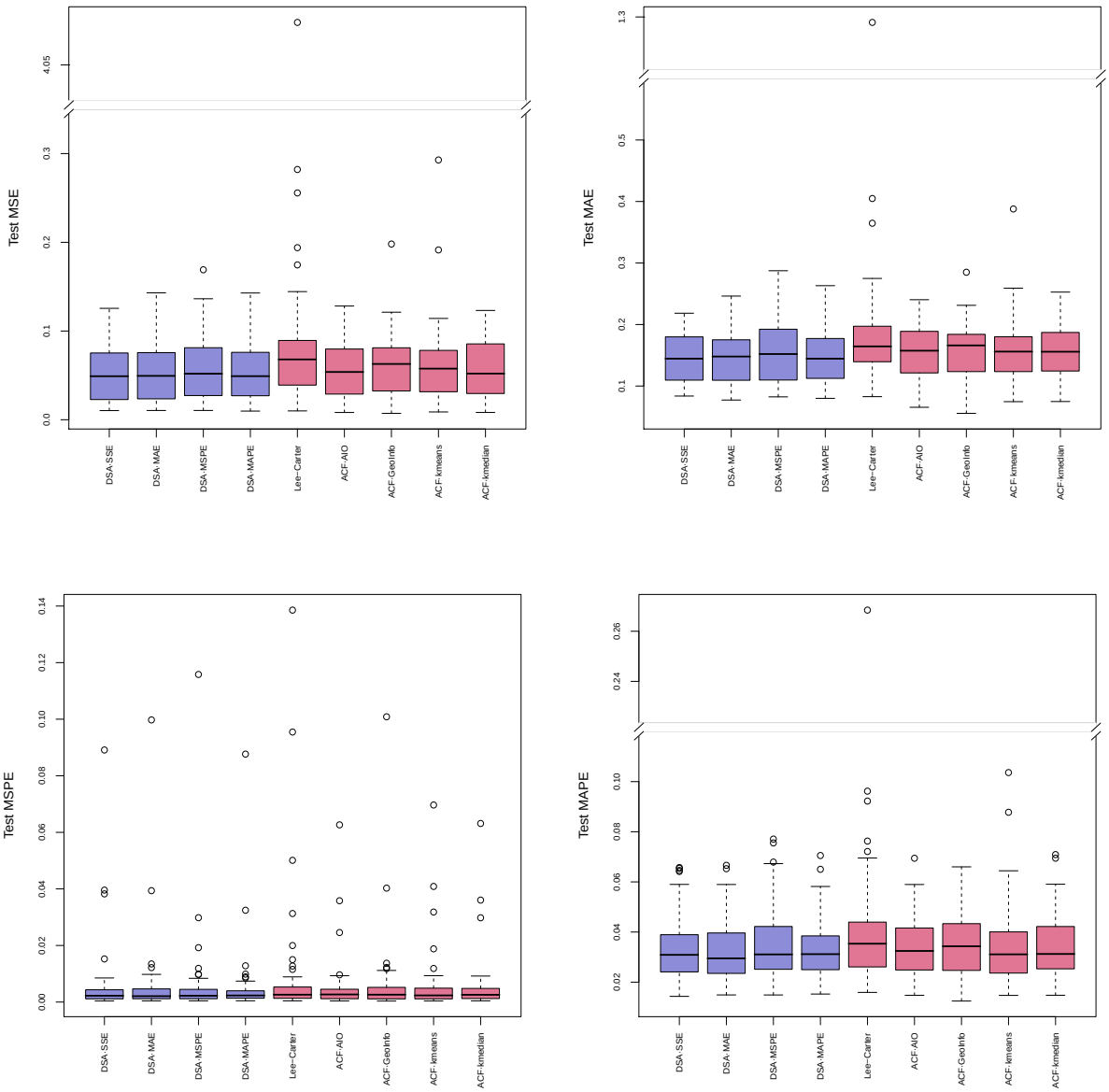


Figure 2.2: Boxplots of test risk values of the 60 populations.

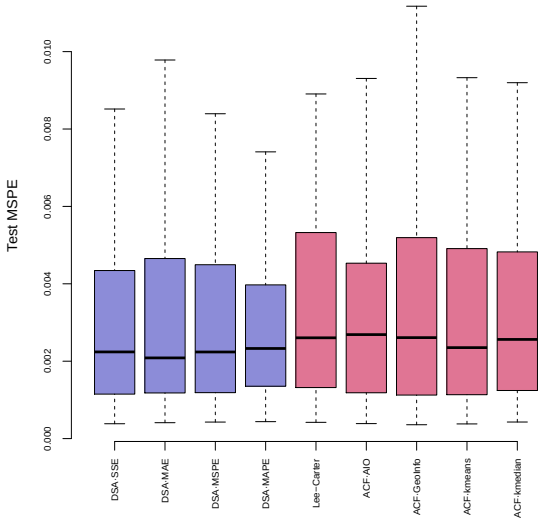


Figure 2.3: Boxplot of test MSPE values with outliers discarded.

to steps 4 and 5 in the numerical procedure described in Section 2.3.2. We do the summation in formulae (2.4)-(2.7) only for age and calculate risk values for each year over the testing period (2003-2010). This gives us 8 test risk values for each prediction model along with each risk metric.

For each metric and each target population, the DM test is conducted pairwise between one DSA model and one benchmark model. To facilitate the exposition of the pairwise test procedure, we label the pair of models in comparison by Model A and Model B. As has been noted earlier, a sequence of 8 test risk values is associated with each prediction model. Two one-sided DM tests are available to us, respectively, corresponding to the null hypothesis, that is Model A is no worse than Model B, and the opposite. When Model A is confirmed to be significantly better than the other in the pair by the test with a p-value smaller than 0.05, we count it as a win by Model A over Model B. Similarly, if Model B is confirmed to perform better by the test with a p-value smaller than 0.05, we count it as a win by Model B over Model A. We do the tests for each of the 60 target populations and total all the wins obtained by one model over the other. The number of wins by Model A and that by Model B do not necessarily sum to 60, because the question is still open as to whether Model B necessarily wins when the test does not confirm a win by Model A.

Table 2.2 reports the number of wins by one model over another under each performance risk metric. Each cell of the table includes two integers. The first integer is the number of wins by the corresponding model in the row over the model on the column, and the second

integer is the number of wins by the model in the column over the one in the row. For example, "(49, 1)" in the first cell from the left of the first row in the panel of MSE means that the DSA·MSE model wins 49 times over the Lee-Carter model, and the Lee-Carter model wins the DSA·MSE model only once among all the 60 comparisons. For the rest of the 10 populations, we are unable to deduce which model performs statistically better.

Table 2.2 clearly indicates, according to the DM test results, the DSA based models generally perform much better than all the benchmark models. The superiority of the DSA based models is all the more evident when compared with the Lee-Carter, the ACF·AIO, and the ACF·kmedian models, regardless of the applied performance metric. The ACF·GeoInfo and the ACF·kmeans have the best performance among the set of benchmark models, but not as good as the DSA based models do in general. However, exceptions do occur when the DSA·MSPE model is applied.

The population-specific comparison is also conducted between the DSA·MSE model and the benchmark ACF·GeoInfo model based on the test MSE on 30 female populations and listed in Table A.2. Based on the results, the proposed method has led to an increase in predicting accuracy for 24 out of 30 female populations with an averaged 15.59% improvement.

2.3.3.3 Comparison upon Gender-specific Populations

We show the boxplots of test risk values for the 30 female populations in Figure 2.4 and those for the 30 male populations in Figure 2.5 in order to investigate the performance of the DSA-based model in relation to the benchmark models when applied to gender-specific populations. In general, while the prediction errors for female mortality seem to be stable, those for male mortality rates are volatile, regardless of the risk metric used for performance measure. The prediction quality of every prediction model is higher in the female population than in the male. Even so, the superiority of the DSA-based models over the benchmarks is still evident on the boxplots from the populations of either gender.

2.3.4 Some Further Observations on the DSA based Models

Figure 2.6 demonstrates the distribution of the optimal group size M_i and the population-specific component size N_i selected by each DSA-based model over the 60 populations. The figure shows that the DSA-based models select varying group sizes and component sizes across the 60 target populations. Only a small portion of the 60 populations are assigned with a group size of one by the DSA-based models. Accordingly, we believe that, in terms

Table 2.2: Wins of one prediction model over another under the criterion of a p-value less than 0.05 from the one-sided DM test. The first element in cell is the number of wins (from the comparison over all the 60 populations) by the corresponding model in the row over the model on the column.

MSE	Lee-Carter	ACF·AIO	ACF·GeoInfo	ACF·kmeans	ACF·kmedian
DSA·MSE	(49, 1)	(34, 9)	(30, 12)	(30, 16)	(35, 9)
DSA·MAE	(47, 4)	(37, 10)	(29, 10)	(27, 20)	(36, 11)
DSA·MSPE	(40, 7)	(27, 19)	(19, 20)	(21, 21)	(27, 17)
DSA·MAPE	(42, 5)	(27, 14)	(27, 16)	(25, 21)	(26, 15)
MAE	Lee-Carter	ACF·AIO	ACF·GeoInfo	ACF·kmeans	ACF·kmedian
DSA·MSE	(45, 2)	(34, 13)	(31, 13)	(23, 17)	(36, 10)
DSA·MAE	(46, 5)	(39, 10)	(31,12)	(25, 17)	(39, 10)
DSA·MSPE	(36, 11)	(29, 20)	(23, 21)	(19, 24)	(30, 21)
DSA·MAPE	(43, 5)	(28, 13)	(22, 20)	(21, 17)	(29, 14)
MSPE	Lee-Carter	ACF·AIO	ACF·GeoInfo	ACF·kmeans	ACF·kmedian
DSA·MSE	(29, 9)	(22, 12)	(26, 10)	(23, 13)	(24, 7)
DSA·MAE	(28, 5)	(18, 8)	(24, 9)	(20, 12)	(21, 7)
DSA·MSPE	(26, 7)	(22, 17)	(18, 13)	(16, 18)	(21, 16)
DSA·MAPE	(24, 9)	(22, 12)	(15, 13)	(18, 15)	(22, 13)
MAPE	Lee-Carter	ACF·AIO	ACF·GeoInfo	ACF·kmeans	ACF·kmedian
DSA·MSE	(37, 6)	(34, 15)	(26, 17)	(22, 18)	(34, 13)
DSA·MAE	(36, 4)	(32, 6)	(29, 13)	(23, 16)	(33, 6)
DSA·MSPE	(30, 10)	(25, 17)	(21, 22)	(18, 22)	(25, 17)
DSA·MAPE	(31, 7)	(24, 15)	(20, 19)	(21, 20)	(25, 14)

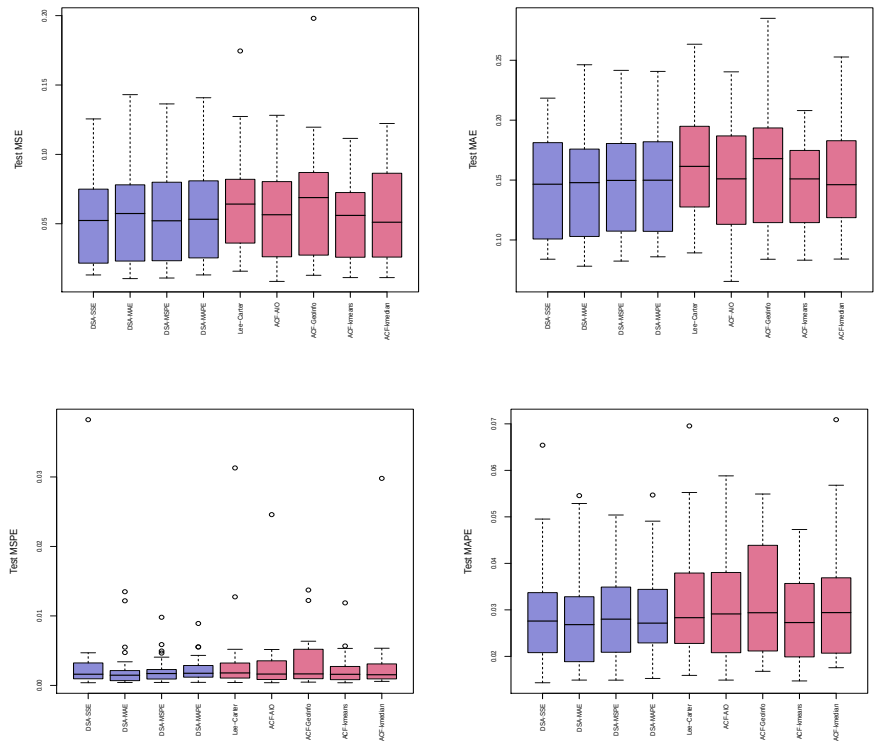


Figure 2.4: Boxplots of test risk values from the 30 female populations.

of mortality prediction, the DSA algorithm considers the single-population model as less competitive for most populations. Moreover, we observe that the DSA-based procedure selects more than one population-specific component, i.e., $N_i > 1$, for a majority of the 60 populations. That is to say, the extension that includes more than one population-specific component in model (2.1) helps to enhance the performance of mortality prediction.

We also explore the membership of the optimal group selected by the DSA-based models for each target population and conduct a comparison with the grouping information from the geographic proximity specified in Appendix A.1. For the sake of brevity, we focus on the DSA·MSE model and present four specific examples in Table 2.3. The membership results ensure that our DSA-based procedure not only produces groups that are consistent with geographic information to a certain degree but also makes further improvements. In some cases, when the group based on geographic proximity is satisfactory, our DSA-based procedure proceeds to choose a further subset as the optimal one. For instance, the group of female Hungarians belongs to this category. In other cases, discrepancies occur between the optimal group from the DSA procedure, and the geographic proximity information, then several representative populations from other parts of the world are added to the optimal group. For example, female Canadians are added to the group of female Taiwanese.

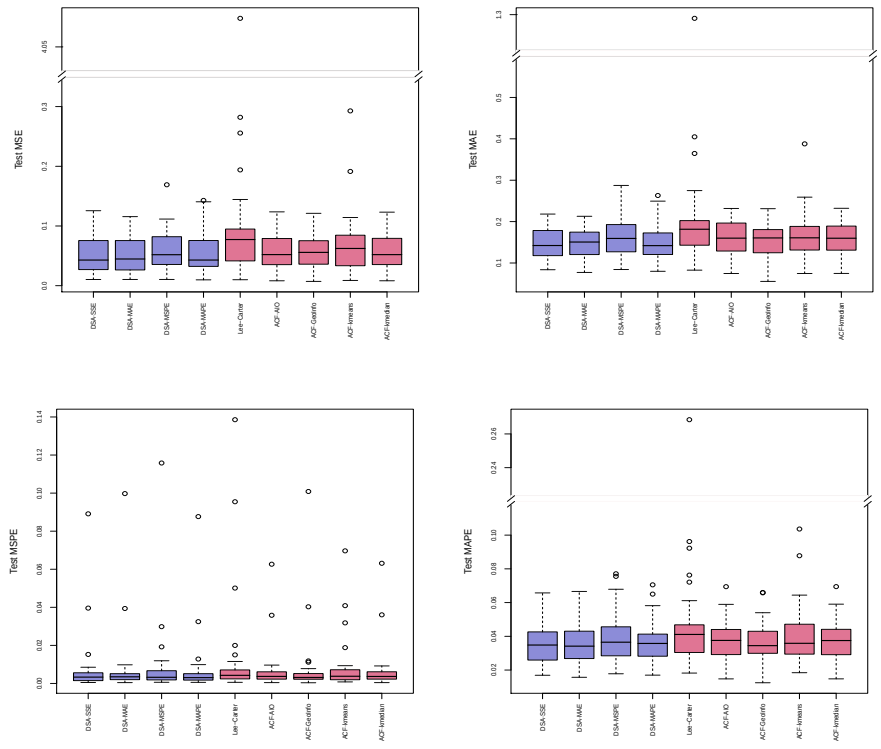
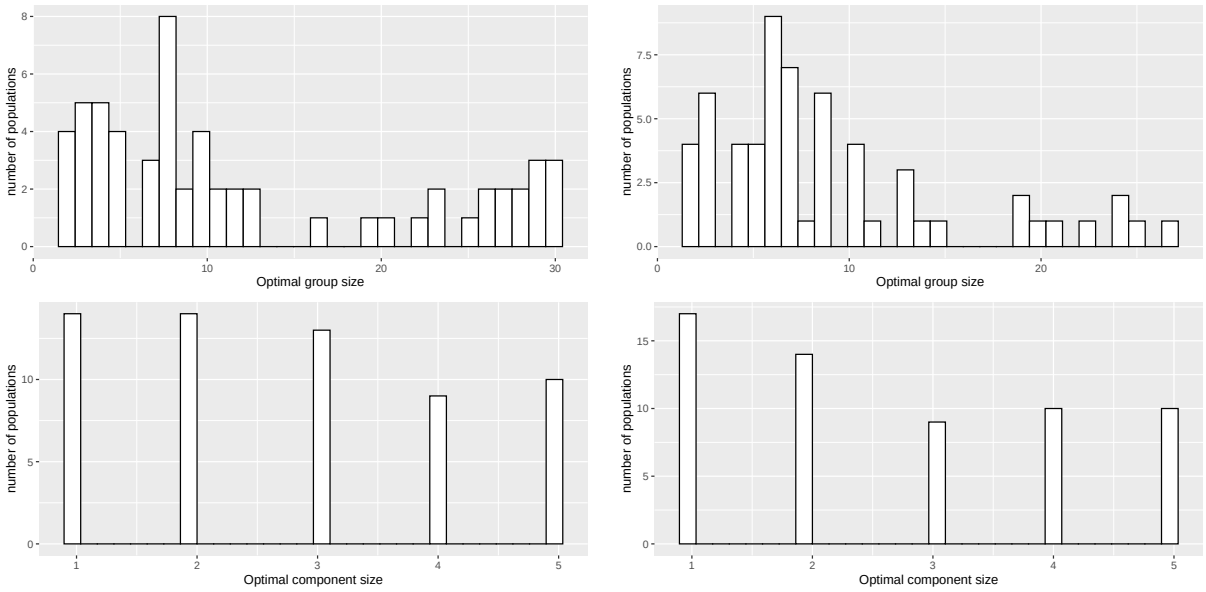


Figure 2.5: Boxplots of test risk values from the 30 male populations.

This happens when the prediction accuracy can be enhanced with the group that contains not only geographically proximate populations but also distant ones with similar mortality development patterns. In further other cases, the target population has relatively unstable mortality patterns. In order to offset the adverse effects of volatility in the mortality rates of the target population and avoid obtaining unstable prediction results, some populations from other parts of the world instead of being geographically proximate are added to the optimal group. One such example is the Bulgarian female population.

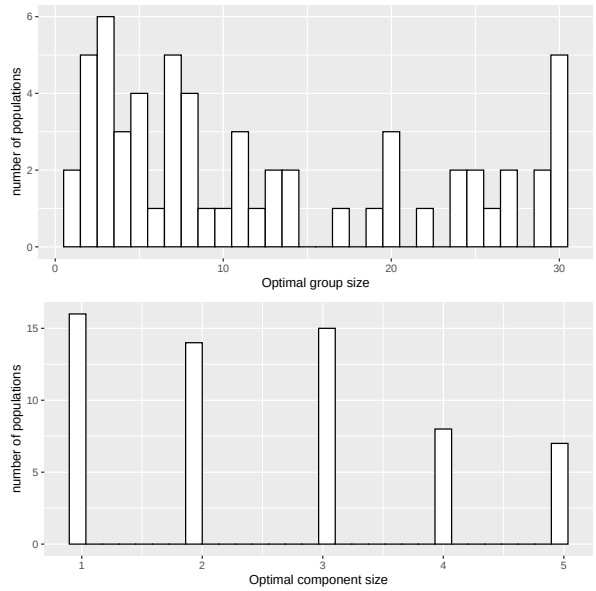
2.4 Concluding Remarks

Our data-driven framework selects populations from any given candidate pool to enhance mortality forecasting of individual populations. The DSA algorithm is the key element of our framework, designed to screen populations for joint modeling. The framework is fully data-driven and flexible to embrace any multi-population mortality model. Numerical analysis with the Human Mortality Database is conducted to confirm that the performance of the proposed DSA-based prediction method is indeed superior to many prevailing benchmark

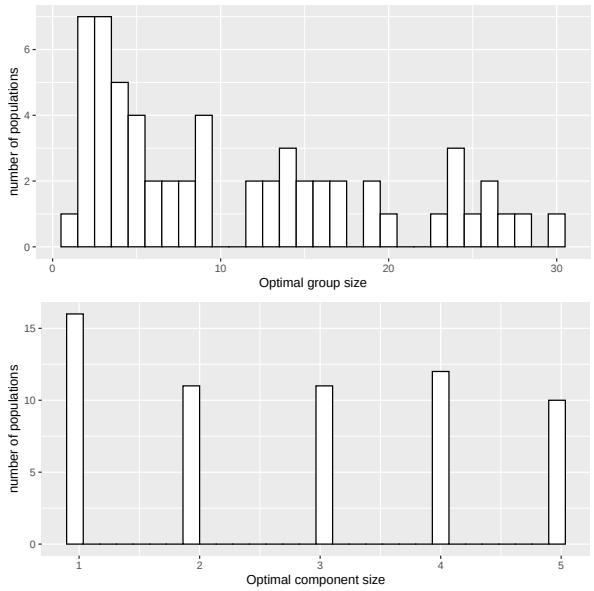


(a)
DSA·MSE

(b)
DSA·MAE



(c)
DSA·MSPE



(d)
DSA·MAPE

Figure 2.6: Distributions of M_i (group size) and N_i (number of population specific components) selected by the DSA based prediction model over the 60 populations.

Table 2.3: Some examples of membership in the optimal group selected by the DSA·MSE model.

Target Population	**Members**	**Test MSE**	
		DSA·MSE	**ACF·GeoInf**
Bulgaria.female	Bulgaria.female Australia.female Czech Republic.female Denmark.male Japan.female Latvia.female Latvia.male Norway.female Russia.male Slovakia.male New Zealand.male Scotland.male	0.0481	0.0961
Hungary.female	Hungary.female Czech Republic.male	0.0529	0.0853
Canada.female	Canada.female Canada.male Austria.male Japan.male Japan.female Netherlands.male Taiwan.female U.S.A.female	0.0168	0.0198
Taiwan.male	Taiwan.male Switzerland.male Portugal.female Czech Republic.male Japan.female	0.0199	0.0360

models.

The implementation of the DSA algorithm in our framework yields a sequence of groups in increasing size. Thus, the algorithm can be treated as a stepwise forward procedure for the selection of populations. It starts with a group consisting exclusively of the target population. Then, more populations are added to gradually increase the group size, though deletion and substitution moves might be repeated in the course of the procedure. An alternative to the DSA algorithm is the Addition-Substitution-Deletion (ASD) algorithm, which reverses the order of the three moves in the DSA algorithm. Different from the DSA algorithm, the ASD algorithm starts with a group composed of the target population and all the candidates from the pool. Then, the group size is gradually decreased by removing certain populations from the group. Both the DSA and the ASD algorithms can eventually provide a sequence of grouping results for model selection in the validation step. Instead of the ASD algorithm, we choose the DSA algorithm for our mortality forecast framework largely because of its computational efficiency. The DSA algorithm allows the users to cut off the search process by setting the parameter COG and controlling the computational demand in running the algorithm. When a COG smaller than the total size of the input candidate pool is set, the algorithm will search for optimal groups in size only up to the designated value of COG, and discard all the groups in size bigger than COG. On the contrary, if we need to control the computational demand in the ASD algorithm and set a value for COG smaller than the total number of the candidate pool, the algorithm will leave out all the groups in size smaller than COG in the search for the best grouping.

Quantification of uncertainty is one of the essential aspects of any forecasting procedure. With respect to the forecasting of mortality rates, the fan charts are commonly employed for uncertainty quantification. The fan charts delineate confidence intervals for the future mortality rates over time. Examples can be found in Hatzopoulos and Haberman [2009], Li et al. [2009] and Cairns et al. [2011b]. The implementation of our DSA framework yields a joint model of mortality rates for the target population along with a set of selected auxiliary populations. We can simulate a great number of paths for future mortality rates from the resulting mortality model, and then generate fan charts from the simulated paths. Since it is quite a conventional procedure to build the fan charts and the most innovative part of our DSA-based framework lies in its flexibility to embrace various joint models and information from a given candidate pool of populations, we choose not to delve into the details of quantification of forecast uncertainty in the chapter.

When conducting a comprehensive evaluation of a mortality forecasting model, many factors are worthy of our attention, and empirical prediction accuracy is merely one of them. For instance, Cairns et al. [2009] and Haberman and Renshaw [2011], among others,

have considered various desirable properties, such as parsimony, transparency, generation of simple paths, cohort effects, nontrivial correlation structure, robust parameter estimates, biological reasonableness. It is only in terms of the empirical prediction accuracy that the present chapter asserts the superiority of the DSA-based models over the benchmarks. On no account do we claim the superiority of our model in any other aspect. For example, the ACF model using stationary times series models, including the random walk with zero drift and AR(1) model, for the population-specific components, can yield coherent mortality forecasting results. By extending the models to the general ARIMA model, the resulting model can no longer generate coherence.

Chapter 3

Bivariate Model Based Ensemble and Time Shifting

3.1 Introduction

In this chapter, we propose an ensemble statistical learning framework that allows for borrowing information from the mortality data of a given pool of auxiliary populations to enhance the accuracy of mortality forecast for a target population. In this ensemble framework, we propose to use bivariate mortality models as the base learners. We establish a sequence of bivariate mortality models between the target population and each auxiliary population from the given pool and then apply a certain averaging strategy to ensemble all the estimates from those established bivariate models. The idea of averaging over a sequence of estimates for the same prediction target was commonly used in meta-analysis in statistics and stacking ensemble methods in machine learning, e.g., Claeskens and Hjort [2008], Raftery et al. [1997], Hansen and Racine [2012], and references therein. Recently, the averaging idea has been implemented in some literature for mortality prediction. First, the model averaging method has been applied by Shang [2012], Shang and Haberman [2018], and Shang and Booth [2020]. Second, the stacking regression ensemble methods have been used in Kessy et al. [2021]. These studies fit different types of single-population mortality models to the mortality data of a given population and then apply different model averaging strategies to aggregate the forecasts from various models to obtain a final forecast. In contrast, we utilize the averaging approach novelly to borrow information from multiple auxiliary populations to enhance the forecasting accuracy for a target population. We aggregate predictions for a target population that are obtained from a cascade of bivariate-population models of the same type (i.e., base learners), each built upon the data of the

target population and one of the auxiliary populations.

Our proposed bivariate model based ensemble (BMBE) framework facilitates borrowing information from multiple populations, and in the meanwhile, circumvents the potential computational difficulties and intractability of a large-dimensional multi-population model. We investigate various "averaging strategies" including a simple average over the whole pool of auxiliary populations, an average within geographical subgroups, and an average within k-means clusters. We also propose a data-driven "rank and average" strategy which ranks auxiliary populations according to their capability of improving the accuracy of the mortality forecast for the target population and averages over those top-performed ones selected by a cross-validation procedure.

Another merit of the proposed BMBE framework is its flexibility in working with different base learners. When bivariate-population models such as ACF and CBD models are adopted as the base learner, the flexibility of our proposed framework allows us to add a parameter Δt into the common mortality trend component to characterize the time by which one population is ahead of or behind the other in their mortality development stages. According to Zhou et al. [2014], some populations tend to be more dominating while others tend to follow their mortality dynamics. For instance, a developing country may demonstrate a similar mortality improvement pattern in the recent decade to what a developed country experienced earlier in the 1990s. Furthermore, as mentioned in Section 5.4 of Li et al. [2015b], for two populations, the period effect estimates, which capture the mortality developing trend in time, usually roughly co-move with one another but are located at different absolute levels. These observations motivate us to study the effect of time shift on the accuracy of mortality forecast under our proposed BMBE framework. Leads and lags in mortality have also been studied in Milidonis and Efthymiou [2017] with the belief that developed populations would lead the mortality changes in the less developed populations.

We apply the proposed BMBE prediction method to mortality data of 24 populations of both genders from the Human Mortality Database (HMD). We conduct two empirical studies. In the first empirical study, we adopt the ACF model (with and without a time shift component) as the base learner and apply the BMBE method with various averaging strategies to predict mortality rates of ages between 0 and 100. In the second empirical study, we take the CBD model (with and without a time shift component, with and without a cohort effect term) as the base learner because we aim to predict mortality rates for seniors aged 55-90. The outperformance of the proposed BMBE over several benchmark prediction methods is profound in the first empirical study where the ACF model is used as the base learner and the inclusion of the time shift component can further improve the forecasting accuracy. The second empirical study reveals some key points of using the model

average approach: 1) If the utilized base learner is a misspecified bivariate-population model for most pairs of populations, the biases arising in individual bivariate-population models cumulate and may lead to unsatisfactory prediction results in the BMBE framework. Lessening misspecification in base learners is the key to success when using the model average approach. 2) A weak base learner with fewer model assumptions predicts less accurately than a strong base learner with more model assumptions if these model assumptions are correctly specified, of which the first empirical study provides a good example. However, the weak base learner performs better if these model assumptions are not satisfied, of which the second empirical study provides a good example. There is a trade-off between prediction accuracy and robustness. 3) The "rank and average" strategy is more robust to the misspecification of the base learner compared to other averaging strategies.

The rest of the chapter proceeds as follows. Section 3.2 describes the ACF and CBD models with the additional time shift parameter. These two models are used as the base learners in our empirical studies, respectively. Section 3.3 introduces various averaging strategies. Section 3.4 presents empirical studies with the Human Mortality Database. Section 3.5 provides some discussion about the interpretation of the time shift parameter and the effect of a cohort effect on the base learners. Finally, Section 3.6 provides concluding remarks and some discussions on possible avenues for future research.

3.2 The Base Learner

In our proposed BMBE framework, the base learner is a bivariate-population model used to link the target population with one auxiliary population from a given pool. In principle, any bivariate-population model can be used as the base learner in our proposed BMBE framework.In this chapter, we use two prevailing bivariate-population models, the ACF and CBD models, as the testbed because of their unique roles in the literature. As mentioned in the preceding section, we add a time shift parameter Δt to the bivariate-population models to investigate how it can affect the resulting forecasting accuracy of our BMBE approach. We label the resulting bivariate-population models by ACF-ts and CBD-ts, respectively, with the suffix to indicate the inclusion of the "time shift" parameter in the models.

3.2.1 ACF-ts Model

Let $\log m_j(x,t)$ denote the central logarithmic mortality rate at age x in year t of the jth population, for $t = 0, 1, \ldots, T$, $x = x_1, \ldots, x_n$, and $j = 1, 2$. The ACF-ts model is

an extension of the bivariate-population ACF model of Li and Lee [2005] with an extra parameter Δt representing the time shift:

$$\log m_1(x,t) = a_1(x) + B(x)K(t) + b_1(x)k_1(t) + \epsilon_1(x,t), \tag{3.1}$$

$$\log m_2(x,t) = a_2(x) + B(x)K(t-\Delta t) + b_2(x)k_2(t) + \epsilon_2(x,t). \tag{3.2}$$

The first population (with subscript 1) is the target population for mortality prediction. The second population (with subscript 2) is the population included to borrow information from, and is hereafter referred to as the "auxiliary population". The time shift parameter Δt, which takes integer values from an interval symmetric about 0, characterizes the advance or delay of "evolution" in years by the target population over the auxiliary. If $\Delta t = 0$, the model degenerates to the bivariate-population ACF model of Li and Lee [2005]. The other components in the model carry the same meaning as those in the ACF model, i.e., the common trend term is a product of $B(x)$, a deterministic age function, and $K(t)$, a stochastic period function, $b_j(x)$ and $k_j(t)$ constitute the population-specific components, and $\epsilon_j(x,t)$ are the white noise terms, $j = 1, 2$. The time-lag structure of the common process in the aforementioned base learner is a special case of a more general Granger-causality method in Milidonis and Efthymiou [2017]. We actually assumes the common process term in population 1 at time t is perfectly correlated with that in population 2 at time $t - \Delta t$.

To avoid the issue of model unidentifiability, we impose the following constraints in parallel to the ACF model:

$$\sum_{l=1}^{n} B(x_l) = 1, \quad \sum_{t\in\mathbb{S}} K(t) = 0, \quad \sum_{l=1}^{n} b_j(x_l) = 1, \quad \text{and} \quad \sum_{t=0}^{T} k_j(t) = 0,$$

where $\mathbb{S} = \{-\Delta t, -\Delta t + 1, \ldots, T\}$ for $\Delta t \geq 0$ and $\mathbb{S} = \{0, 1, \ldots, T - \Delta t\}$ for $\Delta t < 0$. In our study, we fit the common trend sequence $K(t)$ by a random walk process with drift (RWD), a widely used model in the literature. In the meanwhile, we fit the population-specific components $k_j(t)$ with the "best" ARIMA model chosen by the Akaike Information Criterion (AIC). For the calibration of the ACF-ts model, we prespecify a set of integer values for Δt, implement the usual singular value decomposition (SVD) procedure for the calibration of the model with Δt fixed at each integer from the prespecified set, and then apply a validation procedure to choose the best Δt value (see Section 3.2.4). We relegate the step-by-step calibration procedure to Appendix B.1.

3.2.2 CBD-ts Model

Let $q_j(x,t)$ denote the probability that an individual in population j aged x will die between t and $t+1$ given the individual is alive at time t, for $t = 0, 1, \ldots, T$, $x = x_1, \ldots, x_n$, and $j = 1, 2$. The CBD-ts model describes the mortality development of two populations as follows:

$$\text{logit}(q_1(x,t)) = K(t) + (x - \bar{x})k_1(t) + \epsilon_1(x,t), \tag{3.3}$$

$$\text{logit}(q_2(x,t)) = K(t - \Delta t) + (x - \bar{x})k_2(t) + \epsilon_2(x,t), \tag{3.4}$$

where $\text{logit}(q) \equiv \log\left(\frac{q}{1-q}\right)$, $\bar{x} = \sum_{l=1}^{n} x_l/n$. The time shift parameter Δt takes integer values from a symmetric interval centered at 0. This specification extends the original single-population CBD model of Cairns et al. [2006] or equivalently model M5 in Cairns et al. [2009] to a bivariate-population model by introducing a common age-period term $K(t)$ shared by both populations and a time shift parameter to reflect the differences in their mortality development stages. We fit the common trend sequence $K(t)$ with a RWD model and the population-specific components $k_j(t)$ with a bivariate RWD model. The calibration of Δt is also carried out through a validation procedure as in the ACF-ts model (see Section 3.2.4 for details), and that of other components uses the maximum likelihood method (see Appendix B.2 for details).

3.2.3 A Generalized Model Setup

The aforementioned models can be further summarized as a general framework that encompasses many classic bivariate-population mortality models. If $\eta_j(x,t)$ is denoted to represent an age-specific quantity characterizing the mortality level at age x in year t of the jth population, for $t = 0, 1, \ldots, T$, $x = x_1, \ldots, x_n$, and $j = 1, 2$. A generalized bivariate-population model that allows time shift can be specified as follows:

$$\eta_1(x,t) = F(x,t) + f_1(x,t), +\epsilon_1(x,t), \tag{3.5}$$

$$\eta_2(x,t) = F(x, t - \Delta t) + f_2(x,t) + \epsilon_2(x,t). \tag{3.6}$$

$F(x,t)$ is a function characterizing an underlying common mortality trend shared by both populations, Δt takes integer values from a symmetric interval with respect to 0, $f_1(x,t)$ and $f_2(x,t)$ reflect the specific features of the two populations respectively, and $\epsilon_1(x,t)$ and $\epsilon_2(x,t)$ denote white noise terms in the above model.

We have demonstrated how extensions can be spawned from the ACF model and the CBD model under this framework respectively. Extension of some other mortality models is

also available through different specifications of $\eta_j(x,t)$, $F(x,t)$ and $f_j(x,t)$. For example, the cohort effect could be considered in $f_j(x,t)$, to extend those mortality models with a cohort effect. An Example will be provided in Section 3.5.2.

3.2.4 Choice of Δt

As mentioned earlier, the parameter Δt is designed to capture the difference in years of mortality development between the target population and the auxiliary population in the bivariate ACF-ts or CBD-ts model. The way to determine the value of Δt is in the same spirit as choosing values of hyperparameters in statistical learning, which can be realized by applying a validation procedure. As a preparation, the training dataset is partitioned into a modeling set and a validation set and a grid of integer values is pre-specified for Δt. The modeling set is used to calibrate a model with a fixed Δt and this modeling procedure yields a series of fitted models corresponding to each integer in the pre-specified set. The value of Δt is determined as the one that yields the smallest overall sum of squared errors (SSE) calculated based on the validation set for predicting $\log m_j(x,t)$ in the ACF-ts model or $\text{logit}(q_j(x,t))$ in the CBD-ts model.

In the analysis of HMD in Section 3.4, we take a pre-specified set of integers over $[-10, 10]$ for Δt. Given two specific populations, sometimes it is clear the mortality development stage of the target population is in advance of the auxiliary population or the other way around, so a set of positive or negative integers is appropriate. However, we universally adopt the candidate set $\{-10, -9, \ldots, 9, 10\}$ for Δt when jointly modeling any pair of target and auxiliary populations to make the choice of Δt fully driven by data and avoid selection biases due to subjective judgment on the development stages of populations.

3.3 Bivariate Model Based Ensemble for Prediction

The bivariate-population models introduced in Section 3.2 allow us to borrow information from one auxiliary population for the forecasting of the target. To borrow information from multiple populations, it is natural to think of a multi-population model, which can be, however, highly intractable and computationally prohibitive when it comes to calibration. Further, as pointed out in the introduction section, identifying a set of auxiliary populations conducive to enhancing forecast accuracy for the target population is a complex undertaking. So, instead of calibrating a multi-population model, we propose an ensemble framework using the bivariate-population models as the base learners.

The primary idea of the ensemble procedure is to reduce the prediction uncertainty and improve prediction accuracy by aggregating prediction results from multiple predictive models. For a given target population and a pool of auxiliary populations, we fit a bivariate-population model between the target population and each auxiliary population in the pool and obtain the extrapolative results (i.e., forecast) on future mortality rates of the target population from each of the resulting bivariate-population models. We then aggregate all the forecasts to form a final forecast for the mortality of the target population. Let $\eta_1^{(s)}(x,t)$ denote the mortality forecast of the target population for age x and calendar year t using information from the sth auxiliary population, $s = 1, \ldots, S$. Here, $\eta_1(x,t) = \log m_1(x,t)$ if the ACF-ts model was adopted as the base learner, and $\eta_1(x,t) = \log\left[\frac{q_1(x,t)}{1-q_1(x,t)}\right]$ if the CBD-ts model was used. Then the final ensemble prediction takes a general form of

$$\hat{\eta}_1(x,t) = \sum_{s=1}^{S} w_s \eta_1^{(s)}(x,t) \ , \tag{3.7}$$

where w_s is the weight assigned to the sth auxiliary population. We consider the following averaging strategies (i.e., the strategies of assigning the weights w_s):

- **Simple Average (SimAvg)**: This is the most naive strategy where we assign equal weights to all S auxiliary populations, i.e., set $w_s = \frac{1}{S}$, for $s = 1, \ldots, S$, in Equation (3.7). This strategy does not involve any screening over the cascade of bivariate-population models regarding their efficacy in predicting the mortality of the target population.

- **Average Based on Geographic Information (GeoAvg)**: We use geographic proximity as exogenous information to pre-select groups. For a given target population, the final forecast is the average of forecasts from the bivariate-population models with auxiliary populations within the same geographic group as the target population. Specifically, if there are R auxiliary populations within the same geographic group as the target population, we set $w_s = \frac{1}{R}$ in Equation (3.7) for the R auxiliary populations, and $w_s = 0$ for those auxiliary populations outside the geographic group.

- **Average Based on Clustering Results (KmeansAvg)**: This is a data-driven strategy to pre-specify groups using cluster analysis (k-means method) to find populations with similar mortality characteristics [Hatzopoulos and Haberman, 2013]. For a given target population, we compute the final forecast as the simple average of the forecasts from the bivariate-population models built with each auxiliary population located within the same cluster as the target.

- **Rank and Average (RankAvg)**: This strategy is inspired by the notion that "relevant" populations tend to bring in more information while "irrelevant" populations only bring in noises. We rank auxiliary populations according to the validation SSEs of the corresponding bivariate-population models on the validation data set mentioned in Section 3.2.4. This strategy tends to only keep the top u^* bivariate-population models for the final forecast. For any $u \in \{1, 2, \ldots, S\}$, the final forecast based on the top u bivariate-population models is calculated by assigning $w_s = \frac{1}{u}$ for each of the top u bivariate-population models, and $w_s = 0$ for the rest bivariate-population models in Equation (3.7). The value of u^* is selected as the one such that the final forecast yields the smallest validation SSE.

  The RankAvg strategy addresses the challenge of auxiliary populations selection and facilitates a customized selection procedure for each target population. It is expected to have a relatively more robust performance compared with the other strategies since it takes in the whole pool of auxiliary populations and has a mechanism of screening out those that are impotent in improving the prediction accuracy for the mortality of the target population.

3.4 Empirical Analysis

In this section, we evaluate the performance of various averaging strategies and base learners under the proposed BMBE framework by applying them to the Human Mortality Database (HMD) and comparing them with several benchmark models. We conduct two empirical studies. The first study focuses on the mortality forecast for a full range of ages between 0 and 100 and adopts the bivariate ACF model (with and without time shift) as the base learner. The second study focuses on mortality forecasting for the senior age group, ages between 55 and 90, and applies the bivariate CBD model (with and without time shift) as the base learner since the CBD model is known for its superior performance in characterizing mortality development for seniors. In addition, we also study the possible influence to add a cohort effect to the CBD base learner in our BMBE framework.

As one would see shortly, the first empirical study confirms a noticeable improvement in mortality forecasting accuracy by the BMBE method over benchmark models and a positive contribution in reducing forecast error by adding the time shift component to the base learner. The second empirical study still shows an improvement by the BMBE method though not as significant as in the first empirical study, and the inclusion of the time shift parameter only has a marginal effect on the performance of the resulting mortality

forecast. Furthermore, these empirical studies together reveal the importance of avoiding severe misspecification in the base learner for the proposed ensemble method to work well.

3.4.1 Data Description

To strike a balance among different factors, such as the size of the candidate pool, the length of time, and the number of missing values in HMD, we concentrate on the mortality data of 30 populations from 1970 to 2010 in our empirical studies. The mortality data (1970–2010) is split into a training set (1970–2002) and a test set (2003–2010). The training set is further split into a modeling set (1970–1994) and a validation set (1995–2002). We use the modeling set to fit bivariate-population models and create candidate predictive rules and the validation set to determine the value of the parameter Δt in each bivariate-population model with time shift. The validation set is also used to determine the optimal number of top-performed forecasts used in the RankAvg strategy.

According to Brainerd and Cutler [2005], demographic disasters in the form of sharply rising death rates happened among several member countries of the former Soviet Union in the 1990s and beyond. These populations may demonstrate an obviously different trajectory of mortality development from the rest in the HMD, thus one population from the group cannot be seen as an advance or a delay of a population outside the group, which imposes a major challenge in adopting bivariate-population models with a time shift. Before we hastily exclude these populations from our studies, we conduct further analysis to ensure the appropriateness of the criterion of population choice used in our studies. These empirical studies together also reveal the importance.

We illustrate the age-aggregated logarithmic mortality sequence for each of the 30 populations on the left panel of Figure 3.1. We then conduct a k-means cluster analysis to detect structural dissimilarity among these sequences of age-aggregated logarithmic mortality. The k-means cluster procedure is designed to exclude the effect of mortality level on the dissimilarity among different populations. Below are the details of the clustering procedure:

1. Define a dissimilarity matrix $\mathbf{D}$ with entries $D_{i,j}$ being the variance of the sequence $\text{diff}_{i,j}(t) = \sum_x \log_i(x,t) - \sum_x \log_j(x,t)$, $t = 1970, 1971, \ldots, 2002$.

2. Apply Multi-Dimensional Scaling (MDS) to transform the matrix $\mathbf{D}$ into a 2-dimensional objective. According to Cox and Cox [2008], MDS outputs a configuration of points in a 2-dimensional space, where each point represents one original sequence and the distance between each point pair retains the dissimilarity measured by $D_{i,j}$ to the greatest extent.

3. Implement a k-means cluster analysis to the transformed outputs of MDS.

It is worth noting that variance is used in the above procedure as the dissimilarity measure so that the disparity in mortality level does not have an effect on the dissimilarity measure. If two populations i and j have parallel trajectories in their age-aggregated logarithmic mortality, the difference sequence is a constant and thus, $D_{i,j} = 0$; otherwise, we anticipate a positive value for the variance.

We display the clustering results on the right panel of Figure 3.1. From the figure, most member countries (Belarus, Bulgaria, Latvia, Lithuania, Russia, and Ukraine) of the former Soviet Union are classified into a separate group. We view these populations as "outliers" in mortality development patterns. We exclude these six populations from our analysis and focus on the rest of the 24 populations in our empirical studies; see Table 3.1 for the specific names of the 24 populations and their geographic grouping; see Diao et al. [2021] and Richman and Wüthrich [2021].

Table 3.1: Geographic grouping of 24 populations from HMD.

Target Population	**Geographic Group**	**Target Population**	**Geographic Group**
Australia	Oceania	Netherlands	West Europe
Austria	West Europe	New Zealand	Oceania
Japan	Asia	Norway	Scandinavia
Belgium	West Europe	Poland	East Europe
Scotland	Great Britain	Portugal	South Europe
Canada	North America	U.S.A.	North America
Czech Republic	East Europe	Slovakia	East Europe
Denmark	Scandinavia	Spain	South Europe
Finland	Scandinavia	Sweden	Scandinavia
France	West Europe	Switzerland	West Europe
Hungary	East Europe	Taiwan	Asia
Italy	South Europe	England & Wales	Great Britain

3.4.2 Empirical Study with the ACF-ts Model

In this empirical study, we investigate the performance of the BMBE method using the ACF model (with and without time shift) as the base learner, study various averaging strategies

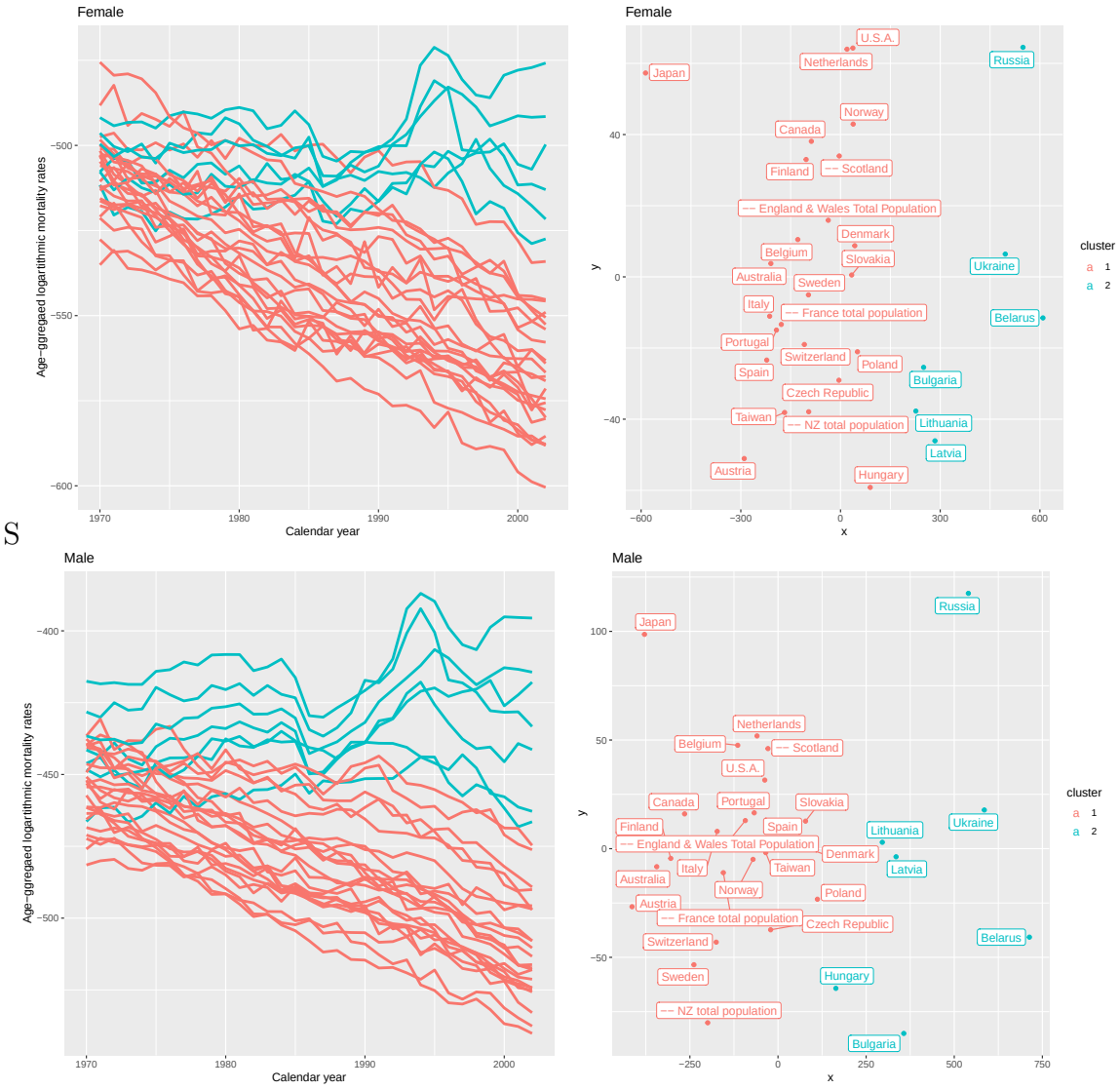


Figure 3.1: Clustering results of age-aggregated logarithmic mortality rates based on MDS (from top to bottom: female and male populations)

and the value of adding the time shift term and compare their performance with the classic ACF models in predicting the mortality rates of ages 0–100.

3.4.2.1 Prediction Models

We take each of the 24 populations in Table 3.1 as the target for mortality forecasting and obtain the forecasts through the BMBE method as described in Section 3.3. Specifically, we develop a bivariate-population ACF-ts model between the target population and each of the remaining 23 populations, resulting in 23 prediction rules for the mortality of the target population. We then apply an averaging strategy to ensemble all of the 23 prediction results into a final forecast. We apply all the four averaging strategies introduced in Section 3.3, and label the resulting forecasts by ACF-ts.SimAvg, ACF-ts.GeoAvg, ACF-ts.KmeansAvg, and ACF-ts.RankAvg, respectively. For ACF-ts.KmeansAvg, we apply a k-means clustering analysis with $k = 8$ and 1,000 independent random initializations. We execute the analysis over the 24 populations for both genders respectively.

To examine the value of including the time shift term in the ACF model, we consider **ACF·RankAvg** model, which is the spacial case of **ACF-ts·RankAvg** with Δt fixed at zero. We also consider the following benchmark models for comparative analysis:

- **Lee-Carter**: The Lee-Carter model is fitted to each population separately. The sequence of $k(t)$ obtained from an SVD procedure is fitted using the `auto.arima` function from the `R` package `forecast` to search for a suitable ARIMA model.
- **ACF·AIO**: The ACF model is fitted to the 24 populations jointly.
- **ACF·GeoInfo**: The ACF model is fitted to each geographic groups in Table 3.1.
- **ACF·kmeans**: The ACF model is fitted to each cluster from the k-means clustering algorithm (with $k = 8$ and 1,000 independent random initializations).

We calibrate each of the above benchmark models using the training set (1970–2002) and then extrapolate the resulting models into the testing period (2003–2010) for mortality forecasting.

3.4.2.2 Prediction Performance

Evaluation Based on Test SSEs

The prediction accuracy is evaluated in terms of test SSEs, which are denoted by $e(t)$ and calculated as follows:

$$e(t) = \sum_{x=0}^{100} \left[\log m_1(x,t) - \log \widehat{m}_1(x,t)\right]^2, \quad t = 2003, \ldots, 2010,$$

where $m_1(x,t)$ is the realized mortality rate of a target population, and $\widehat{m}_1(x,t)$ is its forecast. For a succinct summary about the performance of each forecasting model, we compute the overall test SSE as $\sum_{t=2003}^{2010} e(t)$ for each target population. A smaller test SSE implies a better prediction performance.

Rotating the target population over the 24 in the pool, we obtain 24 overall test SSEs. We report the 1st quartile, median, mean, and the 3rd quartile of the 24 overall test SSEs for each forecasting model in Table 3.2. These results indicate that our proposed BMBE based forecasts (the bottom five in each panel of Table 3.2), particularly ACF-ts·RankAvg and ACF-ts·SimAvg, substantially outperform the benchmark models (the top four models in each panel of Table 3.2) with a smaller median and mean of the resulting test SSEs. The 1st and 3rd quartiles from each of the five BMBE based predictions are also generally smaller than those for the benchmark models.

The 24 populations show a similar downward trend in their trajectories of logarithmic mortality rate as shown by those red curves in Figure 3.1. Therefore, the trajectory of one population may be well approximated by shifting that of another population and the ACF model with a time shift can be considered a good fit for most pairs of populations. When the base learner is a correctly specified model, the SimAvg strategy is expected to perform well as it has the advantage of averaging over the full list of auxiliary populations and makes use of all useful information. The relative underperformance of the GeoAvg and KmeansAvg strategies is due to the fact that they both only borrow information from a small subset of the population pool. The RankAvg strategy gives comparable results to the SimAve strategy. A comparison between ACF-ts·RankAvg and ACF·RankAvg confirms the benefit of including a time shift component in the ACF base learner.

The population-specific comparison is also conducted between the ACF-ts·RankAvg model and the benchmark ACF·GeoInfo model based on the test SSE on 24 male populations and listed in Table B.1. Based on the results, the proposed method has led to an increase in predicting accuracy for 19 out of 24 male populations with an averaged 14.15% improvement.

Evaluation Based on One-sided Diebold-Mariano Tests

Table 3.2: Summary statistics of test SSEs comparing the prediction performance of the BMBE based approaches using ACF or ACF-ts as the base learner versus benchmark models.

	1st Quartile	Median	Mean	3rd Quartile
Female Population				
Lee-Carter	19.24	45.11	44.04	63.86
ACF·AIO	20.49	44.91	43.76	62.09
ACF·GeoInfo	20.95	45.49	42.96	63.77
ACF·kmeans	18.99	45.77	41.63	59.19
ACF·RankAvg	18.54	43.04	38.81	55.81
ACF-ts·SimAvg	16.59	39.65	37.78	54.00
ACF-ts·GeoAvg	19.43	38.21	38.95	56.67
ACF-ts·KmeansAvg	15.11	43.79	39.51	59.47
ACF-ts·RankAvg	15.04	40.84	37.34	54.73
Male Population				
Lee-Carter	28.28	40.85	49.07	65.29
ACF·AIO	28.28	39.54	46.90	65.76
ACF·GeoInfo	23.75	37.43	41.41	64.02
ACF·kmeans	24.35	36.70	42.31	58.82
ACF·RankAvg	21.49	37.15	37.13	54.52
ACF-ts·SimAvg	23.00	33.67	36.66	56.69
ACF-ts·GeoAvg	26.74	35.76	37.97	57.63
ACF-ts·KmeansAvg	24.80	35.82	39.69	58.38
ACF-ts·RankAvg	24.40	31.61	35.58	56.11

Like it has been adopted in Section 2.3.3.2 in the previous chapter to compare the performance of mortality forecasting models, we conduct a one-sided Diebold-Mariano (DM) test to determine if the test SSE from one predictive model is statistically significantly smaller than that of another. We follow the same comparison procedure using the DM test as we did in Section 2.3.3.2. For readers' convenience, we restate the details of the procedure in the context of the present empirical study.

For a pair of models in comparison, Model A and Model B, we apply two one-sided DM tests, respectively, with the null hypothesis that Model A is no worse than Model B and that Model B is no worse than Model A. The hypothesis tests are conducted based on the resulting sequences of test SSEs, $\{e(t), t = 2003, \ldots, 2010\}$ from both models. If Model A is concluded to be significantly worse than Model B by the DM test with a p-value smaller than 0.05, we count it as a win of Model B over Model A in predicting the mortality of the target population. If Model B is confirmed to be significantly worse than Model A by the DM test, we count it as a win of Model A over Model B. There are scenarios in which both tests lead to p-values larger than 0.05, and none of the two models wins the other in statistical significance.

Table 3.3 reports the number of wins for one model (from the group of **BMBE** methods) over another (from the group of benchmark models or the ACF.RankAvg model). Each cell of the table contains two integers recording the comparative results of the model in the row versus the one in the column. The first integer is the number of wins by the model in the row over the one in the column. The second integer is the number of wins by the model in the column. For example, "(23, 1)" in the first column in the panel of Female Population means that the ACF-ts·SimAvg model wins 23 times over the Lee-Carter model and the Lee-Carter model wins the ACF-ts·SimAvg model only once among all of the 24 comparisons for female populations. As shown in Table 3.3, the BMBE based predictive models perform significantly better than benchmark models in terms of the number of wins. The superiority of ACF-ts·SimAvg and ACF-ts·RankAvg is evident when compared with the four benchmark models as these two models frequently win among all 24 comparisons for both female and male populations. Furthermore, the results for both genders also indicate that ACF-ts·RankAvg wins ACF·RankAvg more frequently, which reinforces the benefits of including the time shift component (i.e., the parameter Δt) in the ACF base learner.

3.4.3 Empirical Study with the CBD-ts Model

In this empirical study, we investigate the forecasting performance of the BMBE method for seniors with ages ranging from 55 to 90. The CBD-type models are known as an

Table 3.3: Number of wins for comparison between the BMBE approaches and the benchmark ACF models from DM tests: In each cell, the first integer indicates the number of wins by the model in the row over the model in the column out of 24 comparisons and the second integer is the number of wins by the model in the column over the one in the row.

	Lee-Carter	ACF·AIO	ACF·GeoInfo	ACF·kmeans	ACF·RankAvg
			Female Population		
ACF-ts·SimAvg	(23, 1)	(22, 1)	(17, 0)	(16, 3)	(9, 3)
ACF-ts·GeoAvg	(13, 0)	(15, 1)	(13, 3)	(12, 4)	(3, 9)
ACF-ts·KmeansAvg	(15, 1)	(15, 2)	(15, 3)	(11, 3)	(5, 8)
ACF-ts·RankAvg	(21, 2)	(19, 2)	(16, 1)	(14, 2)	(10, 4)
			Male Population		
ACF-ts·SimAvg	(24, 0)	(20, 0)	(16, 4)	(15, 4)	(8, 9)
ACF-ts·GeoAvg	(16, 3)	(13, 3)	(10, 9)	(11, 7)	(5, 12)
ACF-ts·KmeansAvg	(19, 0)	(17, 0)	(10, 5)	(10, 9)	(6, 12)
ACF-ts·RankAvg	(24, 0)	(21, 0)	(14, 3)	(17, 3)	(11, 5)

improvement of the Lee-Carter/APC-type models for modeling and predicting the mortality of a senior group because of their relatively simple log-linear structure of the mortality curve and parsimonious age effects, see Cairns et al. [2009] and Cairns et al. [2011a]. We use the CBD model and its extension with time shift as the base learner in our BMBE method and compare the resulting forecasting performance with the classic CBD model (Cairns et al., 2006), which is also known as model M5 in Cairns et al. [2009].

3.4.3.1 Prediction Models

We consider the same four averaging strategies as described in Section 3.3 and use them with CBD models with a time shift. We label the resulting predictive models by CBD-ts·SimAvg, CBD-ts·GeoAvg, CBD-ts·KmeansAvg, and CBD-ts·RankAvg, respectively. To study the value of adding a time shift component in the base learner, we also consider the CBD·RankAvg method, which is the special case of the CBD-ts·RankAvg method with Δt fixed at zero in the base learner. For comparison, we consider the classic CBD model (CBD) as a benchmark.

3.4.3.2 Prediction Performance

Evaluation Based on Test SSEs

As in the previous empirical study, we use the overall test SSE $\sum_{t=2003}^{2010} e(t)$ as a measure of prediction accuracy, where

$$e(t) = \sum_{x=55}^{90} \left[\text{logit}\, q_1(x,t) - \text{logit}\, \widehat{q}_1(x,t)\right]^2, \quad t = 2003, \ldots, 2010,$$

$q_1(x,t)$ is the probability that an age-x individual from the target population dies between year t and $t+1$ given the individual is alive at time t, and $\widehat{q}_1(x,t)$ is its prediction.

Table 3.4 summarizes the prediction performance of the various models in the second empirical study. The CBD-ts·RankAvg method is a clear winner over the classic CBD model for both female and male populations. The performance of the SimAvg strategy is disappointing and worse than the classic CBD model, which forms a contrast to its outstanding performance in the first empirical study. The success of the SimAvg strategy in the first empirical analysis is benefited by the fact the trajectories of age-aggregated logarithmic morality rate share similar negative slopes as shown in Figure 3.1 and a ACF-ts model works as a reasonably well-specified base learner for most pairs of populations. The parallelity in trajectories of the aggregate logarithmic mortality rates within the 24 populations was also

confirmed by the MDS-based clustering analysis described in Section 3.4.1. However, the current empirical study considers the trajectories of age-aggregated logit$q_j(x,t)$ over ages 50-90, and these trajectories are not parallel to each other to the same degree as we have in the first empirical study. This implies that the CBD-ts model could be a misspecified base learner for a population pair, and the biases of forecast arising in each misspecified base learner cumulate and lead to deteriorated prediction results in the SimAvg strategy. The misspecification issue is much less severe for the GeoAvg and KmeansAvg strategies as they only average over a small subset of populations.

The superiority of the RankAvg strategy is apparent. The test SSEs from this strategy (from both CBD-ts.RankAvg and CBD.RankAvg) are smaller than those from the CBD model for both male and female populations. This strategy performs well in both empirical studies and exhibits some robustness in its outperformance. It does not require any pre-grouping and uses a fully data-driven mechanism to identify a subset of auxiliary populations for the final forest. When an auxiliary population demonstrates a different development pattern from the target population, the prediction based on the corresponding CBD-ts model may lead to a large prediction error and the RankAvg strategy tends to rank this auxiliary population low and eventually screen it out in the cross-validation procedure.

Furthermore, a comparison between CBD·RankAvg and CBD-ts·RankAvg shows that the inclusion of the time shift parameter in the base learner improves the results slightly for male populations but not the female populations. The results of the two empirical studies suggest that there is a trade-off between prediction accuracy and robustness. When we include a time-shift term in the base learner, we implicitly assume that the trajectory of one population is roughly a shift of the other. If the assumption is satisfied, the base learner with the time-shift term is more efficient and the final forecast is more accurate and stable. Nevertheless, if the assumption is severely violated, we pay a heavy price with a biased base learner and get a deteriorated final forecast. The base learner without time-shift terms gives mediocre prediction results if the time-shift assumption is satisfied, while they do not have the pain point of being sensitive to misspecification.

Similarily, the population-specific comparison is conducted between the CBD-ts·RankAvg model and the benchmark CBD model based on the test SSE on 24 male populations and listed in Table B.2. 16 out of 24 male populations benefit from the proposed method with an averaged 3.76% improvement.

Evaluation Based on One-sided Diebold-Mariano (DM) Tests

We also apply the DM test to compare the performance of the various predictive models

Table 3.4: Summary statistics of test SSEs comparing the prediction performance of BMBE based approaches using CBD or CBD-ts as the base learner versus the CBD model.

	1st Quartile	Median	Mean	3rd Quartile
Female Population				
CBD	3.61	5.87	7.37	11.67
CBD·RankAvg	3.76	6.06	6.31	7.66
CBD-ts·SimAvg	5.13	7.96	10.28	13.32
CBD-ts·GeoAvg	4.60	6.50	8.76	11.95
CBD-ts·KmeansAvg	3.97	5.76	7.61	11.70
CBD-ts·RankAvg	3.18	5.76	6.94	10.32
M6	5.37	7.53	8.45	10.40
M6·RankAvg	8.14	33.96	59.19	103.29
Male Population				
CBD	2.32	2.88	3.08	3.63
CBD·RankAvg	2.08	2.93	2.94	3.44
CBD-ts·SimAvg	2.08	3.04	5.95	4.38
CBD-ts·GeoAvg	2.12	2.60	2.84	3.69
CBD-ts·KmeansAvg	2.22	2.54	2.84	3.56
CBD-ts·RankAvg	2.04	2.67	2.85	3.38
M6	2.13	3.14	3.20	3.73
M6·RankAvg	3.70	7.49	16.64	14.97

and report the results in Table 3.5. The table suggests that all the BMBE methods still beat the CBD model in terms of the number of wins, except when CBD-ts·GeoAvg is applied to female populations. The DM test also confirms the superiority of CBD-ts·RankAvg among all the predictive models.

Table 3.5: Number of wins for comparisons between a CBD-ts model and the CBD model based on a pairs of one-sided DM tests: In each cell, the first integer indicates the number of wins of the model in the row over the model in the column out of 24 comparisons and the second integer is the number of wins of the model in the column over the one in the row.

	CBD	
	Female	Male
CBD·RankAvg	(12, 9)	(9, 8)
CBD-ts·SimAvg	(10, 7)	(12, 9)
CBD-ts·GeoAvg	(7, 8)	(12, 6)
CBD-ts·KmeansAvg	(11, 5)	(6, 5)
CBD-ts·RankAvg	(11, 3)	(10, 5)

3.5 Further Discussions

3.5.1 Interpretation of Δt values

3.5.1.1 Interpretation of Δt Values in ACF-ts Models

For each target population, 23 Δt values are resulted from the 23 bivariate ACF-ts models; each is built between the target and one reference from the rest 23 populations. We report the mean of the 23 Δt values for each target population in Table 3.6. Since Δt is designed to characterize the advance or delay of "evolution" in years by the target population over the auxiliary, the sign and magnitude of the parameter are expected to reflect the target population's relative position in the mortality development compared to the reference population. To be more specific, we expect a target population at a more developed level to

have a positive Δt, while a target population at a less developed mortality level to have a negative sign for Δt. The results in Table 3.6 are expected in some sense. We can see that populations, such as Canada, France, Japan, England and U.S.A., which are widely recognized to hold a leading position in the evolution of human mortality, have a positive mean value of Δt while those relatively less developed populations, such as Czech Republic and Hungary, have a negative mean value of Δt.

Table 3.6: Mean of 23 Δt values for each target population based on the ACF-ts model.

Target	Female	Male	Target	Female	Male
Australia	0.13	0	New Zealand	-1.70	-0.87
Austria	-1.22	-1.91	Norway	-2.00	-0.83
Belgium	0.26	0.13	Poland	0.17	0.70
Canada	3.09	0.83	Portugal	-0.09	-0.39
Czech Republic	-1.35	-1.96	Slovakia	0.57	-0.61
Denmark	-0.83	0	Spain	0.48	-0.70
Finland	-0.39	-0.78	Sweden	1.65	0.39
France	0.74	0.74	Switzerland	-1.30	-1.61
Hungary	-1.52	-1.74	Taiwan	-1.52	-1.65
Italy	0.04	1.35	England & Wales	1.91	2.91
Japan	2.83	4.61	Scotland	-1.70	-1.96
Netherlands	0.17	1.43	U.S.A.	1.57	1.91

To demonstrate the relationship between Δt and mortality development stage, we use the mean age-aggregated logarithmic mortality over the training period as a proxy for the relative mortality level and draw a scatterplot of the mortality level versus the mean value of Δt from the 24 populations; see the left upper graph in Figures 3.3 and 3.4 for female and male populations, respectively. Since a positive and large value of Δt should indicate a leading position in mortality development, we expect that a smaller value of the mean age-aggregated logarithmic mortality is associated with a larger Δt value. It is interesting to note that the Pearson correlation coefficients between the two variables take the values of -0.397 and -0.478 in the female and male populations, respectively, indicating a modestly negative relationship between the mean values of Δt and the mortality levels for the 24 populations.

While we observed a negative correlation between the Δt value and the mortality level

over the 24 populations, we are mindful of the fact that Δt directly reflects the mortality level only when the population-specific effects are relatively small compared with the common mortality trend in the ACF-ts model (3.5)-(3.6). For a better interpretation on how the Δt value is related to the mortality level, we define the concept of Relative Scale to measure how much population-specific effects contribute to the mortality development of a pair of popualtions compared to the common trend in a base learner model.

For a bivariate-population model with population-i as the target and population-j as the reference, its Relative Scale, denoted $RS(i, j)$, is calculated as follows:

$$\mathrm{RS}(i,j) = \frac{1}{2}\frac{\mathrm{mean}[\Delta k_i(t)]}{\mathrm{mean}[\Delta K(t)]} + \frac{1}{2}\frac{\mathrm{mean}[\Delta k_j(t)]}{\mathrm{mean}[\Delta K(t)]}, \tag{3.8}$$

where the mean operator is applied for the corresponding sequence over the time period involved in the model, and

$$\Delta k_i(t) = k_i(t+1) - k_i(t), \;\; \Delta k_j(t) = k_j(t+1) - k_j(t), \;\; \Delta K(t) = K(t+1) - K(t).$$

From the definition, a small $RS(i, j)$ value implies that the common mortality trend is the major driving force of mortality development for both populations, and a large $RS(i, j)$ value means that at least one of the two involved populations in the base learner model has its mortality development dictated largely by its population-specific effects. Figure 3.2 demonstrates the $RS(i, j)$ value of each pair from the 24 populations under our analysis and the figure reveals that the RS value is rather large for some pairs of populations, creating the risk of using the Δt value to represent the mortality level of a population. As a relevant note, the matrices in Figure 3.2 are symmetrical since we have $RS(i, j) = RS(j, i)$ by the definition of the RS measure.

To dig deeper into the intricate relationship between the Δt value and the mortality level of invovled populations, we start from either a lower or an upper triangle in the matrices of Figure 3.2, and set a threshold $\gamma \in \{0.25, 0.5, 1, \infty\}$ for RS to delete populations from the set until all of the remaining $RS(i, j)$ entries in the matrices are smaller than the threshold. The deletion proceeds in an interative way, and the population with the the largest $RS(i, j)$ value in the remaining entries of the matrices is deleted in each iteration until the criterion is met (i.e., all of the remaining entries are smaller than the threshold γ). It is worth noting that $\gamma = \infty$ means no deletion of any population from the matrices, and the remaining entries with a smaller γ from the deletion procedure should correspond to a stronger negative correlation between the Δt value and the mortality level for the remaining populations. This is confirmed by the scatterplots of the mortality level versus the mean value of Δt in Figures 3.3 and 3.4, where each graph only contains the scatterplots from

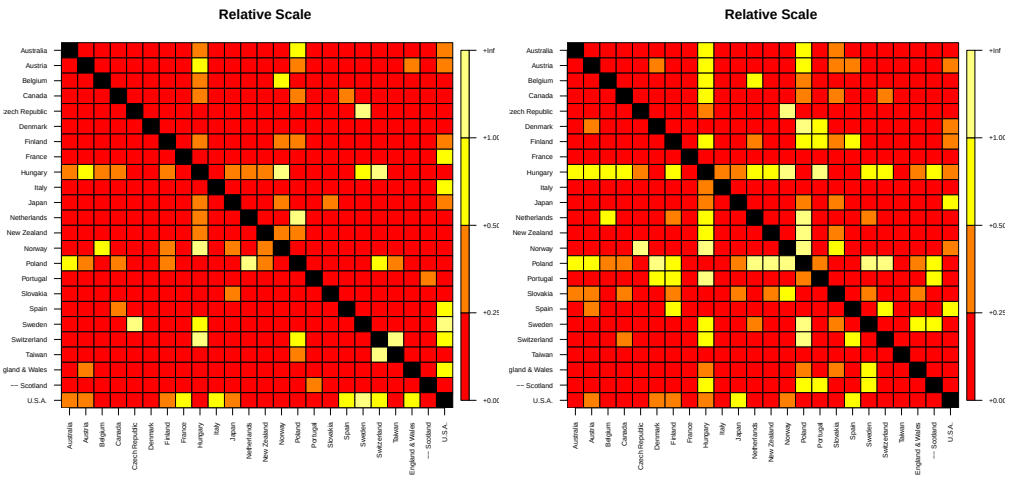


Figure 3.2: Values of Relative Scale based on ACF-ts models. Left panel: female data. Right panel: male data.

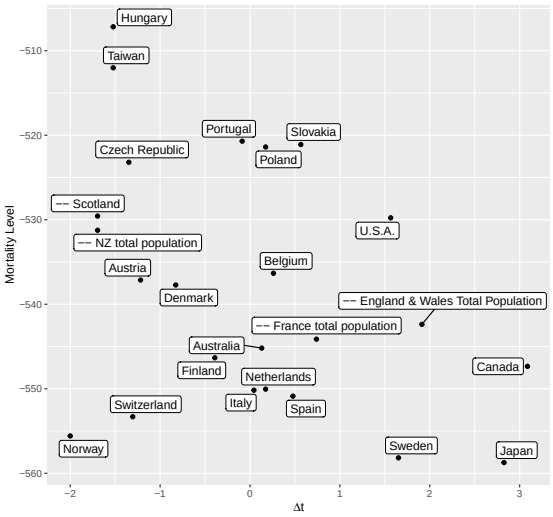
the remaining populations with RS value smaller than the corresponding threshold value γ. With a decrease in prespecified γ, the negative relationship between the mean value of Δt and the mortality level appears stronger. The Pearson correlation coefficient between the two variables decreases from -0.397 to -0.671 for the female populations and from -0.478 to -0.793 for the male populations respectively, with γ decreasing from ∞ to 0.25.

3.5.1.2 Interpretation of Δt Values in the CBD-ts Models

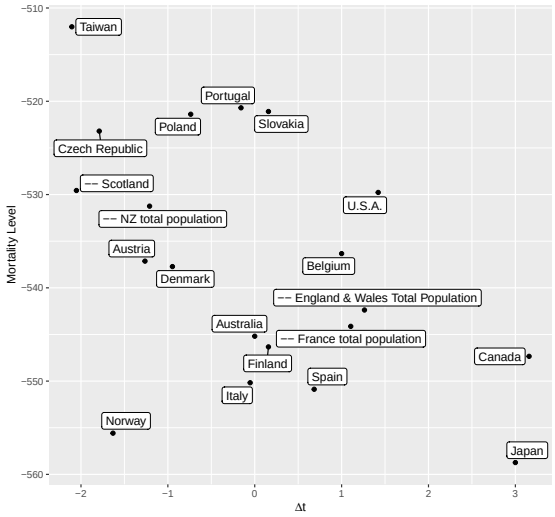
Similar to Section 3.5.1.1, the mean of the 23 Δt values, resulted from the 23 bivariate CBD-ts models with different auxiliary populations, for each target population is calculated and reported in Table 3.7.

To analyze the relationship between the calibrated Δt values in the bivariate CBD-ts models and the mortality development level, the RS measure, aimed to capture the degree of dominance by the common trend process over the population-specific effects, has been applied in the study for the CBD-ts models. As demonstrated in Figure 3.5, the resulting RS values for all of the involved pairs of populations are quite small. Thus, the common trend process dominates for the development of mortality in all the involved bivariate-population CBD-ts models, and a high correlation between the Δt value and the mortality level of a population is expected.

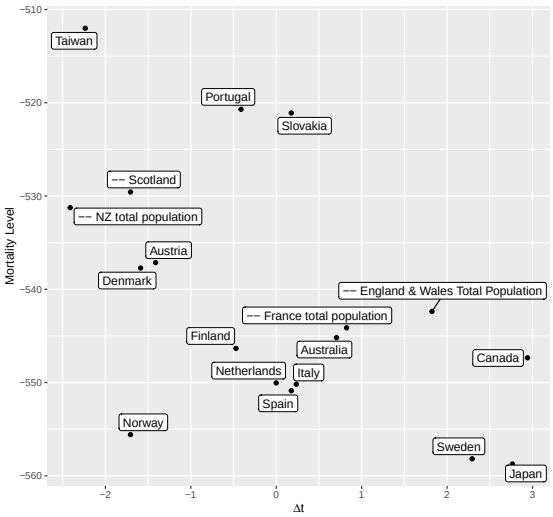
The scatter plots of the mean of age-aggregated logarithmic mortality rates versus the mean of Δt values for all the 24 target populations are provided in Figure 3.6, exhibiting an obvious negative relationship between the mean value of Δt and the mortality level.



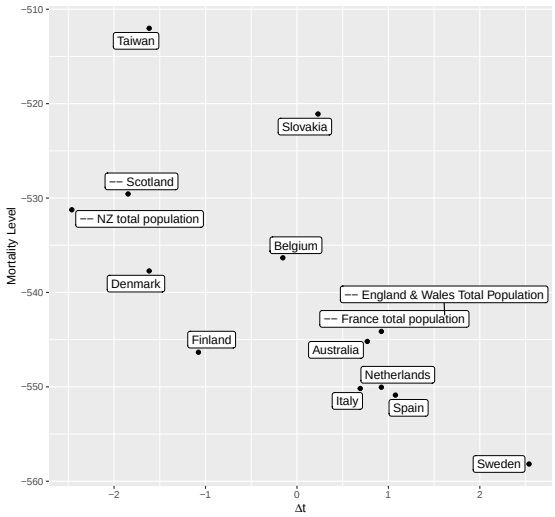
(a) $\gamma = \infty$, Corr $= -0.397$



(b) $\gamma = 1$, Corr $= -0.506$



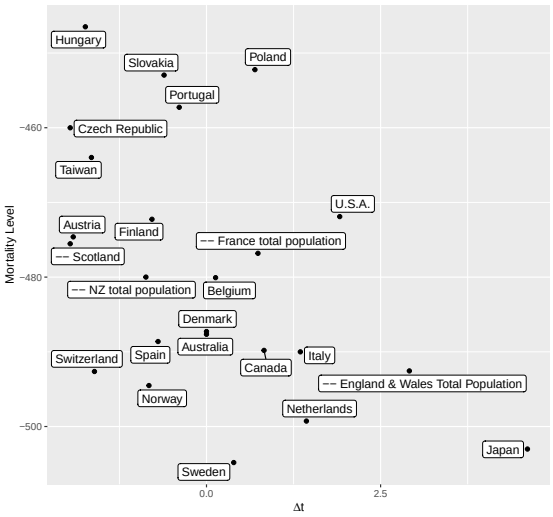
(c) $\gamma = 0.5$, Corr $= -0.560$



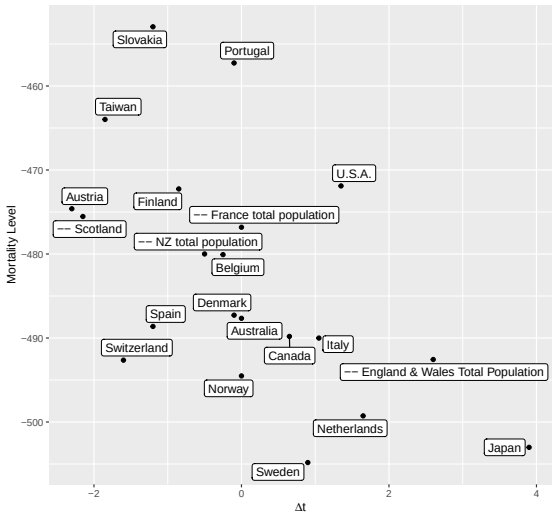
(d) $\gamma = 0.25$, Corr $= -0.671$

Figure 3.3: Female Data: relationship between the mean value of Δt and the relative mortality level based on ACF-ts models with different threshold values of γ.

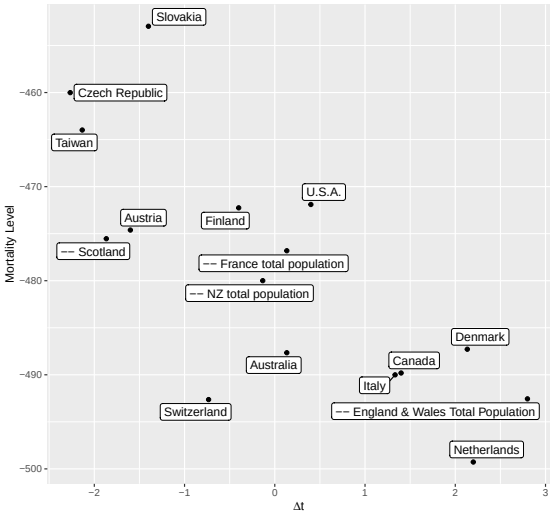
Populations with lower mortality tend to have a positive mean value of Δt, and those with higher mortality tend to have a negative mean value of Δt. The Pearson correlation coefficients between the two variables for the female population and male populations are as high as -0.896 and -0.971 respectively, suggesting a very strong negative linear correlation. These observations confirm the intended interpretation of the parameter Δt in our bivariate CBD-ts models: a positive and large value implies the advance of the target population in



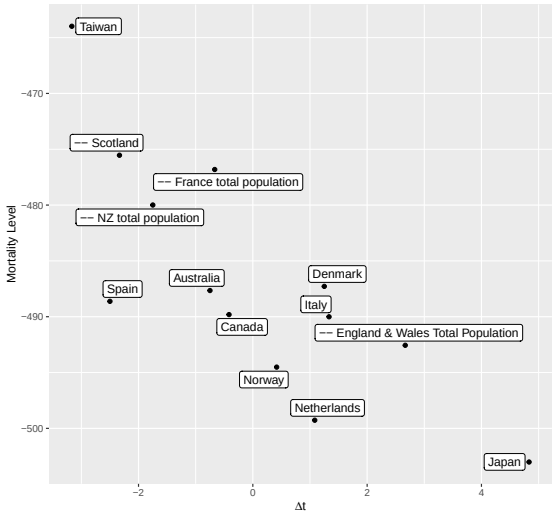
(a) $\gamma = \infty$, Corr $= -0.478$



(b) $\gamma = 1$, Corr $= -0.555$



(c) $\gamma = 0.5$, Corr $= -0.772$



(d) $\gamma = 0.25$, Corr $= -0.793$

Figure 3.4: Male Data: relationship between the mean value of Δt and the relative mortality level based on ACF-ts models with different threshold values of γ.

mortality development relative to the auxiliary population, and a negative value implies a delay of the target population in mortality development.

Moreover, since the common mortality trend is the major driving force of mortality development for both populations in every bivariate-population system involved in the current empirical study as validated by the observed small RS values, the value of the Δt also quantifies the number of years of difference in mortality development between the target

Table 3.7: Mean of 23 Δt values for each target population based on the CBD-ts model.

Target	Female	Male	Target	Female	Male
Australia	4.48	4.43	New Zealand	-0.13	2.22
Austria	1.17	-0.83	Norway	1.65	2.61
Belgium	1.65	-1.13	Poland	-7.17	-7.78
Canada	3.91	5.17	Portugal	-0.04	-1.87
Czech Republic	-7.30	-8.39	Slovakia	-6.26	-8.57
Denmark	-6.04	-2.74	Spain	8.26	5.43
Finland	1.48	-2.70	Sweden	0.91	6.43
France	7.65	4.13	Switzerland	7.04	6.57
Hungary	-7.91	-9.48	Taiwan	-4.96	-0.91
Italy	6.00	3.87	England & Wales	-2.39	-0.39
Japan	9.48	8.65	Scotland	-6.04	-5.65
Netherlands	-3.17	-0.30	U.S.A.	-2.26	1.22

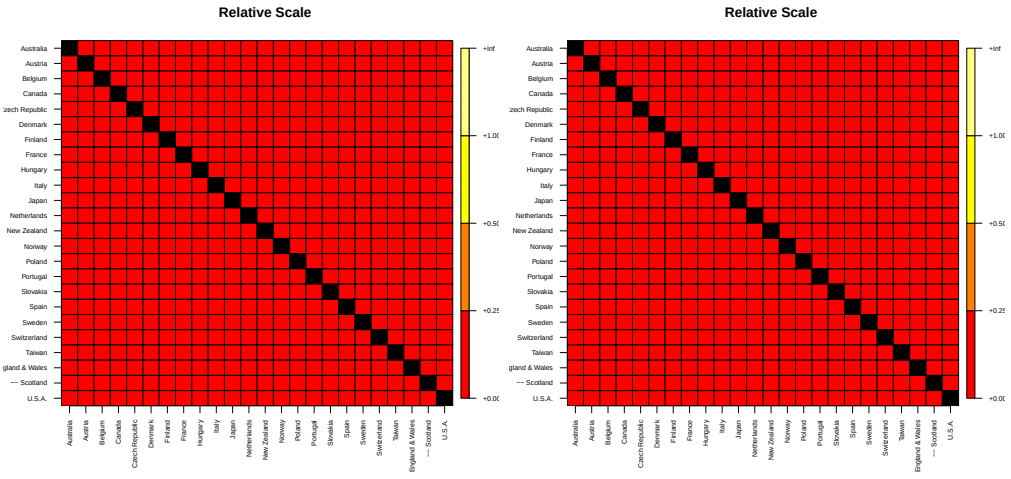
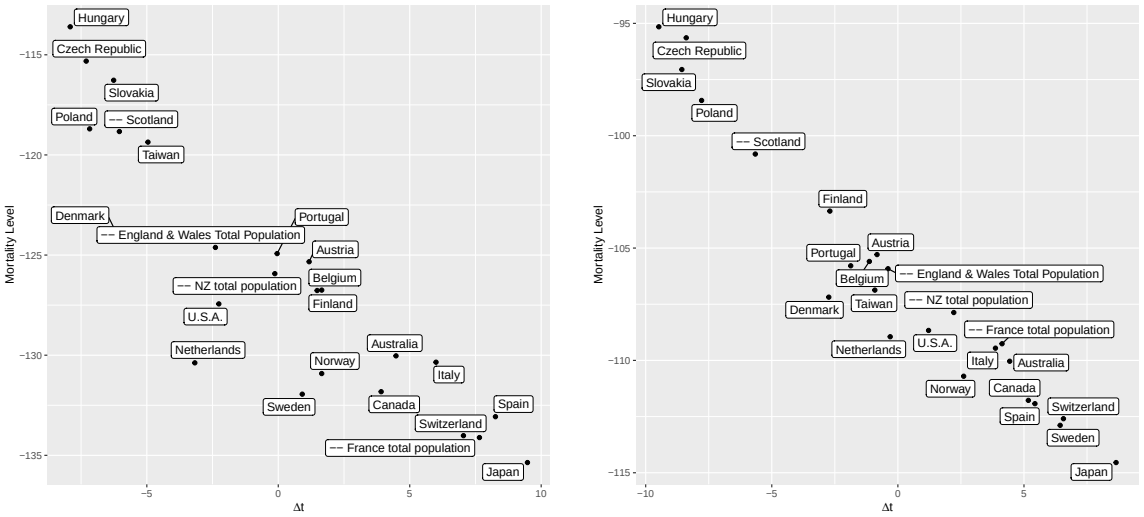


Figure 3.5: Values of Relative Scale based on CBD-ts models. Left panel: female data. Right panel: male data.



(a) Female: Corr = −0.896

(b) Male: Corr = −0.971

Figure 3.6: Relationship between the mean value of Δt and the relative mortality level based on CBD-ts models.

population and the auxiliary population. For instance, the calibrated Δt value between the US male population (as the target) and the Canadian male population (as the auxiliary) is -6, meaning that the Canadian male population is 6 years earlier than the target US male population in terms of the mortality development stage. We show the age-aggregated logarithmic mortality sequences of the two populations in Figure 3.7 for a graphical illustration of the relationship between the two populations.

3.5.2 Inclusion of Cohort Effect

Cohort effect, which addresses the mortality differences between people with different years of birth, has been widely documented in the mortality literature, see Renshaw and Haberman [2006], Cairns et al. [2009], Plat [2009] and Hunt and Blake [2021]. In this section, we study the possible influence of the cohort effect on mortality forecasting under our BMBE framework. We consider the following CBD-type model with a cohort effect as the base learner:

$$\log\left[\frac{q_j(x,t)}{1-q_j(x,t)}\right] = K(t) + (x-\bar{x})k_1(t) + \gamma_1(x,t) + \epsilon_1(x,t), \tag{3.9}$$

$$\log\left[\frac{q_j(x,t)}{1-q_j(x,t)}\right] = K(t-\Delta t) + (x-\bar{x})k_2(t) + \gamma_2(x,t) + \epsilon_2(x,t), \tag{3.10}$$

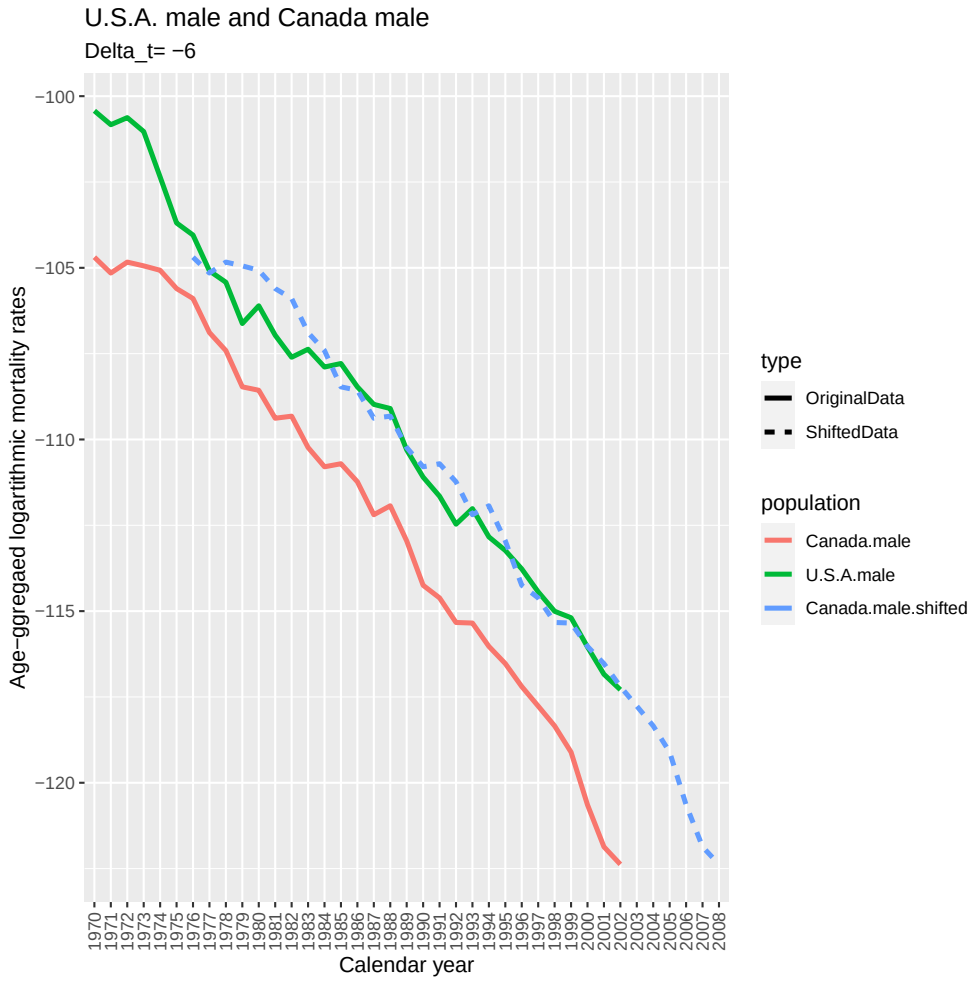


Figure 3.7: Age-aggregated logarithmic mortality sequences for the US male population and the Canadian male population for age from 55 to 90.

where $q_j(x,t)$, $K(t)$, $k_j(t)$, $\bar{x}$, $\epsilon_j(x,t)$ and Δt carry the same meaning as specified in (3.3) and (3.4) for CBD-ts model. This specification extends the original single-population M6 model in Cairns et al. [2009] to a bivariate-population model by introducing a common $K(t)$ and a time-shift term facilitating borrowing information from the auxiliary population across time. We label this model by "M6-ts". The difference of "M6-ts" from the CBD-ts model is the inclusion of the extra cohort effect terms $\gamma_j(x,t)$, $j = 1, 2$.

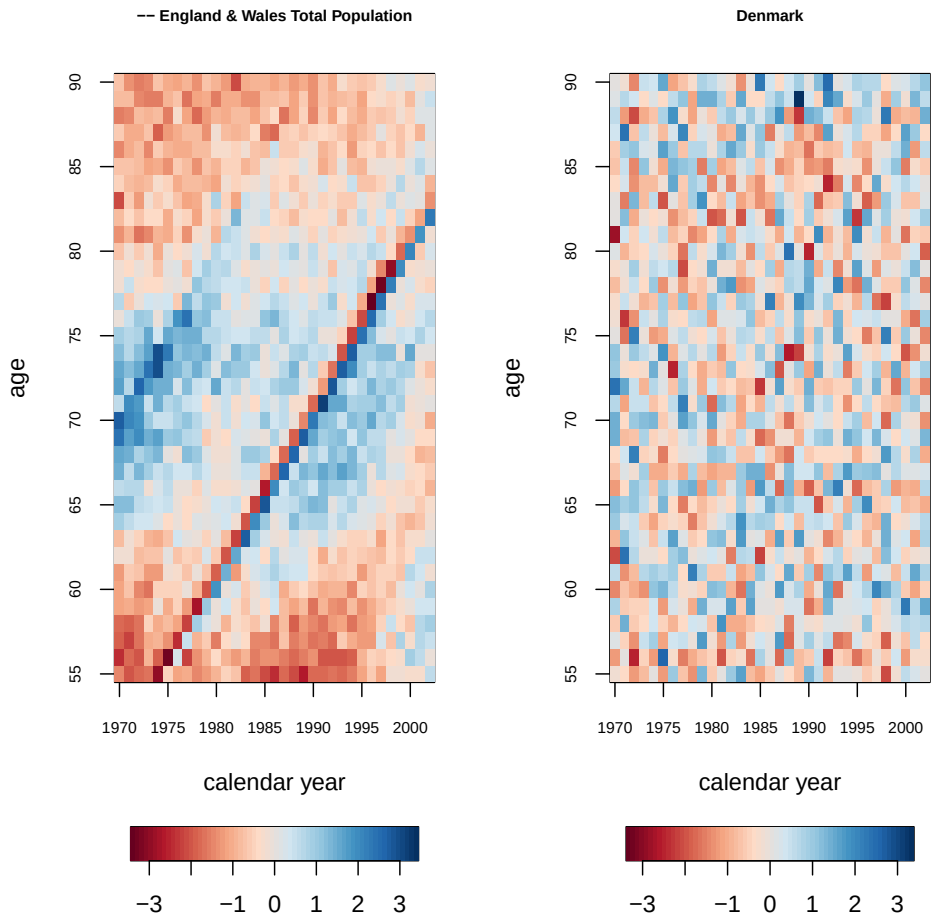


Figure 3.8: Deviance residual of the CBD model on different populations. Left: male of ages 55 to 90 for England & Wales; right: male of ages 55 to 90 for Demark.

Table 3.8 summarizes the test SSEs from the single population M6 model, and the BMBE method using the model in (3.9) and (3.10) as the base learner. The RankAvg averaging strategy is applied in the BMBE method, and therefore, the resulting predictive model is labeled by M6-ts·RankAvg. A comparison between Table 3.4 and 3.8 indicates that the test SSEs of M6-ts·RankAvg are significantly larger than those of the other models, especially

for female populations. The inclusion of both the cohort effect term and the time shift term in the base learner is detrimental to the overall forecasting accuracy in the study.

Table 3.8: Summary statistics of test SSEs for BMBE approach (using RankAvg averaging strategy and M6 as the base learner) versus the plain M6 model.

	1st Quartile	Median	Mean	3rd Quartile
Female Population				
M6	5.37	7.53	8.45	10.40
M6·RankAvg	8.14	33.96	59.19	103.29
Male Population				
M6	2.13	3.14	3.20	3.73
M6·RankAvg	3.70	7.49	16.64	14.97

There are considerable differences in the strength of the cohort effect across different populations, and strong cohort patterns have only been found in some populations, see Li et al. [2016]. Figure 3.8 contains two demonstrative examples regarding the disparate strength of the cohort effect. Strong diagonal signals in the residuals of the CBD model indicate the existence of a cohort effect for the male seniors from England & Wales in the left subfigure. In contrast, the residuals of the CBD model exhibit pure randomness so male seniors from Denmark in the right subfigure do not suggest the existence of a cohort effect. Therefore, the M6-ts model containing cohort effect terms for both populations is a misspecified model for this pair of populations. Adding cohort effects to both populations in the base learner may only benefit a target population having a strong cohort effect itself and can borrow information from the ones which also have strong cohort effects in the BMBE method. It jeopardizes the overall forecasting performance as it is a misspecified base learner as long as one in each population pair does not include a noticeable cohort effect. This suggests a complicated parametric model with many model assumptions is not suitable to serve as a base learner under the BMBE framework.

3.6 Concluding Remarks

We propose a general and flexible ensemble framework for mortality forecasting that allows for borrowing information across populations via a cascade of bivariate mortality models, thus called bivariate mortality based ensemble (BMBE). The framework ensembles forecasts from all the bivariate mortality models using various averaging strategies for the final forecast, among which the "Rank and Average" averaging strategy yields robust outperformance in forecast accuracy. The "Rank and Average" strategy also turns out to outperform benchmark models in our empirical studies. One salient advantage of the "Rank and Average" strategy lies in its no requirement of pre-grouping populations and its ability to recognize useful information and exclude irrelevant noise in a data-driven manner. We also study the effect of a time shift component and a cohort effect term on the base learner on the resulting mortality forecasting accuracy within the ensemble framework. Our empirical studies suggest that avoiding severe misspecification in the base learner is the key to success when utilizing the BMBE approach.

Chapter 4

Enhancing Mortality Prediction via Selection of Age Bands

4.1 Introduction

A general problem in front of actuaries is obtaining reliable estimates of future mortality rates for insurance product design, pricing, and risk management. As discussed in the previous chapters, borrowing information across populations has become quite prevailing and has been confirmed helpful if an appropriate procedure is applied. In this chapter, we move our focus to borrowing information across ages.

Almost all the existing mortality models, e.g., The Lee-Carter model [Lee and Carter, 1992] and the CBD model [Cairns et al., 2006], exploit joint modeling across ages that can be perceived as a venue for borrowing information across ages for mortality prediction. As we have illustrated in Figure 1.1 in Chapter 1, the mortality development patterns turn to have more similarities within some groups of ages (e.g., young ages, middle ages, senior ages) and more differences across different age groups. This motivates us to study the problem of selecting age bands to be used in mortality models for enhanced prediction performance. Indeed, some literature has confirmed the different predicting results with different age ranges used in a mortality predictive model. Shang and Haberman [2020] conducted comparisons between the full-age range model that uses data from all accessible ages and the partial-age range model that uses data from ages of interest only. Furthermore, Tsai and Cheng [2021] incorporated statistical clustering methods into mortality models to dig into similarities of age-specific development patterns among ages and improve predicting performance. Wang et al. [2021] considered the "neighboring" effect under the assumption

that future mortality development of a target age is influenced by the development patterns of its "neighboring" ages.

In this chapter, we aim to throw more light upon the idea of borrowing information across ages and design an innovative age-specific age band (ASAB) based framework for mortality prediction with enhanced accuracy by detecting similarities in mortality development patterns across ages and fully utilizing useful hidden information with the help of data mining tools. The framework is built on splitting the overall predicting goal into multiple individualized predicting tasks, seeking potential improvement in predicting accuracy for each target age to the largest extent, and then aggregating all the results to meet the overall predicting task. Different data mining tools are incorporated to determine the specific form of an optimal pool of ages or age sets to borrow information through an existing mortality model. We propose considering an age-specific age band for different target ages to borrow information from neighboring ages. The chosen age band is expected to be capable of embracing the potential benefits of the similarity shared by those ages in their mortality development patterns and maintaining a relatively simple structure.

The proposed ASAB based mortality prediction framework for selecting age bands is flexible to embed most existing mortality forecasting models, whereas we will exploit the Lee-Carter model as the embedding mortality model in this chapter for an illustrative purpose. We apply the proposed ASAB based prediction method to the mortality data of 24 populations of both genders from the Human Mortality Database (HMD), together with some benchmarks based on statistical learning approaches like clustering. The empirical study results confirm the effectiveness of our proposed ASAB based method in that our proposed method secures an overall improvement in predicting accuracy for most of the 48 populations. Meanwhile, the proposed ASAB based framework has demonstrated its potential to be extended to combine with borrowing information across populations by embedding multi-population models.

Based on the assumption that nearby target ages should borrow information from a similar set of ages to achieve desirable prediction accuracy enhancement, a smoothing procedure is incorporated in the ASAB based framework. The effects of different choices of the desired smoothness have been examined in our empirical study. Based on the empirical results, we recommend setting a moderate degree of smoothness as the desired level to ensure the effectiveness of the smoothing procedure. Furthermore, we discuss the capability to ensure a certain degree of age coherence of the proposed ASAB based method. The results confirmed that the ASAB based methods, both the smoothed and the non-smoothed version, have demonstrated their abilities to maintain a desirable level of age coherence.

The rest of the chapter proceeds as follows. Section 4.2 provides a detailed description of our proposed ASAB based prediction framework and the data-mining procedures we utilize to determine the age band selection. Section 4.3 presents empirical studies of our proposed methods and compares them with benchmark models considered in our study by analyzing data from the Human Mortality Database. Finally, Section 4.4 provides some concluding remarks.

4.2 Age-Specific Age Band (ASAB) Based Framework for Mortality Prediction

For the prediction of a mortality table, a conventional way is to calibrate a mortality model with the full age-range data target and obtain the predicting results. The similarities among different ages, which can further enhance the predicting accuracy, however, are not fully utilized with a full age-range model. To embrace the potential benefit of borrowing information across ages that share similarities in their mortality development pattern, some data-mining tools have been introduced to help detect similarities and formulate mortality prediction. In this section, we propose the age-specific age band (ASAB) based framework for mortality prediction, in which the overall predicting target is split into multiple individual target ages and focus on the mortality prediction of each target age with help of borrowing information across ages by searching for an optimal age band.

4.2.1 General Age-Specific Age Set Based Framework

Literature has confirmed the positive potential of utilizing similarities among different ages for acquiring enhanced mortality prediction. Shang and Haberman [2020] proved by empirical studies that partial-age range models, i.e., models calibrated with data of ages of interest only, can provide better mortality prediction accuracy for the retiree group. Tsai and Cheng [2021] pointed to certain similarities shared by some specific set of ages in their mortality development patterns and obtained improved mortality prediction performance by incorporating these similarities through clustering analysis.

In contrast to the above literature, we propose a general age-specific age set based mortality predicting paradigm, for which the ASAB framework that we will focus on in this chapter is a particular case. In the age-specific age set based paradigm, we split the overall predicting target into multiple individual target ages and separately consider the mortality prediction for each target age, say $x_0 \in \{x_L, x_L+1, \ldots, x_U\}$ of a target population,

to fully take advantage of borrowing hidden information among different ages. For each individual target age x_0, the general paradigm aims to determine an age-specific age set $\mathcal{A}^*_{x_0}$ to be used in the calibration of a mortality model through a certain searching procedure. Assuming that mortality data in the study lies in an age-year window $\mathcal{X} \times \mathcal{T}$, with $\mathcal{X} = \{x_L, x_L + 1, \cdots, x_U\}, \mathcal{T} = \{t_L, t_L + 1, \cdots, t_U\}$. Given the age-specific age set $\mathcal{A}^*_{x_0}$ for target age $x_0 \in \mathcal{X}$, the Lee-Carter model or the CBD model is calibrated only using the data of all the ages within the set $\mathcal{A}^*_{x_0}$:

$$\log m(x,t) = a_x + b_x k_t + \epsilon_{x,t}, \quad x \in \mathcal{A}^*_{x_0}, \;\; x_0 \in \mathcal{X}, \tag{4.1}$$

$$\operatorname{logit} q(x,t) = \log\left[\frac{q(x,t)}{1-q(x,t)}\right] = K_t + (x - \bar{x})k_t + \epsilon_{x,t}, \quad x \in \mathcal{A}^*_{x_0}, \;\; x_0 \in \mathcal{X}, \tag{4.2}$$

with $t \in [t_L, t_U]$, and $\epsilon_{x,t}$'s are the independent and identically normally distributed with zero mean and age-specific variance σ_x^2. The future mortality prediction for each target age x_0 is then obtained by extrapolation. Although we have introduced both the Lee-Carter model and the CBD model under the proposed framework, we would focus on the study of Lee-Carter model in the subsequent empirical study in Section 4.3 for illustration. We adopted the SVD-based calibrating procedure for Lee-Carter model. Then for each specifc target age x_0, we fit the calibrated period sequence k_t by a random walk process with drift (RWD) for the corresponding Lee-Carter model and extrapolate based on the calibrated time series model, just like what we did in Chapters 2 and 3.

4.2.2 Age-Specific Age Band Method

One specific candidate to determine $\mathcal{A}^*_{x_0}$ for the predicting target age x_0 is to consider borrowing information from its adjacent ages under the rationale that the mortality development patterns for adjacent ages turn out to be more similar than ages far away from each other. This motivates us to consider $\mathcal{A}^*_{x_0}$ as an age band for an individual target age x_0 that consists of its adjacent ages from $\tilde{x}_L$ to $\tilde{x}_U$, where $\tilde{x}_L \leq x_0 \leq \tilde{x}_U$. In the following two subsections, we provide two specific forms of age bands that we will use in our empirical studies in the sequel and discuss how to determine the specific ranges of age bands.

4.2.2.1 Symmetric Age Band

Definition 4.2.2.1 (Symmetric Age Band) *For a given predicting target age x_0, a symmetric age band of x_0 with a band radius r, denoted as $\mathcal{A}(x_0, r)$, consists of ages from $\tilde{x}_L$ to*

$\tilde{x}_U$, *where*

$$\begin{cases} \tilde{x}_L & = \max(x_0 - r, x_L), \\ \tilde{x}_U & = \min(x_0 + r, x_U). \end{cases} \tag{4.3}$$

As the name suggested, a symmetric age band includes an equal number of ages from both sides of the target age. For a given target age x_0, different values of hyperparameter r result in unique age bands $\mathcal{A}(x_0, r)$, representing different extents of borrowing information from the neighboring ages. Therefore, the task of determining the "optimal" symmetric age band for each target age x_0 can be accomplished by determining the age-specific "optimal" value of the hyperparameter r, denoted as $r^*_{x_0}$. It is also expected that the value of $r^*_{x_0}$ should not change much as the value of x_0 changed to some adjacent values in most cases. The similarity of mortality development patterns among nearby ages usually leads to a similar amount of information for nearby target ages that need to be borrowed to achieve the optimal prediction accuracy enhancement.

4.2.2.2 Asymmetric Age Band

Definition 4.2.2.2 (Asymmetric Age Band) *For a given predicting target age x_0, an asymmetric age band of x_0 with a lower radius r_L and an upper radius r_U, denoted as $\mathcal{A}(x_0, r_L, r_U)$, consists of ages from $\tilde{x}_L$ to $\tilde{x}_U$, where*

$$\begin{cases} \tilde{x}_L & = \max(x_0 - r_L, x_L), \\ \tilde{x}_U & = \min(x_0 + r_U, x_U). \end{cases} \tag{4.4}$$

The length of its age band, denoted as $l = \tilde{x}_U - \tilde{x}_L + 1$, can also serve as an index of the amount of information borrowed from other ages. However, a given target age x_0 and a given length l does not necessarily yield a unique asymmetric age band since all $\mathcal{A}(x_0, r_L, r_U)$ with $r_L + r_U + 1 = l$ satisfy the constraint on the length of age band.

To ensure the optimal asymmetric age band is determined by the selection of the optimal value of hyperparameters l, we need a searching procedure to search for the optimal asymmetric age band with different given values of l in a computationally efficient manner. We design a fully data-driven DSA algorithm to fulfill the task in a computationally efficient manner. The idea of the DSA algorithm takes the same spirit as the one from Diao et al. [2021] (or Chapter 2), where a DSA algorithm has been developed for the population selection problem in multi-population mortality modeling by providing a recommendation for the proper choice of reference populations. The DSA algorithm for determining the age

band in the current context starts with an age set of size one containing only the target age x_0 itself and searches iteratively on the two-dimensional covariate (r_L, r_U) from $(0, 0)$ to a pre-specified upper limit (A, B). We apply the in-sample sum of squared errors (SSE) as the risk function in the DSA algorithm, while we may also try with other risk functions such as Mean absolute error(MAE), Mean squared percentage error (MSPE), and Mean absolute percentage error(MAPE), those that have been studied in Chapter 1. Assuming the training data has a time window $\mathcal{T}_{train}$ (which does not necessarily consist of consecutive years as $\mathcal{T}$ does), we calculate the risk for a given age band $\mathcal{A}(x_0, r_L, r_U)$ as follows:

$$f(\mathcal{A}(x_0, r_L, r_U)) = \sum_{t \in \mathcal{T}_{train}} \left(\eta_{x_0,t} - \hat{\eta}_{x_0,t}\right)^2, \tag{4.5}$$

in which $\eta_{x_0,t}$ can be $\log m(x_0, t)$ or $\operatorname{logit} q(x_0, t)$, depending on the choice of mortality model (i.e., Lee-Carter or CBD models), and $\hat{\eta}_{x_0,t}$ is the calbriated value of $\eta_{x_0,t}$ using the age band $\mathcal{A}(x_0, r_L, r_U)$.

Assuming the current asymmetric age band for target age x_0 at the beginning of each iteration has a length of $l = a + b + 1$ with $r_L = a$ and $r_U = b$, the algorithm updates the current asymmetric age band by checking whether the outputs from the following three moves (see Figure 4.1 for graphical illustration) can yield better result compared with the current records:

- **Deletion:** Delete one age from the left or right of the current age band. This would result in two new age bands with a length of $l = a + b$: $(r_L, r_U) = (a - 1, b)$ if $a > 0$ or $(r_L, r_U) = (a, b - 1)$ if $b > 0$. The better one between these two outputs would be chosen to compare with the current recorded "best" age band with a length of $l = a + b$.

- **Substitution:** Substitute one age from the left or right of the current age band. This would result in two new age bands with a length of $l = a+b+1$: $(r_L, r_U) = (a-1, b+1)$ if $a > 0, b < B$ or $(r_L, r_U) = (a + 1, b - 1)$ if $a < A, b > 0$. The better one between these two outputs would be chosen to compare with the current recorded "best" age band with a length of $l = a + b + 1$.

- **Addition:** Add one more age to the left or right of the current age band. This would result in two new age bands with a length of $l = a + b + 2$: $(r_L, r_U) = (a + 1, b)$ if $a < A$ or $(r_L, r_U) = (a, b + 1)$ if $b < B$. The better one between these two outputs would be chosen to compare with the current recorded "best" age band with a length of $l = a + b + 2$.

Similar to the DSA algorithm in Chapter 2, the search follows the order of deletion-substitution-addition and would stop when $a = A$ and $b = B$.

Implementing the DSA algorithm generates a sequence of (asymmetric) age bands in consecutively increasing lengths from $l = 1$ to $l = A + B + 1$, and each age band in the sequence is the one that minimizes the value of the exploited risk function over all examined age bands of the same length in the algorithm. Instead of directly examining all combinations of $(r_L, r_U) \in \{(a, b), a \in [0, A], b \in [0, B]\}$, the DSA algorithm provides an effective "greedy" searching scheme that yields age bands with the smallest value of the exploited risk function compared to a large portion of age bands of the same length in a much more computational friendly manner. Therefore, the "optimal" age band can be chosen among the resulting sequence by determining the value of l that yields to the smallest risk function value. We denote the resulting optimally selected value of l by $l^*_{x_0}$. Similar to what we commented about the $r^*_{x_0}$ sequence in the preceding subsection, the value of $l^*_{x_0}$ is expected to stay relatively stable as the value of x_0 changes to some adjacent values.

4.2.2.3 Selection of Age Band

As mentioned in previous subsections, the optimal age band can be chosen among these inputs of age bands with different radii r or lengths l by determining the optimal value of corresponding hyperparameters, r (for the symmetric age band) or $l = r_L + r_U + 1$ (for the asymmetric age band), which can be realized by using a general validation procedure. In order to make full use of the data, We propose a block cross-validation-style method, inspired by Bergmeir and Benítez [2012] and [Bergmeir et al., 2014]. Block cross-validation is regarded as a variant of conventional cross-validation for time series data. As mentioned in [Bergmeir et al., 2014], making full use of the data brings in the advantage of more precise error estimates of the prediction for the blocked cross-validation scheme and meanwhile assures independence by leaving a margin of a certain distance in time between the training and validating blocks. The method recently has been in favor in the mortality community as a new tool to provide robust validation error estimates, see [Kessy et al., 2021] and [SriDaran et al., 2022] as examples.

In the block cross-validation method, all available training data is partitioned into k blocks sequentially. Each block of sequential data is used as the validation set once to evaluate the performance of different models, while the rest $k-1$ blocks are used as the training blocks after s-year data are left out from both sides of the validating block. The omission of some data on borders is applied to ensure the approximate independence between training and validating data. The choice of k depends on the trade-off between the computational

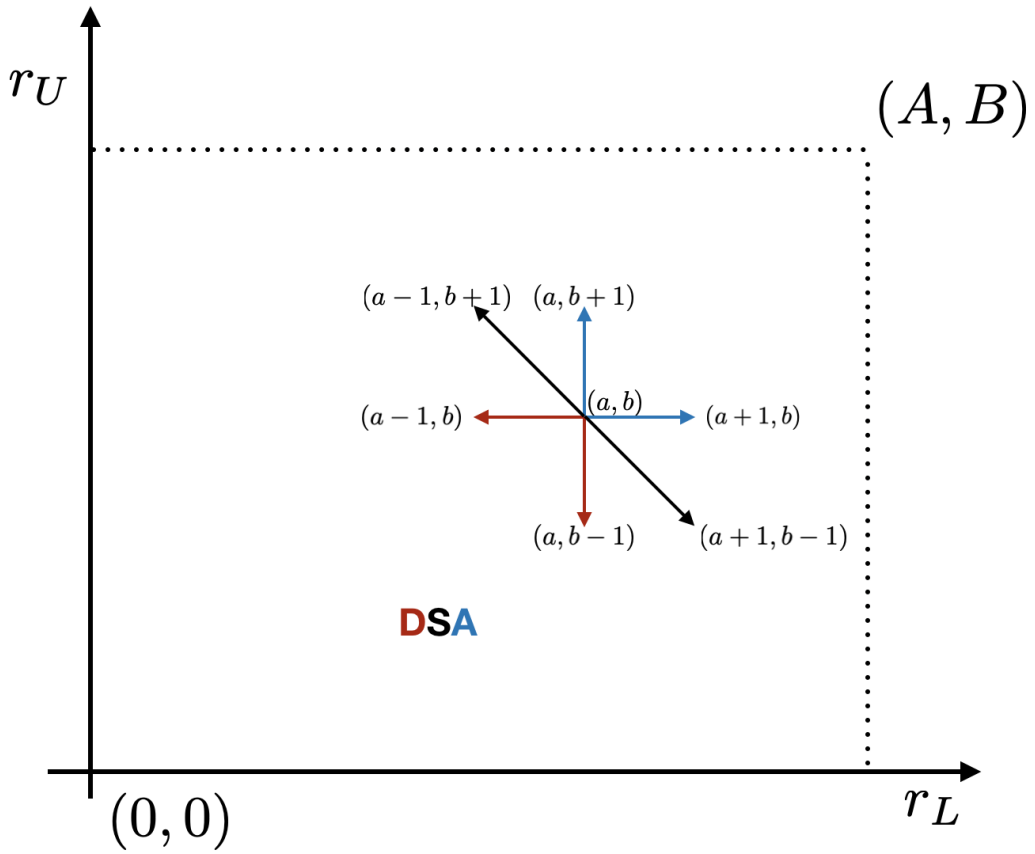


Figure 4.1: Illustrative figure of the DSA iterative updating scheme for creating a sequence of asymmetric age bands.

cost and the amount of available data. A larger k leads to an increase in the number of models to be estimated while a smaller k leads to a smaller training set. Similarly, the choice of s (i.e., the size of the data omitted from the validation block of data) depends on the trade-off between the degree of independence and the amount of available data. A large s leads to a considerable loss of data and may even result in an insufficient amount of data for model estimation while a small s can not ensure the approximate independence between training and validating data. An illustrative example can be found in Figure 4.2, which demonstrates how the data is split into training blocks and validating blocks for a 4-fold block cross-validation in our empirical study.

After choosing a validating block with a time window $\mathcal{T}_{validate}$ and the training blocks with a time window $\mathcal{T}_{train}$, the age-specific coefficients a_x and b_x in the Lee-Carter model (4.1) can be estimated using a standard SVD procedure with the training blocks. The SVD procedure also gives the values of the period index k_t over the training time window but no values for k_t over the validating time window $\mathcal{T}_{validate}$. These missing values are imputed by fitting a random walk with drift using all observable period indices on both sides of the validating block as

$$k_t = k_{t-1} + d + \epsilon_t, \tag{4.6}$$

where d serves as the drift parameter and ϵ_t as the noise term. A forward/backward fill approach, shown in Equation (4.8) would be then utilized to impute the missing values of k_t in $\mathcal{T}_{validate}$ based on estimated $\hat{d}$ and realized values of k_t in $\mathcal{T}_{train}$:

$$\textbf{forward fill}: \hat{k}_t = k_{t-\alpha} + \alpha\hat{d}, \tag{4.7}$$

$$\textbf{backward fill}: \hat{k}_t = k_{t+\alpha} - \alpha\hat{d}. \tag{4.8}$$

For each chosen validation block, the validation SSEs of logarithmic mortality rates of target age x_0 for a given age band $\mathcal{A}(x_0, \cdot)$ of any forms are calculated as

$$f(\mathcal{A}(x_0, \cdot)) = \sum_{t \in \mathcal{T}_{validate}} \left(\log m_{x_0,t} - \log \hat{m}_{x_0,t}\right)^2. \tag{4.9}$$

The validation SSEs are then summed up as the BCV-SSE. Finally, $\mathcal{A}^*_{x_0}$, the optimal age band for target age x_0 is selected among all available age bands $\mathcal{A}(x_0, \cdot)$ as the one with the least BCV-SSE.

The above model fitting and cross-validation procedure are described for the Lee-Carter model. The procedures can be applied in parallel for other mortality models, e.g., CBD model.

4.2.2.4 Smoothing

With the aforementioned blocked cross-validation procedure, the optimal age band for target age x_0 has been decided with a sequence of values of age-specific hyperparameters $r^*_{x_0}$ or $l^*_{x_0}$ being determined by comparing the values of the realized BCV-SSE. However, the randomness of the realized values may potentially influence the results of the comparison and thus result in steep ups and downs in the $r^*_{x_0}$ or $l^*_{x_0}$ as the target age x_0 changes. As mentioned in previous discussions, nearby target ages usually borrow similar amounts of information to achieve the optimal prediction accuracy enhancement, leading to a relative stableness in the chosen values of $r^*_{x_0}$ or $l^*_{x_0}$ with respect to different values of x_0. Smoothing, as a standard tool to create an approximating function that attempts to capture important patterns in the sequence while leaving out the influence of noise, is considered to be applied to the $r^*_{x_0}$ or $l^*_{x_0}$ sequence to adjust their final values in an attempt to retain the stableness with different values of x_0.

Many different smoothing algorithms can be considered while we adopt the following "moving average" scheme, where each point in the original sequence is replaced with the average of h adjacent points:

$$r^*_{x_0.\mathbf{smooth}} = \frac{1}{2h+1}\left(r^*_{x_0-h} + r^*_{x_0-h+1} + \cdots + r^*_{x_0} + \cdots + r^*_{x_0+h}\right) \tag{4.10}$$

where h is known as the radius of the smoothing window. For values at the beginning and the end of the sequence, the averaging function is applied to a smaller sections of the array, from $r^*_{x_L}$ to $r^*_{x_0+h}$, or from $r^*_{x_0-h}$ to $r^*_{x_U}$. In this chapter, different values of h have been considered to represent the various degrees of desired smoothness.

4.3 Empirical Study

In this section, we empirically evaluate the performance of our proposed ASAB based prediction methods via analysis of the Human Mortality Database (HMD) and make comparisons with clustering-based methods proposed in Tsai and Cheng [2021]. As one would see shortly, the empirical study confirms a noticeable improvement in overall mortality predicting accuracy by our proposed ASAB based method over the clustering-based benchmark models. More specifically, the ASAB based method reduces predicting error for a majority of ages, especially for the adult and retiree age groups that contain ages 30 to 89, while it yields a comparable predicting performance with the benchmark models for the other age groups.

4.3.1 Empirical Setting and Predictive Models

We consider the mortality data from 24 populations as listed in Table 3.1 in Section 3.4 of Chapter 3 for a age-year window $\mathcal{X}\times\mathcal{T}$ with $\mathcal{X} = \{0, 1, \cdots, 100\}$ and $\mathcal{T} = \{1970, 1971, \cdots, 2010\}$. The mortality data is split into a training set (1970–2002) and a test set (2003–2010). Each involved prediction model is trained using the training data set (1970–2002) and the resulting models are extrapolated to the testing period (2003–2010) to obtain mortality prediction.

The prediction models we consider in this empirical study include:

- **Lee-Carter**: As a benchmark model, the Lee-Carter model is fitted to each population separately with the full-age band from 0 to 100 for each gender.
- **ASAB Based Lee-Carter Model**: The ASAB based approach embedded with the Lee-Carter model is applied for the mortality prediction of each of the 24 populations. The details of the implementation procedure have been described in Section 4.2. We temporarily consider symmetrical age bands in the study and will investigate the effect of asymmetrical age bands later.

  To adopt the block cross-validation procedure, the training data of 33 years (1970–2002) are decomposed into 4 non-overlapping blocks with a length of roughly 8 years for each. Each of the four blocks would be used as the validating block once to evaluate the performance of prediction models with different age bands, and the data in the closest $s = 5$ years on both sides of the validating block are omitted. The remaining data after omission are used to train the prediction model and the final choice of optimal age band for target age x_0 depends on the overall performance on all validating blocks. Figure 4.2 shows how the 4-fold block cross-validation works on the training set (1970–2002).

  Depending on whether we apply the smoothing step in determining the age band size, we have two prediction models from the ASAB based method:

  - **LC-ageband**: Use the aforementioned 4-fold BCV procedure to search for the optimal symmetric age band for each age among symmetric age bands with different band radii r, including both the age band containing the target age x_0 itself only and the full-age band including all ages from 0 to 100.
  - **LC-ageband-smooth**: Based on the age band sizes determined in the **LC-ageband** model, a smoothing step of moving average scheme with the length of smoothing window $h = 4$ is applied. That is, for each target age x_0, the smoothed

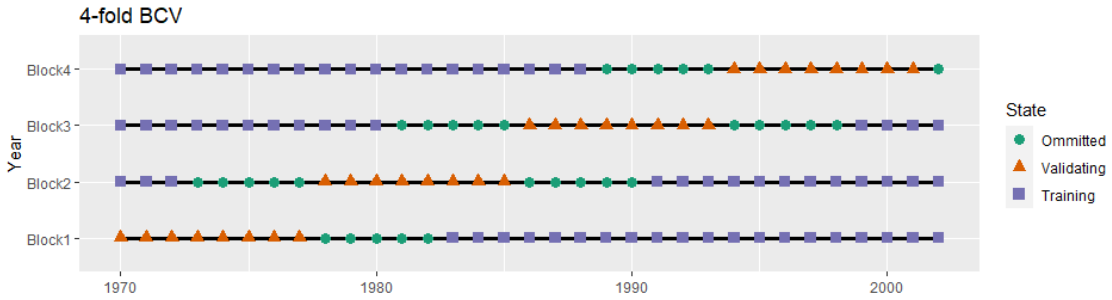


Figure 4.2: Illustrative figure of a 4-fold blocked cross validation (BCV). Blue squares represent points from the time series used for model training, orange triangles represent points used for validation, and green circles represent the omitted data points.

value $r^*_{x_0.\mathbf{smooth}}$ computed as the average of $\{r^*_{x_0-4}, r^*_{x_0-3}, \cdots, r^*_{x_0+4}\}$ is used as the radius of the selected age band.

- **Clustering-based models**: The clustering-based methods use the results of clustering analysis to determine ages that should be regarded as in the same cluster to be included in the calibration of a mortality model. As we have claimed in previous sections, we apply the Lee-Carter model to the data for each resulting cluster of ages from the clustering analysis as Tsai and Cheng [2021] does. Specifically, we let $Y_{x,t}$ be the annual increment of age-specific mortality level of age x from year t to year $t-1$, i.e.,

$$Y_{x,t} = \log m(x,t) - \log m(x,t-1), \quad t \in [t_{L+1}, t_U]. \tag{4.11}$$

We take time series $Y_{x,t}$ as the clustering feature for each age x. With a normal assumption for each element in the age-specific $Y_{x,t}$ sequences with mean μ_x and variance σ_x^2, we can assume age x_1 and x_2 should be classified into the same cluster if the pairs $(\mu_{x_1}, \sigma^2_{x_1})$ and $(\mu_{x_2}, \sigma^2_{x_2})$ are similar. Therefore, the clustering algorithms are implemented on the set of two-dimensional objects $\{(\hat{\mu}_x, \hat{\sigma}^2_x) : x \in \mathcal{X}\}$, where

$$\hat{\mu}_x = \bar{Y}_{x,t} = \frac{1}{n-1} \sum_{t_{L+1}}^{t_U} Y_{x,t}, \tag{4.12}$$

$$\hat{\sigma}^2_x = \frac{1}{n-2} \sum_{t_{L+1}}^{t_U} \left(Y_{x,t} - \bar{Y}_{x,t}\right)^2. \tag{4.13}$$

Depending on which specific clustering method is used, we have three different prediction models with their labels corresponding to a clustering method as follows:

– **LC-Kmeans**: Use the K-means clustering method to determine each age cluster. We use the recommended value of K from `NbClust` function in `R` package `NbClust` [Charrad et al., 2014] to determine the number of clusters.

– **LC-HCluster**: Use Ward's agglomerative clustering method [Ward Jr, 1963] to determine each age cluster, in which a bottom-up merging approach using the sum of squared errors as the objective function on the two-dimensional objects $(\hat{\mu}_x, \hat{\sigma}_x^2)$ is applied. At each step, the pair of clusters that leads to a minimum increase in total within-cluster variance is merged. The optimal number of clusters, as another important feature of this clustering algorithm, is decided by the recommended number of clusters from applying `NbClust` function in `R` [Charrad et al., 2014].

– **LC-GMM**: Use the Gaussian mixture model (GMM) clustering method to determine each age cluster. The GMM clustering method assumes a mixture of Gaussian distribution to the data. By iteratively modifying the parameters of the mixture distribution until they best fit the underlying data, each component distribution of the mixture models represents a cluster. The GMM clustering algorithm consists of three cores: initialization via other clustering results, maximum likelihood estimation via the EM algorithm, and the selection of models and the number of clusters via approximate Bayes factors with the BIC. We follows the strategy proposed by Fraley and Raftery [2002] to fit the data with a Gaussian mixture structure using `mclust` function in `R` package `mclust` [Scrucca et al., 2016]. A detailed description of the method can be found in Fraley and Raftery [2002] as well as Tsai and Cheng [2021].

4.3.2 Predicting Performance

Similar to those empirical studies in the previous chapters, the overall prediction accuracy is evaluated in terms of test SSEs, which are denoted by $e(t)$ and calculated as follows:

$$e(t) = \sum_{x=0}^{100} \left[\log m(x,t) - \log \widehat{m}(x,t)\right]^2, \quad t = 2003, \ldots, 2010,$$

where $m(x,t)$ is the mortality rate for the target population, and $\widehat{m}(x,t)$ is its predicted quantity. For a succinct overview, the overall test SSE as $\sum_{t=2003}^{2010} e(t)$ for each target population over the 24 populations in the pool is computed and we obtain 24 overall test SSEs. The 1st quartile, median, mean, and the 3rd quartile of the resulting 24 overall test SSEs for each predicting model are reported in Table 4.1 as a general comparison in predicting

accuracy among different models, in which a smaller value implies a better prediction performance. The results in the table indicate that our proposed ASAB based method with smoothing (i.e., LC-ageband-smooth) substantially outperforms the benchmark Lee-Carter model as well as the three clustering-based models with a smaller median and mean of the resulting test SSEs, for both female and male populations. The 1st and 3rd quartiles of Test SSEs from ASAB based method with smoothing are also generally smaller than those from other prediction models. The ABAS based method without smoothing (i.e., LC-ageband) also shows comparable predicting power as the LC-GMM method, which is the best among all the three considered clustering based methods, a result also confirmed by empirical studies with mortality data for both genders of the US and the UK in Tsai and Cheng [2021].

We also apply the one-sided Diebold-Mariano (DM) test (see Diebold and Mariano [1995] and Harvey et al. [1997]) to compare the predicting accuracy between the group of the two ASAB based methods and the group consisting of the Lee-Carter model and the three clustering based models. The comparison procedure is the same as what we conducted in Section 2.3.3.2 of Chapter 2 and Section 3.4.2.2 of Chapter 3. We quickly summarize the comparison procedure again below for readers' convenience. For a pair of models in comparison, two one-sided DM tests with the corresponding null hypothesis that one model is no better than the other are conducted based on the sequences of age-aggregated test SSEs from both models for the same population in the 24 populations. If one model is concluded to be significantly better than the other by the corresponding test with a p-value smaller than 0.05, we count it as a win of that model. The comparison is made on the number of wins obtained between the pair of models, a larger number indicates a more advanced position in improving predicting accuracy for the corresponding models. Table 4.2 reports the comparison results. Each cell of the table contains two integers recording the comparison results of the model in the row versus the one in the column, with the first integer as the number of wins of the model in the row, and the second as the number of wins of the model in the column. As Table 4.2 clearly indicates, the ASAB based models significantly outperform the benchmark models in terms of providing more accurate predictions, especially the one with a smoothing procedure applied to the selected age band sizes.

We are also interested in investigating whether the relative outperformance of our ASAB based methods may vary over different ages. To this end, we decompose the test SSEs with a finer partition in ages. We draw the boxplots of test SSEs (in its logarithmic scale) of 24 populations obtained for different age groups under various prediction models in Figure 4.3. The full age range is decomposed into the following six non-overlapping age groups:

Table 4.1: Summary statistics of test SSEs of 24 female populations and 24 male populations comparing the prediction performance of different predicting methods.

	1st Quartile	Median	Mean	3rd Quartile
Female Population				
Lee-Carter	19.37	44.73	44.64	63.90
LC-GMM	16.80	44.44	45.18	66.18
LC-Kmeans	18.01	46.84	45.15	71.06
LC-HCluster	16.99	45.43	45.02	72.40
LC-ageband	17.44	44.49	44.00	67.36
LC-ageband-smooth	14.81	41.94	40.57	59.22
Male Population				
Lee-Carter	28.40	41.31	49.79	67.97
LC-GMM	24.70	35.98	40.60	59.77
LC-Kmeans	27.21	38.55	45.97	62.67
LC-HCluster	27.15	38.92	46.30	62.67
LC-ageband	16.78	37.67	39.14	60.66
LC-ageband-smooth	16.01	34.99	37.44	53.90

Table 4.2: Number of wins for comparisons between the proposed models (in the row) and the benchmark models (in the column) based on a pairs of one-sided DM tests: In each cell, the first integer indicates the number of wins of the model in the row out of 24 comparisons while the second integer is the number of wins of the model in the column.

	Lee-Carter	LC-GMM	LC-Kmeans	LC-HCluster
Female Population				
LC-ageband	$(10, 4)$	$(10, 6)$	$(13, 5)$	$(12, 5)$
LC-ageband-smooth	$(17, 0)$	$(19, 1)$	$(19, 1)$	$(17, 1)$
Male Population				
LC-ageband	$(18, 1)$	$(11, 4)$	$(15, 1)$	$(15, 2)$
LC-ageband-smooth	$(22, 0)$	$(19, 3)$	$(19, 0)$	$(17, 0)$

- **Child:** ages 0 to 9;
- **Teenage:** ages 10 to 19;
- **Young Adult:** ages 20 to 29;
- **Adult:** ages 30 to 59;
- **Retiree:** ages 60 to 89;
- **Elder:** ages larger than 89.

As shown in Figures 4.3 and 4.4, the extent of improvement in predicting accuracy by our proposed ASAB based methods varies over genders and age groups. For female populations (refer to the first column in Figures 4.3 and 4.4), major improvement in predicting accuracy of adopting the proposed age-band-base methods has been observed in the adult and retiree age groups that together contain ages 30 to 89, while the predicting performance remains just comparable to other methods for other age groups. For male populations (refer to the second column in Figures 4.3 and 4.4), all the ages except the groups of Child (ages 0 to 9) and Elder (ages larger than 89) embrace an obvious benefit of reduced predicting error brought in by the proposed ASAB based methods.

Table 4.3: Population-specific test SSEs comparing the prediction performance of the ASAB based method versus benchmark models.

	LC-GMM	LC-ageband-smooth	Change%
	Male Population		
England & Wales	15.03	10.98	-26.97%
France	12.57	10.57	-15.89%
New Zealand	60.92	52.68	-13.52%
Scotland	62.91	64.13	1.94%
Australia	29.90	26.28	-12.08%
Austria	33.28	33.39	0.34%
Belgium	37.01	37.80	2.11%
Canada	16.70	12.29	-26.41%
Czech Republic	37.82	36.59	-3.27%
Denmark	76.36	71.81	-5.97%
Finland	63.09	62.57	-0.81%
Hungary	45.62	48.16	5.57%
Italy	32.65	24.53	-24.85%
Japan	9.90	6.78	-31.52%
Netherlands	32.93	32.78	-0.44%
Norway	68.57	66.73	-2.69%
Poland	14.73	13.16	-10.66%
Portugal	91.13	86.57	-5.00%
Slovakia	48.19	44.63	-7.38%
Spain	34.94	25.00	-28.46%
Sweden	56.70	54.49	-3.89%
Switzerland	59.39	53.70	-9.58%
Taiwan	27.36	16.96	-38.02%
U.S.A.	6.78	6.03	-11.01%
MEAN	**40.60**	**37.44**	**-11.19%**
MEDIAN	**35.98**	**34.99**	**-8.48%**

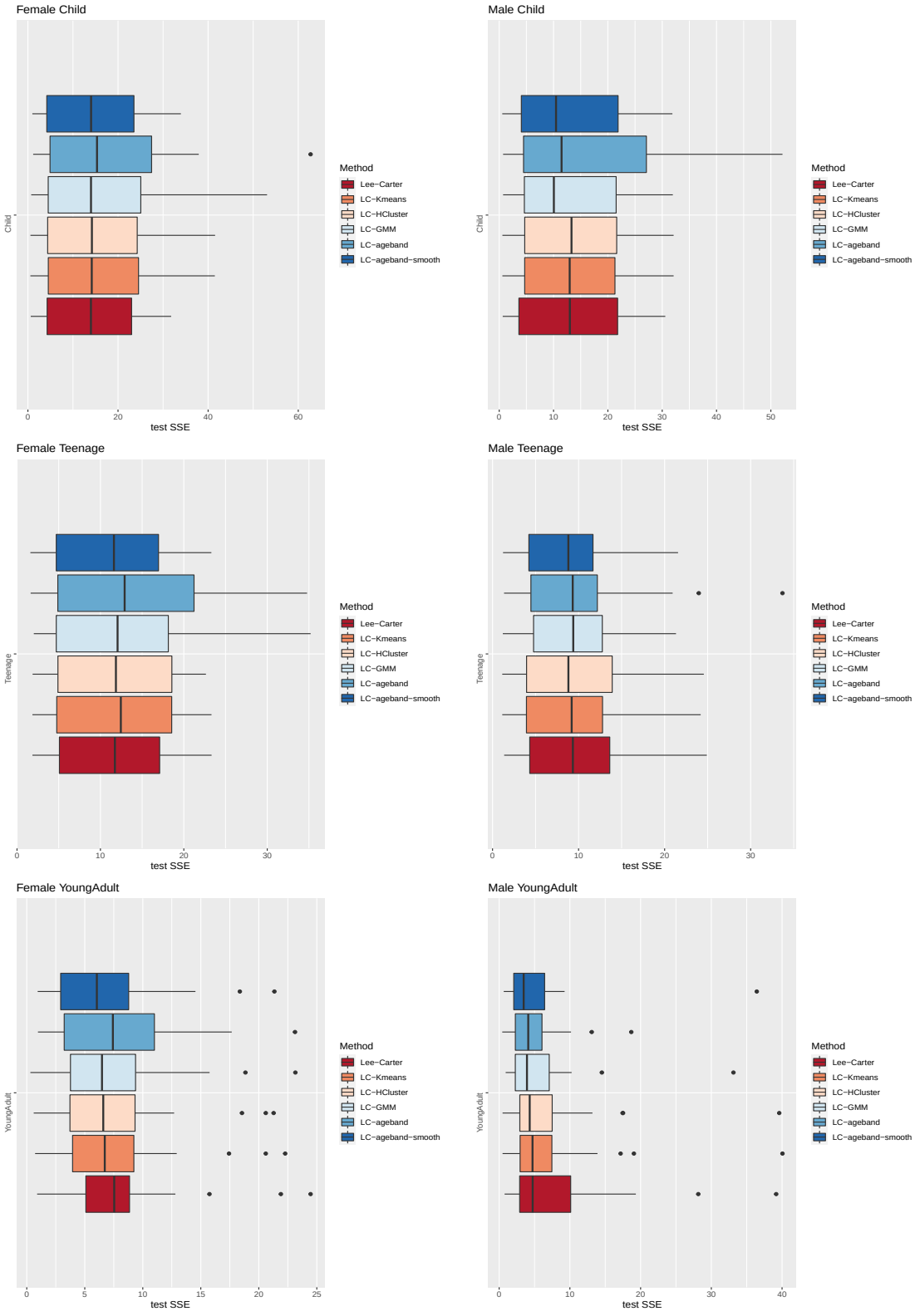


Figure 4.3: Boxplots of test SSEs of 24 female populations (left panel) and 24 male populations (right panel) comparing the prediction performance of different methods for different age groups: Child, Teenage, and Young Adult.

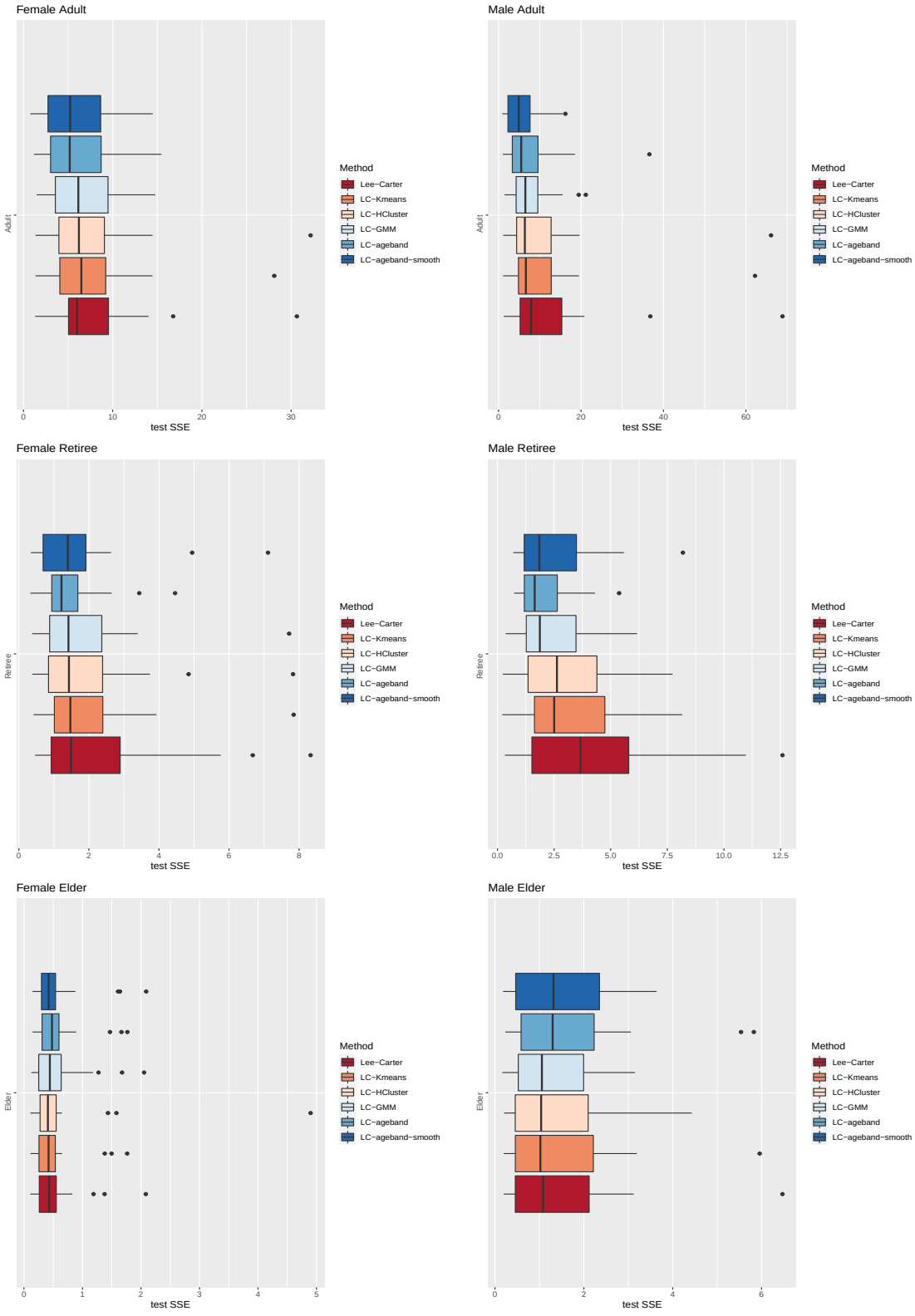


Figure 4.4: Boxplots of test SSEs of 24 female populations (left panel) and 24 male populations (right panel) comparing the prediction performance of different methods for different age groups: Adult, Retiree, and Elder.

4.3.3 Influence of Smoothing

In the proposed ASAB based method, a smoothing step was proposed to denoise and capture how the band radius $r^*_{x_0}$ sequence changes with respect to different values of x_0. The numerical results in previous subsections have confirmed that the smoothing step with a specific degree of smoothness has led to an improvement in prediction accuracy. A natural question to answer is how different degrees of smoothness, represented by different values of h, would influence the extent of improvement in prediction accuracy. In fact, a large range of values for the length of smoothing window h, representing different scenarios of no smoothness($h = 0$), week smoothness($0 < h < 3$), moderate smoothness($3 \leq h < 10$) and strong smoothness ($h = 20$ or even $h = 50$), have been considered in our empirical study while we only reported results for $h = 4$ in the preceding subsection. To illustrate how the performance of our ASAB based method may vary over the different choices of the smoothing window length h, we demonstrate the resulting average realized test SSEs over the 24 populations in Figure 4.5 for both genders. We also include results of the original Lee-Carter model and results of another smoothing scheme based on median, known as "3RS3R" (see [Tukey et al., 1977]) in Figure 4.5 for comparison. According to the figure, the resulting overall prediction accuracy of the method is quite stable under moderate smoothness with $3 \leq h < 10$, and the resulting overall realized SSEs are the most favorable. While we did not report the specific realized SSEs under smoothness window lengths different from 4, their overall magnitudes are quite similar for different values of h with moderate and median smoothness. The median smoothness yields larger but very slightly realized SSEs. We thus recommend considering smoothing with moderate smoothness while different choices of smoothing algorithms appear to be less important to the goal of optimally enhancing predicting accuracy.

Another interesting question is whether there is any pattern in the selected band size as the target age varies from 0 to 100. To address this question, we compute the average of the selected smoothed age-specific band radius $r^*_{x_0.\mathbf{smooth}}$ over the 24 populations, and illustrate its values in Figure 4.6. It clearly indicates that a smaller band radius is preferred for middle ages while a larger band radius is preferred for both younger and elder ages, which provides the insight that middle ages only need to borrow information from a moderate number of their neighboring ages to benefit their future mortality prediction while the young and elder ages, in general, need to borrow information from more ages.

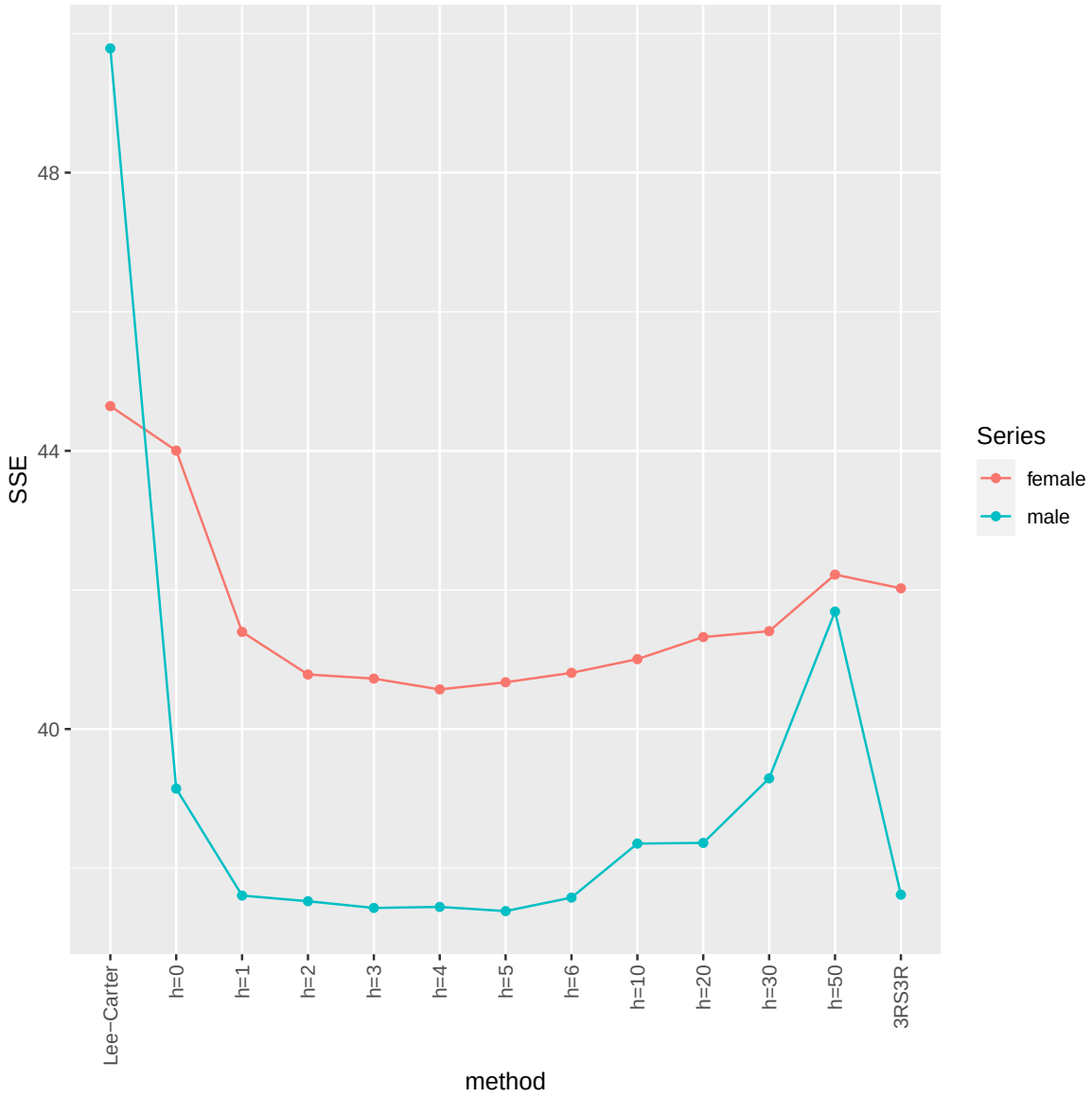


Figure 4.5: The Averaged values of test SSEs of all 24 populations with different smoothing settings.

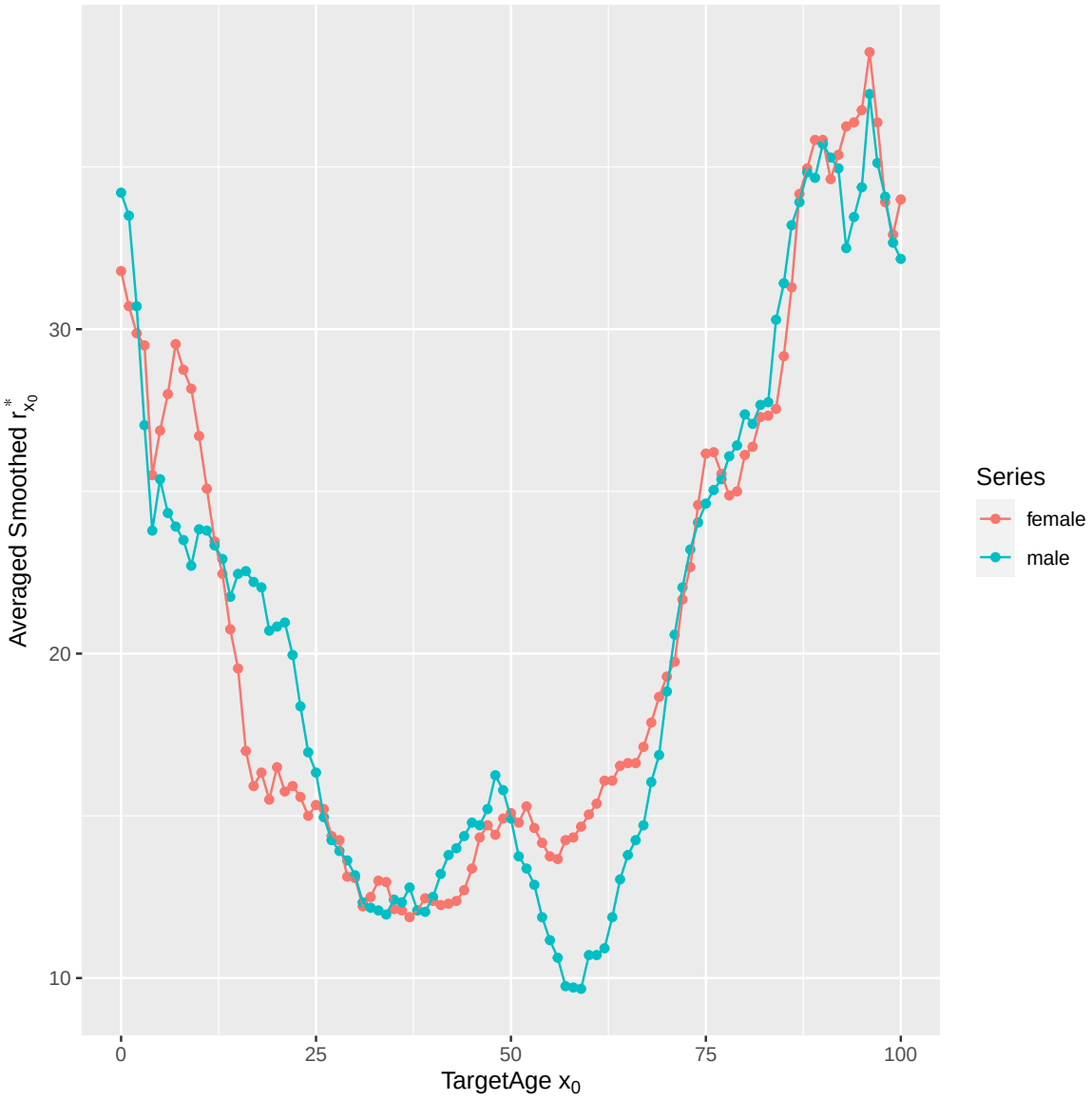


Figure 4.6: The averaged values of smoothed optimal band radius $r^*_{x_0.\mathbf{smooth}}$ with respect to different target ages.

4.3.4 Consistency of Age-Band-Selection-Based Methods

There exist recent discussions about the issue of age coherence. Age coherence, as the name suggested, refers to the phenomenon that mortality rates at different ages do not diverge in the long run [Li and Lu, 2017]. Some literatures have viewed age coherence as a desired property for the long-term forecast of mortality data. To resolve issues regarding age coherence, Li et al. [2013] proposed a "rotation" approach that let the age effect terms become time-variant and gradually converge to an ultimate structure. Gao and Shi [2021] allowed for the geometric and hyperbolic decayed relative speed for the change of the age effect terms over the out-of-sample forecasting steps. However, many classic mortality models do not have a specially designed structure that satisfies the age coherent constraints, e.g., Lee-Carter model, CBD model, and ACF model. In this subsection, we are not providing resolutions to the age coherence issue but conduct numerical examination on whether the proposed ASAB based methods can ensure a comparable degree of consistency in its mortality forecasts when comparing to the classic benchmarks.

The numerical comparison is based on a proposed index named the crossing ratio, where a specific form of violation of consistency is examined. A crossing happens when the relative order of age-specific logarithmic mortality rates of two ages x_1 and x_2 at two different time spots t_1 and t_2 are different, i.e., a crossing happens when

$$(\log m_{x_1,t_1} - \log m_{x_2,t_1}) \times (\log m_{x_1,t_2} - \log m_{x_2,t_2}) < 0. \tag{4.14}$$

The crossing ratio then records the proportion of amount that a crossing happens among all possible combinations of (x_1, x_2) over the age-window $\mathcal{X} = \{0, 1, \cdots, 100\}$ for pre-specified t_1 and t_2 in the year-window $\mathcal{T} = \{1970, 1971, \cdots, 2010\}$. The crossing ratio, whose value ranges from 0 to 1, is used as an index to demonstrate the degree of age coherence. A value close to 0 indicates strong consistency whereas a value close to 1 indicates a severe violation of consistency. In our empirical study, the crossing ratios for forecasts from different prediction models applied to the 24 populations are calculated, with a fixed ending time point $t_2 = 2010$, which is the last year of the testing period. The average crossing ratios over the 24 populations are reported in Table 4.4. to examine the average performance of each method in terms of maintaining age coherence, with the different starting time points t_1 indicating both a short-term case (when $t_1 = 2002$, which is the last year of the training period) and a long term case (when $t_1 = 1970$, which is the first year of the training period). For the row "Real Data" in the table, the realized data are used for the logarithmic mortality rates $\log m_{x_1,t_2}$ and $\log m_{x_1,t_2}$ in Equation (4.14) defining a crossing, while predictive values are used for the other rows in the table. As Table 4.4 indicates, a reasonable level

of consistency has been generally observed, both in real data and all predictive mortality models. In particular, the ASAB based methods have achieved comparably smaller cross ratios, compared to Lee-Carter model and even the real data in both the short-term and the long-term comparisons. This confirms that the ASAB based methods, both the smoothed and the non-smoothed version, have demonstrated their abilities to maintain a desirable level of consistency of its mortality forecasts.

Table 4.4: Averaged crossing ratio with different pre-specified t_1 and t_2.

	$(t_1, t_2) = (2002, 2010)$	$(t_1, t_2) = (1970, 2010)$
Female Population		
Real Data	2.42%	3.21%
Lee-Carter	1.75%	3.13%
LC-ageband	1.70%	3.18%
LC-ageband-smooth	1.72%	3.10%
Male Population		
Real data	1.95%	2.90%
Lee-Carter	1.45%	2.81%
LC-ageband	1.32%	2.77%
LC-ageband-smooth	1.39%	2.70%

4.3.5 Asymmetric Age Band

In the empirical studies of previous subsections, the ASAB based method using a symmetric age band has been compared with other benchmark models and proved its outperformance. A natural direction of possible extensions is to consider adopting asymmetric age bands, which would in theory add extra flexibility in the form of the age band for each target age and therefore potentially lead to additional benefits for prediction enhancement. As mentioned earlier in this chapter, the selection of the "optimal" asymmetric age band can be achieved by the following two steps:

1. Adopt a fully data-driven DSA-based algorithm to search for the "optimal" asymmetric age band with different given values of l, ranging from 0 to a specific upper limit L;

2. Adopt a BCV procedure to determine the best asymmetric age band among those "optimal" asymmetric age bands with different given values of l that have been obtained in the previous step.

To explore the potential of asymmetric age bands, we apply the same data splitting scheme as we explained at the beginning of Section 4.3.1 to ensure a comparable final result. We set the upper limit $L = 100$, which makes it possible to include both the age band containing the target age x_0 itself only and the full age band containing all ages from 0 to 100 in the output of the first step and then being considered in the second step. The search for the optimal asymmetric age band for each age is conducted with the same 4-fold BCV procedure and followed by a smoothing step of a moving average scheme with the length of smoothing window $h = 4$ as we did previously. The realized overall test SSEs for 24 populations are summarized in Table 4.5. The theoretical advantage of adopting asymmetric age bands, according to the results in Table 4.5, in general, does not brings extra superiority in providing improved predicting performance in practice.

Table 4.5: Comparing the prediction performance between symmetric ageband method, denoted as LC-ageband-smooth and asymmetric ageband method, denoted as LC-DSA-smooth.

	1st Quartile	Median	Mean	3rd Quartile
Female Population				
LC-ageband-smooth	14.81	41.94	40.57	59.22
LC-DSA-smooth	14.70	42.04	43.03	65.58
Male Population				
LC-ageband-smooth	16.01	34.99	37.44	53.90
LC-DSA-smooth	15.70	35.70	37.77	57.71

4.4 Concluding Remarks

In this chapter, we propose an age-specific age band based mortality predicting framework that aims to fully utilize the beneficial information hidden across ages in their similarities of mortality development patterns to enhance mortality prediction accuracy. The overall predicting goal is decomposed into multiple individualized predicting tasks in which an optimal age band is to be selected to determine the number of neighboring ages to borrow information. Extensive empirical studies with the Human Mortality Database were conducted to compare the proposed age-specific age band based prediction models with other benchmark models that also consider utilizing the similarities in mortality development patterns among different ages. An overall improvement of our proposed method in predicting accuracy has been observed for the majority of ages, especially for adults and retiree groups. Among all methods, we recommend using the age-specific age band based method with a smoothing procedure for the band sizes because this method shows consistently superior performance in predicting accuracy in all of our empirical studies.

Chapter 5

Borrowing Information from Multiple Aspects

5.1 Introduction

Previous studies in the preceding chapters have triggered a more profound consideration for effectively borrowing information from multiple aspects simultaneously. This chapter serves as a comprehensive fusion of previous chapters to actively explore proper strategies that consider hidden information across both ages and populations to formulate future mortality predictions for a potentially improved prediction accuracy. This chapter will provide three different specific approaches to the stated end and implement them with data from the Human Mortality Database (HMD). All of them are motivated by the question of how to extend the age-specific age set paradigm proposed in the preceding chapter to a more complex multi-population setting. The first approach is a distance-based approach, based on a variance-based distance measure that defines the "neighborhood" of the target age among all ages from multiple populations. The second approach is an ensemble-based approach built upon the distance-based approach. The third approach is the ACF model-based approach, which extends the age-specific age band (ASAB) based method from the last chapter to the ACF model framework. Performance comparisons are conducted among the proposed approaches and the single population ASAB based method as the benchmark based on empirical studies with the Human Mortality Database. The empirical studies show a noticeable improvement in prediction accuracy from the distance-based approach for male populations and a comparable predicting accuracy for female populations compared with the single ASAB based model. The ensemble-based method yields a similar prediction accuracy to the multi-population distance-based approach and performs quite robustly with different

combinations of hyperparameters. The ASAB method applied with the ACF model performs the best among all the three approaches for female populations but is less desirable for male populations. Additionally, this chapter also discloses some interesting stylized facts from the HMD about how ages would be selected from multiple populations by the distance-based approach.

The rest of the chapter proceeds as follows. Section 5.2 introduces the proposed predicting strategies, framework, and the data-mining procedures that we utilize to determine the optimal age sets. Section 5.3 presents empirical studies of our proposed methods and compares them with benchmark models considered in our study. This section also includes a summary of some stylized facts about how ages from multiple populations are selected by the distance-based approach. Finally, Section 5.4 provides some concluding remarks.

5.2 Three Predicting Approaches

5.2.1 Distance-based Approach

When it comes to extending the age-specific age set paradigm to a multiple population setting, the designed procedure should be capable of detecting the useful information hidden from a specific combination of populations and ages and, in the meanwhile, maintaining computational efficiency. One sensible way is to view each age-specific mortality rate time series from different populations as a different candidate in the age candidate pool and then consider selecting those ages in the "neighborhood" of the target age for mortality modeling. However, unlike the single-population case where the concept of "neighborhood" has its natural meaning due to the natural ordering of ages, we have to define what constitutes a "neighborhood" of a given target age in the whole pool of candidate ages. In this section, we propose a variance-based quantity to measure distance and define the "neighborhood" in order to extend the age-specific age set paradigm to the scenario of borrowing information from multiple populations.

As the name suggests, the distance-based approach works similarly to the well-known k-nearest neighbors algorithm (KNN) (Fix and Hodges [1989]), bringing in useful information from those "neighboring" ages sharing enough homogeneity in terms of a certain "distance" measure with the target age x_0 for potential improvement of prediction accuracy. So, in the distance-based approach, the desired optimal age set $\mathcal{A}^*_{x_0}$ is determined as a group of k age "neighbors" of the predicting target age x_0 with $k \in \{0, 1, \cdots, K\}$, where K represents the maximum allowed number of ages to borrow information in the age-specific age set based

paradigm.

The implementation of the approach involves designing a proper "distance" measure that is capable to capture similarities in mortality developing trends. To exclude the effect of mortality level, a general "distance" measure between two age-specific mortality sequences of age i and j (i and j may come from different populations) is designed in the following way:

1. A difference sequence is first defined as:

$$\text{diff}_{i,j}(t) = \eta(i,t) - \eta(j,t), \quad t \in \mathcal{T},$$

   which represents the difference between the two age-specific mortality sequences from age i and age j at time t. The choice of η depends on the chosen mortality model. $\eta_{i,t}$ can be $\log m(i,t)$ for Lee-Carter model and $\text{logit}\, q(i,t)$ for CBD model.

2. A dissimilarity matrix $\mathbf{D}$ is then defined with entries $D_{i,j}$ as the variance of the sequence $\text{diff}_{i,j}(t)$, $t \in \mathcal{T}$, i.e.,

$$D_{i,j} = \text{Var}\left(\text{diff}_{i,j}(t)\right).$$

The designed dissimilarity matrix is able to represent the extent of similarity shared by two different ages in the sense that a smaller value of $D_{i,j}$ means more paralleled the two age-specific mortality sequences, meaning a closer relationship in their mortality development pattern over time. The "neighboring" system can be formed based on values of elements in $\mathbf{D}$ for each target age x_0. For the choice of the optimal age set $\mathcal{A}^*_{x_0}$ associated with the target age x_0, a validation procedure is required for determining the desirable size of the set $\mathcal{A}^*_{x_0}$, denoted by k_{x_0}, conducive to ensuring mortality predicting accuracy. This number k_{x_0} means that the set $\mathcal{A}^*_{x_0}$ consists of the target age x_0 and the k_{x_0} ages closest to the target age x_0 judged by the adopted variance-based distance measure. Similar to the age-specific age band method introduced in Section 4.2.2, we apply a 4-fold BCV procedure to determine the optimal k_{x_0} value over $\{0, 1, \cdots, K\}$. The pre-specified value for the maximum allowed number of ages K may impact the performance of the proposed distance-based approach. A larger value of K would lead to more potential to achieve desired improvements with the help of a richer age candidate pool but raise the computational hurdle for implementation. A smaller value of K would impose stronger assumptions that the most beneficial information is hidden among ages with strong similarities only. Our empirical studies in Section 5.3 set K to be 100 and also investigate the marginal gaining predicting accuracy by a few larger values for K.

Similar to the argument in Section 4.2.2.4, a smoothing step is desired to avoid steep ups and downs in the size of $\mathcal{A}^*_{x_0}$ as the target age x_0 changes with the belief that two neighboring target ages should generally use similar age sets in the mortality modeling so that $\mathcal{A}^*_{x_0}$ should have a relatively stable size as the target age x_0 varies from one to a neighbor age. In our empirical studies, the same moving average smoothing scheme as introduced in Chapter 4 with radius $h = 4$ is adopted on the outcomes of the age-specific sequence of k_{x_0}s from the 4-fold BCV results for each population.

We adopted the conventional extrapolating procedure based on time series models where we fit the period sequence in the underlying mortality model by a proper time series model and extrapolate based on the calibrated time series model. Since the Lee-Carter model is chosen as the underlying model for the distance-based approach in the empirical experiment of this chapter, a random walk process with drift (RWD) is used to describe the period effect term.

The above description of the distance-based approach is in the context of borrowing information from multiple populations and thus, we call it the multi-population distance-based approach and the corresponding prediction model multi-population distance-based model. Similarly, the approach can be applied to borrow information from ages within the target population only. In this case, all the K candidate ages in the above-described procedure all from the same population as the target age x_0, we call it the single-population distance-based approach and the corresponding prediction model single-population distance-based model.

5.2.2 Ensemble-based Approach

As one can see later in Section 5.3.1, directly increasing the maximum allowed number of ages K to form a richer age candidate pool to be selected can lead to further improvement in predicting performance but raise the computational hurdle for implementation. To maintain computational feasibility, we propose an ensemble-based approach, where a mild size subset from the pool is selected randomly, and a fixed size of ages from the subset are chosen, based on the same "distance" criterion, to borrow information. With repetition of random subsampling for N times, we obtain N predictions of mortality for the target age, and each is based on one sampled subset. Different averaging strategies from Chapter 3 could then be applied for the formulation of the final prediction. We adopt a simple average scheme in this chapter for the purpose of simplicity in implementation and illustration. There are three hyperparameters to specify in implementing the ensemble method: N represents the repetition time of random subsampling, n represents the size of each randomly chosen

subset, and m represents the pre-specified size of "neighboring" ages in each subset chosen to borrow information. Different combinations of the values of these three hyperparameters are compared in their predicting performances in the subsequent Section 5.3.3. Cross-validation is not needed in the ensemble-based approach and thus it saves a significant amount of computational demands compared to the distance-based method previously introduced. The same extrapolating procedure in Section 5.2.1 is adpoted. A random walk process with drift (RWD) is again used to describe the period effect term.

5.2.3 ACF Model-based Approach

Another factor worth consideration for further improvements in the effectiveness of borrowing information across ages and populations is the choice of the underlying mortality model. In our previous subsection, the Lee-Carter model was embedded in the multi-population distance-based approach, which may not sufficiently reflect the commonality and disparity in mortality development patterns across populations. We thus propose the ACF model-based approach, where an age set selection procedure is implemented before fitting the mortality data with the ACF model for the prediction of a prediction target age.

Clearly, there are a lot of different designs regarding how the age set selection procedure can be done and what grouping over the pool of populations to enforce in fitting the ACF model. In this chapter, we consider the age band based framework as introduced in Chapter 4, and apply the same cross-validation procedure to determine the optimal band size for each target age as the one that yields the smallest validation SSE for the target age itself. In doing so, the Lee-Carter model is replaced by the ACF model fitted to each group of populations from the same geographical location. After the age band is selected, the prediction of mortality prediction for a target age is obtained based on the derived geographical grouping based ACF model by the conventional extrapolation as we did in Chapter 2.

5.3 Empirical Studies

The empirical studies for the performance of the three proposed methods in this section would be conducted on the same 24 populations from the Human Mortality Database (HMD) with the same data splitting scheme as we used in Sections 3.4 and 4.3.

5.3.1 Performance Evaluation for the Distance-based Approach

As we have previously mentioned, the distance-based approach can be applied to borrow information across multiple populations or within the target population. For the approach applied within the target population, the age candidate pool consists of the ages 0-100 from the same target population. In our empirical study, we consider the following two single-population distance-based prediction models:

- **LC-distance**: For each target age x_0, the other ages are ranked with the proposed distance measure, instead of the numerical order of the ages. The optimal age set $\mathcal{A}^*_{x_0}$ is determined as the group of the k "neighboring" ages of the predicting target x_0, where the value of k is then decided by a 4-fold BCV procedure as previously explained. The maximum allowed value of k is set to be 100, which allows every $\mathcal{A}^*_{x_0}$ could be the full-age set containing all ages from 0 to 100.
- **LC-distance-smooth**: Similar to **LC-distance**, but the size of $\mathcal{A}^*_{x_0}$ for each target age x_0 is obtained using the moving averaging smoothing procedure with the length of smoothing window $h = 4$.

For the application of the distance-based approach to borrow information across all 24 populations, the overall age candidate pool consists of all the ages 0-100 from these 24 populations, which amounts to $2424 (= 101 \times 24)$ different "ages". In our empirical study, we consider the following two multi-population distance-based models:

- **Multi-LC-distance**: The distance-based method is applied with an age candidate pool consisting of all ages from all the 24 populations in a total size of 2424 as explained above, and all the candidate ages are ranked by their distances to the target age x_0. The optimal age set $\mathcal{A}^*_{x_0}$ is determined as the group of the k "neighboring" ages of the predicting target x_0. The value of k is also decided with a 4-fold BCV procedure. The maximum allowed number of ages K is set to be 100 with the belief that most beneficial information is hidden among ages with strong similarities.
- **Multi-LC-distance-smooth**: Similar to **Multi-LC-distance**, but the size of $\mathcal{A}^*_{x_0}$ for each target age x_0 is obtained using the moving averaging smoothing procedure with the length of smoothing window $h = 4$.

Similar to the empirical studies in the preceding chapters, the overall prediction accuracy is evaluated in terms of test SSEs for the logarithmic mortality rates over the 24 populations

in the pool and the same set of summary statistics of the resulting 24 overall test SSEs for each predicting model are reported in Table 5.1. Further, the DM test results for pairwise model comparison regarding prediction performance are summarized in Table 5.2.

According to Table 5.1, adopting the distance-based method under the single population scenario has resulted in similar predicting performance compared to the age-band-based method, whenever the smoothing step is applied or not for the size of the selected age set. The comparable performance between the two types of prediction methods can be observed for both female and male populations via these statistics tabulated in Table 5.1. This observation is not surprising because the variance distance of an age close to the target age values tends to be smaller than that of an age farther away from the target age. Therefore, the selected age sets in the final prediction models are similar between these two approaches. Furthermore, a comparison of the last two rows with the other rows in each panel of Table 5.1 indicates that the relative performance between the multi-population distance-based approach and the single-population distance-based approach is not the same for populations of the two different genders. For female populations, the extra data searched in the multiple-population approach does not lead to obvious improvement in prediction accuracy compared with the single-population approach. However, the multiple-population approach does lead to a noticeable improvement in prediction accuracy for male populations as the approach allows the selected age set to include ages from populations other than the one where the target age is located.

We also apply the DM test based procedure to compare the Multi-LC-distance-smooth model with the Lee-Carter, and the LC-ageband-smooth models in their prediction errors. The DM test based procedure was conducted in the same way as we described in Section 2.3.3.2 of Chapter 2 and Section 3.4.2.2 of Chapter 3. The numbers of wins by one model over the other from the DM test procedure are reported in Table 5.2. The results in the table show that the distance-based method outperforms the standard Lee-Carter model over both female and male populations in terms of DM tests. However, its relative performance compared with the age band based method is disparate between the female and male populations. Over female populations, the age band based method beats the distance-based method more times than the other way around (13 versus 5). Over male populations, however, the distance-based method works better than the age band based method (11 versus 5).

We also investigate the effect on the prediction performance from the maximum allowed number K for the age set size in the distance-based approach. We consider three different values for K (100, 200, and 400) and apply the multi-population distance-based method with smoothing to all 24 populations of both genders. The same set of statistics over the

Table 5.1: Summary statistics of test SSEs from the 24 populations comparing the prediction performance between distance-based models, Lee-Carter model, and age band based models from Chapter 4.

	1st Quartile	Median	Mean	3rd Quartile
Female Population				
Lee-Carter	19.37	44.73	44.64	63.90
LC-ageband	17.44	44.49	44.00	67.36
LC-ageband-smooth	14.81	41.94	40.57	59.22
LC-distance	17.13	44.65	44.15	67.17
LC-distance-smooth	16.82	43.67	42.78	65.86
Multi-LC-distance	16.50	41.91	45.78	72.24
Multi-LC-distance-smooth	16.96	40.95	42.14	66.85
Male Population				
Lee-Carter	28.40	41.31	49.79	67.97
LC-ageband	16.78	37.67	39.14	60.66
LC-ageband-smooth	16.01	34.99	37.44	53.90
LC-distance	16.38	38.07	39.25	60.15
LC-distance-smooth	15.55	36.35	37.40	57.05
Multi-LC-distance	15.43	36.09	37.14	57.02
Multi-LC-distance-smooth	14.04	34.68	35.24	53.10

Table 5.2: Number of wins for comparisons between the proposed models (in the column) and the benchmark models (in the row) based on a pairs of one-sided DM tests: In each cell, the first integer indicates the number of wins of the model in the row out of 24 comparisons while the second integer is the number of wins of the model in the column.

	Multi-LC-distance-smooth
Female Population	
Lee-Carter	$(7, 14)$
LC-ageband-smooth	$(13, 5)$
Male Population	
Lee-Carter	$(1, 20)$
LC-ageband-smooth	$(5, 11)$

resulting 24 SSEs from the 24 populations is reported in Table 5.3. The mean of the 24 SSEs shows a decreasing trend as the maximum allowed number of ages K increases. In other words, it seems that further improvement in predicting performance can be achieved by increasing the value of the maximum allowed number of ages K to form a richer age candidate pool to be selected over the cross-validation step in the distance-based approach. The caveat, however, is that such brutal-force implementation of the proposed multi-population distance-based method would become computationally challenging as the increase of the value of the maximum allowed number of ages K results in an explosion in computational demands for all the needed model fitting, cross-validation, and other necessary calculation. This is the reason we consider the ensemble method that we described in Section 5.2.2.

5.3.2 Stylized Facts from the Distance-based Approach

As mentioned earlier, the proposed multi-population distance-based approach would return an age-specific age set $\mathcal{A}^*_{x_0}$ for each specific target age x_0 from a given target population. The optimal age set $\mathcal{A}^*_{x_0}$ contains age $x_1, x_2, \cdots, x_k$, some of which come from the same population as x_0 while the rest do not. For referral convenience, we call those selected ages for a target age as its reference ages. It is interesting to investigate some stylized facts of

Table 5.3: Summary statistics of test SSEs of 24 populations comparing the prediction performance of the distance-based models with different values of the maximum allowed number of ages K.

	1st Quartile	Median	Mean	3rd Quartile
Female Population				
Multi-LC-distance-smooth, $K = 100$	16.96	40.95	42.14	66.85
Multi-LC-distance-smooth, $K = 200$	17.39	40.86	40.58	62.25
Multi-LC-distance-smooth, $K = 400$	17.85	41.07	40.13	59.38
Male Population				
Multi-LC-distance-smooth, $K = 100$	14.04	34.68	35.24	53.10
Multi-LC-distance-smooth, $K = 200$	14.50	33.91	34.79	51.98
Multi-LC-distance-smooth, $K = 400$	15.48	33.60	34.64	50.53

how the reference ages are chosen differently with respect to different target ages x_0.

Firstly, for each target age x_0 from a given target population, we calculate the mean of all chosen reference ages $x_1, x_2, \cdots, x_k$ in the age set $\mathcal{A}^*_{x_0}$, denoted as $\bar{x}_i$, then the absolute value of the difference between $\bar{x}_i$ and x_0 is calculated. We do this for every x_0 of the 24 target populations, and then take an average from these 24 target populations. The left panel of Figure 5.1 demonstrates this averaged absolute distance with respect to different values of the target ages x_0 to present the average discrepancy between the chosen reference ages and the target. A small value indicates the chosen reference ages do not differ much from the target age on their average value while a large value indicates the existence of a noticeable discrepancy between the chosen reference ages and the target age. As the left panel of Figure 5.1 reveals, the discrepancy seems prominent when the target age x_0 is smaller than 30 or close to 100 but unapparent for the middle ages. The results validate that the selected ages spread out from the target age more when a young or an elder age is the target for mortality prediction than an age from other bands.

Secondly, for each target age x_0 from a given target population, the standard deviation of all chosen reference ages $x_1, x_2, \cdots, x_k$ in the age set $\mathcal{A}^*_{x_0}$ is calculated. The age-specific standard deviations are then averaged over all 24 target populations and plotted in the right

panel of Figure 5.1, to indicate how volatile the chosen reference ages are with different values of the target ages x_0. A low standard deviation indicates that the chosen reference ages do not differ much from each other while a high standard deviation indicates that chosen reference ages are spread out over a wider range. With an observable decreasing trend for both genders as the value of the target ages x_0 increases, the figure confirms that younger ages would generally borrow information from ages with a wider range while the reference ages for adult/retiree ages are chosen to be more concentrated. Meanwhile, an increase of the standard deviation at the right end (when the target age x_0 is large) is also observed in both genders, providing evidence of the very old ages borrow information from ages widely in the distance-based approach.

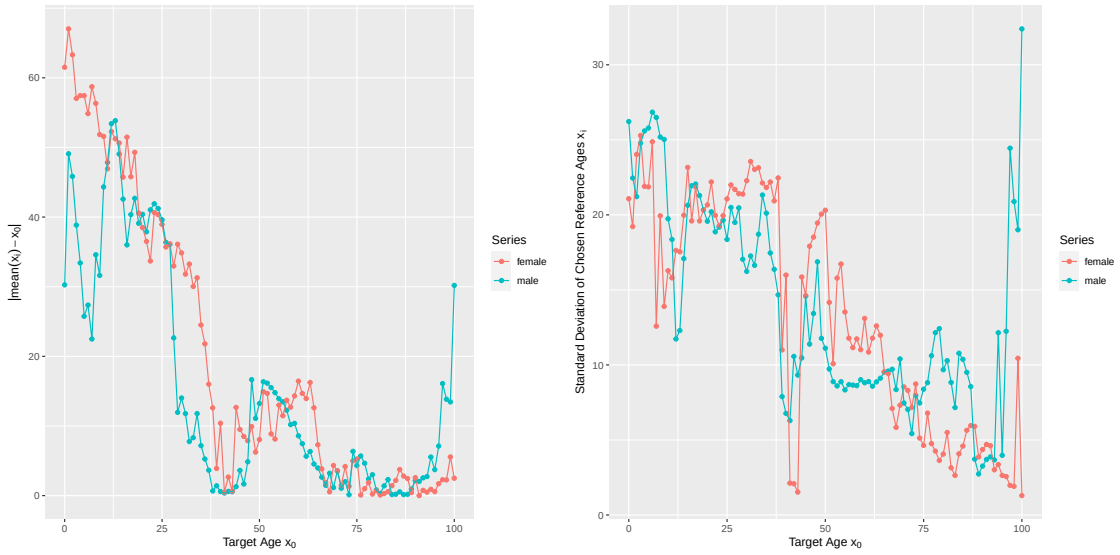


Figure 5.1: Stylized facts of chosen reference ages x_i with respect to different target ages x_0. Left: absolute difference between the averaged chosen reference ages $\bar{x}_i$ and the target age x_0. Right: standard deviation of chosen reference ages x_i with respect to different target ages x_0.

It is also interesting to have a closer look at the composition of chosen ages over different target ages x_0 to examine the difference between the number of domestic ages (selected reference ages from the same target population) and the number of foreign ages (selected reference ages from other reference populations). The averaged number of the selected domestic ages and that of the selected foreign ages (i.e., those contained in set $\mathcal{A}^*_{x_0}$) are illustrated, respectively, by the two curves in Figure 5.2 to show how they vary over the target age x_0. The curves show a similar trend over both female and male populations. Roughly speaking, the number of selected domestic ages shows a first increasing and then decreasing trend while the number of selected foreign ages presents a first decreasing and

then increasing trend as the value of the target age x_0 increases. In other words, for both female and male populations, younger and older ages tend to borrow information from more ages out of the population where the target age is located, while those middle ages tend to use more ages from the domestic population in our distance-based approach.

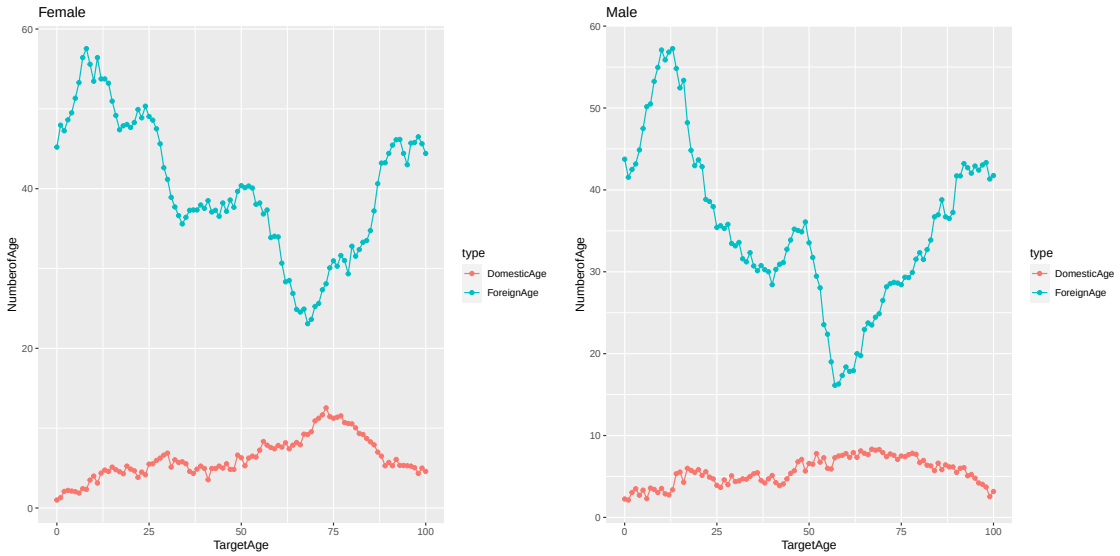


Figure 5.2: Number of selected ages: Domestic versus Foreign.

Furthermore, as every target age of every target population selects its age-specific age set among all the ages of all the 24 populations in the distance-based approach, it is interesting to explore what relative "role" that one specific age/population has undertaken in the system of the information "flow". For a specific target age x_0 from a specific population i, the optimal age set $\mathcal{A}^*_{x_0}$ consists of age $x_1, x_2, \cdots, x_k$, some of which come from the same population i while the rest do not. Regardless of the population they come from, the numbers of times of age 0 to 100 being selected as the members of $\mathcal{A}^*_{x_0}$ for target age x_0 are recorded. The results are then aggregated among all 24 populations with respect to different values of x_0 and demonstrated in Figure 5.3 as a concise summary for the general status of the difference in selection frequency of different ages. The thickness of the arrows represents the differences in selection frequency of different ages. A thicker arrow means the age represented by the starting point of the arrow has been chosen more frequently by the age represented by the ending point of the arrow in the distance-based method, while the thinner arrow means the selection happens less frequently. Moreover, the overall layout of the arrows can provide insights into the difference in the relative "role" that one specific age has undertaken in the system of the information "flow". A large amount of arrows inflow to the point would indicate the age represented by the point majorly serves as the role to "borrow" information from others, while a large amount of arrows outflow from the point

would indicate the age represented by the point are "providing" information to others.

As Figure 5.3 shows, for both female and male populations, there exists a large amount of information inflow when a very young or very old age, usually less than 10 or above 80, is fixed as the target age. It is very frequent to include information that comes from the same very young or very old age group. This observation has again confirmed that young and old ages borrow information from more ages compared with a target age in other bands.

Similarly, the numbers of ages from each of the 24 populations being selected as the members of $\mathcal{A}^*_{x_0}$ for target age x_0 of a given target population are recorded, regardless of the specific values of the chosen ages. The results are then aggregated among all target ages within the same target population. Figure 5.4 demonstrates how frequently each target population with an age selects an age from another population. The thickness of the arrows represents the selection frequency of the population at the starting point by the population at the ending point.

As Figure 5.4 shows, there exist populations that mainly borrow information from ages within themselves, like the US and Japan, and populations that borrow from other populations more, like Austria, Scotland, and New Zealand for both genders. Further population-specific results can also be obtained from Figure 5.4. For instance, the US female mainly borrows information from ages within itself and meanwhile provides information to females in Finland, Netherlands, Canada, and Sweden. Another illustrative example is the France male, which receives information mainly from the US male and itself and provides information for males in Belgium, Switzerland, Austria, and Italy.

5.3.3 Empirical Evaluations over Different Approaches

In this section, we implement the three different approaches described in Section 5.2 to all 24 populations for both genders separately and compare their relative performance using the resulting test SSEs for the logarithmic mortality rate with their summary statistics reported in Table 5.4. In the ensemble-based approach, we exploit hyperparameters $(n, m, N) = (150, 100, 100)$. The results confirm a similar predicting accuracy between the ensemble-based method and the multi-population distance-based method. Furthermore, the ACF model-based method has a noticeable improvement in terms of predicting accuracy over the other two approaches when female populations are considered for mortality prediction. However, it does not lead to an improvement for male populations. There was indeed a slight deterioration in predicting accuracy with the ACF-based approach over male populations.

Finally, we also experimented to check how the triplet hyperparameters (n, m, N) may

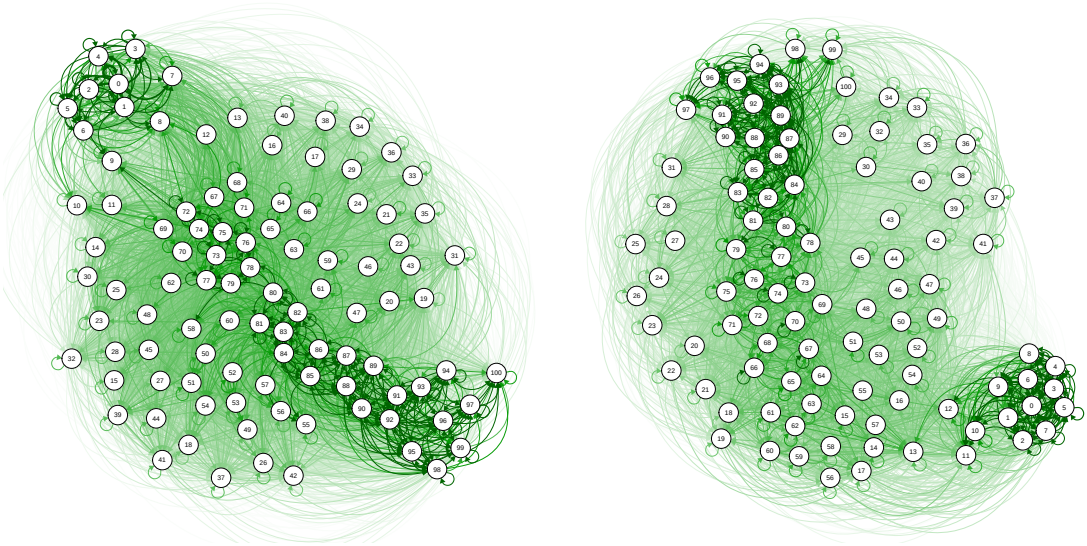


Figure 5.3: Visualized summary of relations between different ages. A thicker arrow means the age represented by the starting point of the arrow has been chosen more frequently in the age-specific age set for the mortality prediction of the age represented by the ending point of the arrow in the distance-based method while the thinner arrow means the selection happens less frequently. The left panel demonstrates results for female populations and the right panel demonstrates results for male populations.

affect the predicting performance of the ensemble-based approach. We reported the prediction summary results in Table 5.5 for different combinations of the three parameters, tabulated separately by gender. The results in the table indicate the predicting performance of the ensemble-based approach is rather robust when m is set at 100, a number almost the same as the number of full age range considered in a single-population method. Different choices of parameters N and n yield similar magnitudes of SSEs.

Extra combinations of hyperparameter values in the ensemble-based approach were also considered with the experimental outcomes demonstrated in Figure 5.5, where each curve shows how the resulting average SSE from all the 24 populations responds to the value of one or two parameters when the other parameter(s) is fixed. Figure 5.5, we have the following observations:

- The upper left panel shows how the average SSE varies over different values for n, the size of subsamples in the ensemble-based approach when the other two parameters are fixed with $m = 100$ and $N = 100$. The curve in the panel indicates that an overly large value for the subsampling size is detrimental to the resulting predicting accuracy. One interpretation of this observation is that a larger value of n tends to have more

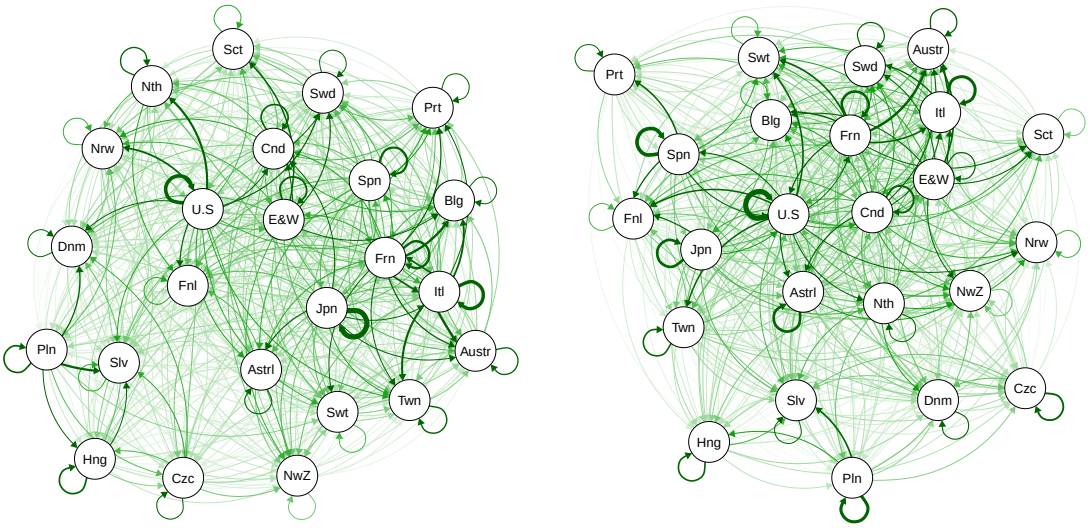


Figure 5.4: Visualized summary of relations between different populations, represented by different points. A thicker arrow means the population represented by the starting point of the arrow has been chosen more frequently to help predict the population represented by the ending point of the arrow in the distance-based method while the thinner arrow means the selection happens less frequently. The left panel demonstrates female results while the right panel demonstrates male results.

overlapping across the subsampled age set for ranking and this will lead to a more similar predicting rule from each subsampling in the ensemble-based method while maintaining relatively independence among predictions from different subsamples is essential to ensure the efficiency of a general ensemble prediction method.

- The upper right panel gives changing trend of the average SSE over different values of m when the other two parameters are fixed at $n = 200$ and $N = 100$. The curve in the panel shows that the best predicting accuracy can be obtained by taking the size of the selected age set the same as the size of each subsample in the ensemble-based approach.

- The experiment for results in the lower left panel is motivated by our observation from the upper right panel of the figure which shows that the best predicting accuracy was reached when $m = n$. The experiment results in the lower left panel indicate that the best choice is roughly between 100 and 200 for m and n when they are set equal to each other in the ensemble-based method.

- The lower right panel of Figure 5.5 gives the curve of the average SSE versus the value of N, the number of subsampling in the ensemble-based approach. As one can expect

from a general understanding of an ensemble method, an increase in the subsampling time would not detrimental to the resulting predicting performance. The curves in the panel indicate that the marginal benefit of increasing the subsampling time is rather tenuous.

Table 5.4: Summary statistics of test SSEs of the 24 populations comparing the prediction performance of different proposed strategies.

	1st Quartile	Median	Mean	3rd Quartile
Female Population				
Multi-LC-distance-smooth, $K = 100$	16.96	40.95	42.14	66.85
Multi-LC-distance-smooth, $K = 200$	17.39	40.86	40.58	62.25
Multi-LC-distance-smooth, $K = 400$	17.85	41.07	40.13	59.38
Multi-LC-ensemble $(150, 100, 100)$	18.67	40.71	41.04	61.40
ageband-ACFGeoInfo	15.40	38.87	37.23	54.29
Male Population				
Multi-LC-distance-smooth, $K = 100$	14.04	34.68	35.24	53.10
Multi-LC-distance-smooth, $K = 200$	14.50	33.91	34.79	51.98
Multi-LC-distance-smooth, $K = 400$	15.48	33.60	34.64	50.53
Multi-LC-ensemble $(150, 100, 100)$	15.41	33.80	34.66	51.76
ageband-ACFGeoInfo	19.23	33.11	36.49	49.71

5.4 Concluding Remarks

In this chapter, the task of improving future mortality prediction accuracy is incorporated with a more comprehensive setting where information across different ages and different populations is considered simultaneously. This chapter proposes three approaches: a distance-based approach as an extension of the age-specific age set paradigm in Chapter 4 to the multi-population scenario with a specially designed "distance" measure, an ensemble-based

Table 5.5: Summary statistics of test SSEs of the 24 populations comparing the prediction performance of ensemble-based approach with different values of hyperparameters, where N is the number of subsampling, n is the size of subset subsampled each time, and m is the size of the chosen age set using the distance-based method from each subsampled age set.

Multi-LC-ensemble (n, m, N)	1st Quartile	Median	Mean	3rd Quartile
Female Population				
Multi-LC-ensemble $(125, 100, 100)$	18.51	39.33	40.98	62.01
Multi-LC-ensemble $(150, 100, 100)$	18.67	40.71	41.04	61.40
Multi-LC-ensemble $(175, 100, 100)$	18.80	42.71	41.22	61.08
Multi-LC-ensemble $(200, 100, 100)$	18.88	43.88	41.40	60.80
Multi-LC-ensemble $(200, 150, 100)$	18.91	43.99	41.32	60.32
Multi-LC-ensemble $(200, 100, 50)$	18.87	43.91	41.42	60.80
Multi-LC-ensemble $(200, 100, 200)$	18.88	43.90	41.40	60.78
Multi-LC-ensemble $(200, 100, 300)$	18.89	43.91	41.40	60.79
Multi-LC-ensemble $(200, 100, 400)$	18.88	43.90	41.41	60.78
Multi-LC-ensemble $(200, 100, 500)$	18.88	43.90	41.41	60.79
Male Population				
Multi-LC-ensemble $(125, 100, 100)$	15.09	33.86	34.69	52.04
Multi-LC-ensemble $(150, 100, 100)$	15.41	33.80	34.66	51.76
Multi-LC-ensemble $(175, 100, 100)$	15.79	33.62	34.68	51.77
Multi-LC-ensemble $(200, 100, 100)$	17.85	32.68	35.81	51.60
Multi-LC-ensemble $(200, 150, 100)$	16.17	33.24	34.65	51.19
Multi-LC-ensemble $(200, 100, 50)$	16.03	33.30	34.75	51.64
Multi-LC-ensemble $(200, 100, 200)$	15.99	33.34	34.73	51.72
Multi-LC-ensemble $(200, 100, 300)$	15.98	33.34	34.72	51.66
Multi-LC-ensemble $(200, 100, 400)$	15.98	33.32	34.72	51.69
Multi-LC-ensemble $(200, 100, 500)$	15.99	33.29	34.72	51.72

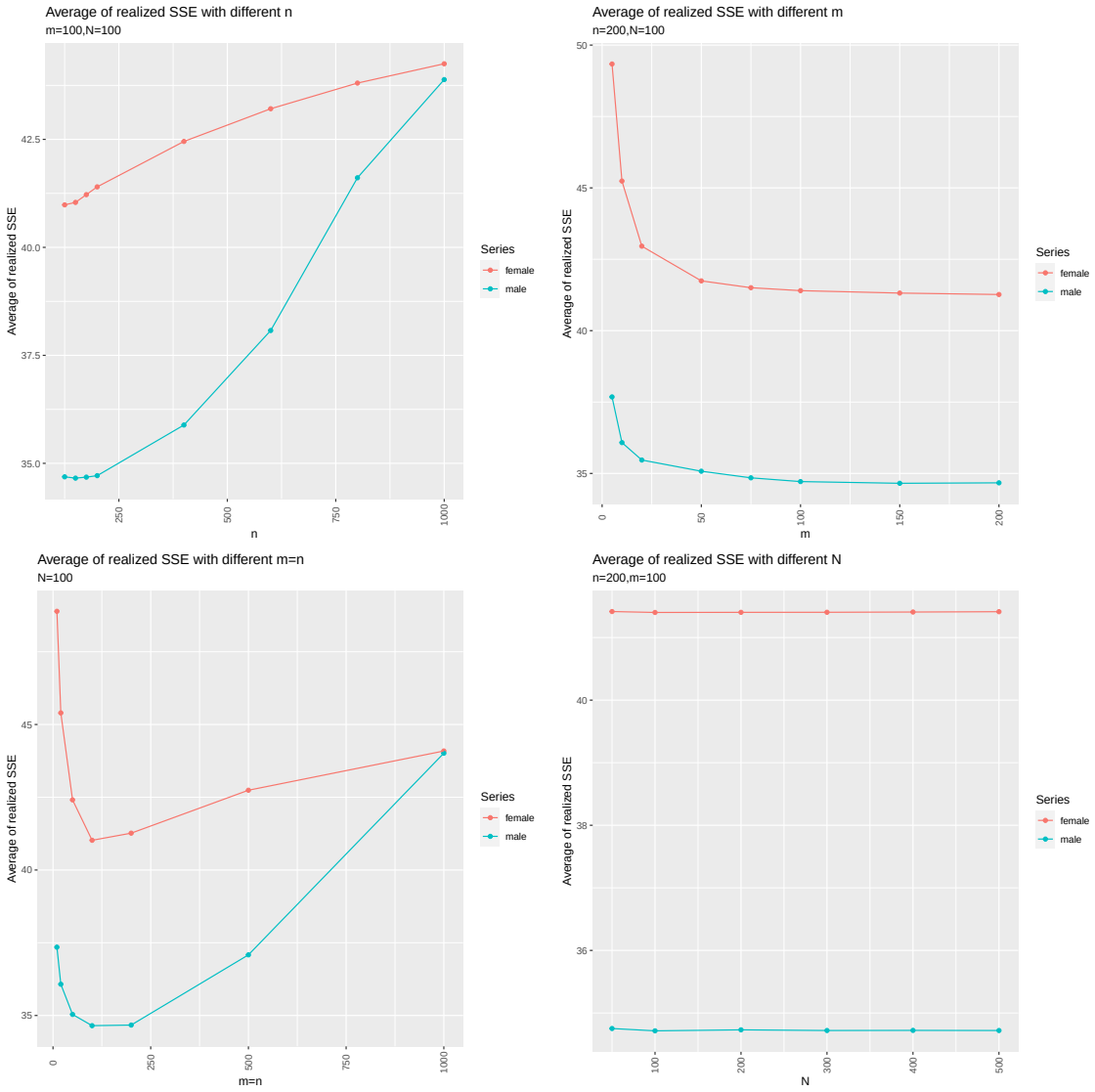


Figure 5.5: Average of realized test SSEs for ensemble-based method with different values of hyperparameters. Topleft: average of realized test SSEs for ensemble-based method with different values of n and fixed m and N. Topright: average of realized test SSEs for ensemble-based method with different values of m and fixed values of n and N. Bottomleft: average of realized test SSEs for ensemble-based method with different values of $m = n$ and fixed N. Bottomright: average of realized test SSEs for ensemble-based method with different values of N and fixed m and n.

approach that averages results from multiple random subsets, and an ACF model-based approach that combines the age band based framework with the ACF model using exogenous geographic information.

Thorough comparisons among the proposed approaches are conducted based on extensive empirical studies. The results of numerical studies with the Human Mortality Database (HMD) have confirmed the potential capability to consider borrowing information among different ages from different populations with a certain level of improvement in predicting accuracy for all three approaches. The multiple-population distance-based approach has led to a noticeable improvement in prediction accuracy for male populations. The ensemble-based method yields results similar to the multi-population distance-based method in predicting accuracy with much less computational demand. Another intriguing property of the ensemble-based method is its capacity to obtain robust results with a range of values of the triplet hyperparameters. As for the ACF model-based approach, it has been suggested by the empirical results that a more complex model, e.g., the ACF model is more appropriate for female populations while a simpler Lee-Carter model works better for male populations. Additionally, several general stylized facts of how ages from multiple populations are borrowed by the distance-based method are provided.

Chapter 6

Future Works

This thesis studies the problem of borrowing helpful information from multiple aspects to enhance human mortality predicting accuracy using prudently designed statistical learning approaches. By the momentum of the research outcomes in this thesis, there are many directions worth further exploration for future research, and this chapter aims to describe some of these in two separate sections. Section 6.1 states some specific aspects that are worth further exploration following up the research in each of Chapters 2-5. Section 6.2 proposes an imputation framework for forecasting any interesting quantities associated with the future liabilities of life insurance products.

6.1 Follow-up Research from Each Chapter

Chapter 2 provides a DSA algorithm based solution for the problem of selecting a "proper" group of populations to ensure the success of multi-population mortality modeling. There are several directions to further the application of the DSA based framework for mortality prediction. Firstly, because of the flexibility in its design, there is enough room to adjust elements in the DSA algorithm, such as the risk function, the threshold, and even the order of the three moves (Deletion, Substitution, Addition), to consider a broader scope of usage. Secondly, while the DSA algorithm is applied with the extended ACF model in Chapter 2, it is interesting to investigate its performance along with other multi-population models such as the CBD model proposed by Cairns et al. [2006] or the common sparse age-period model introduced by Hatzopoulos and Haberman [2013]. Finally, other classic machine learning methods have also been introduced into the area of mortality forecasting recently. For example, Richman and Wüthrich [2021] extended the Lee-Carter model to multiple populations by using neural networks for the automatic selection of optimal model structure,

and Nigri et al. [2019] applied a recurrent neural network with a long short-term memory architecture to the Lee-Carter model for improved predictive capacity. A comparison or combination of the DSA-based procedure and other machine learning techniques can be potentially effective in mortality forecasting.

Chapter 3 designs an alternative framework that addresses major potential issues that arise with direct usage of a high-dimensional multi-population model. The proposed bivariate model based ensemble (BMBE) framework can ease computational hurdles and allow extra flexibility in the structure of the underlining base learner mortality model. There are several aspects worth further exploration of the BMBE framework. Firstly, while the bivariate ACF and bivariate CBD models were used as the base learners in Chapter 3, other bivariate-population models are worth consideration in the BMBE framework, for example, the gravity model in Dowd et al. [2011], or the VAR and VECM models introduced by Zhou et al. [2014]. Secondly, although the BMBE framework does not impose any constraints on the base learner, a base learner that captures more desirable characteristics of the mortality data is conducive to improving the prediction accuracy of the general framework. It could be rewarding to consider a set of base learner candidates (instead of a fixed base learner for every population pair as we do in the thesis) and apply a data-driven model selection procedure for the fit of each population pair before ensembling base learners for the final forecast. An automated procedure of identifying a suitable base learner for each population pair is imperative if a large set of base learner candidates would be implemented, particularly when the pool of auxiliary populations is also large. Finally, as another exciting extension, it would be interesting to investigate further the design of base learner candidates. We may consider other model features in addition to time shift and cohort effect that have been investigated in Chapter 3. Furthermore, since the strength of the cohort effect varies from population to population, it could also be interesting to consider a base learner candidate in the set of candidate pools with an "asymmetric" structure where a cohort effect term is included for one population but not for the other.

In chapter 4, the idea of borrowing information has been altered to a different scope in which similarities among different ages are detected and considered as a source of information for future predicting accuracy enhancement. A novel age-specific age set framework is introduced to decompose the overall predicting goal into multiple individual tasks and search for individual age bands to ensure the mortality prediction of each target age can receive the benefit of borrowing information across ages to the largest extent. Furthermore, Chapter 5 approaches the problem at a more comprehensive level where the selection of both ages and populations are considered jointly. Three different approaches are proposed in this chapter, including one distance-based approach, an ensemble approach (based on

the distance-based approach), and an ACF-based approach. As follow-up research, it would be interesting to explore the performance of these approaches by altering some elements in their designs. For example, different types of ensemble methods and different specifications of the ACF model can be considered. Moreover, a challenging yet promising venture is to extend the DSA-based procedure or the BMBE framework to select not only population groups but also time periods and age bands in a computational-friendly way so that information from multiple aspects can be collected in a more effective way to ensure a further improvement in mortality prediction.

6.2 Imputation Method for Forecasting Future Liabilities of Life Insurance Products

We propose an imputation method for forecasting any interesting quantities associated with the future liability of life insurance products. To convey the idea, let us use the n-year term life annuity immediate as an example and consider forecasting a quantile of its future liability.

Let $a_{x,t}(n)$ denote the time-t pure value of an n-year term life annuity. It can be computed from the individual death rates through:

$$a_{x,t}(n) = \sum_{k=1}^{n} v(k) s_{x,t}(k), \tag{6.1}$$

where $s_{x,t}(k)$, denoting the k-year survival probability for an individual aged x in year t, is calculated as

$$s_{x,t}(k) = \prod_{j=0}^{k-1} p_{x+j,t+j} = \prod_{j=0}^{k-1} (1 - q_{x+j,t+j})$$

with $p_{x,t} = s_{x,t}(1)$ and $q_{x,t} = 1 - p_{x,t}$. Assume we have historical mortality rates data:

$$q_{x,t}, \text{ for } x = x_0, x_0 + 1, \ldots, x_w, \text{ and } t = 1, 2, \ldots, T,$$

where x_w represents the age limit so that $q_{x_\omega,t} = 1$ for every t. These historical data enable us to compute realized values of $a_{x,t}(n)$ up to time $T - n + 1$. One conventional way to estimate the risk measure Value at Risk of

$$a_{x,t}(n) \text{ for } t = T + 1, T + 2, \ldots, ...$$

is to establish a dynamic mortality model for $q_{x,t}$ and then apply a simulation procedure to get an estimate of the risk measure. However, this existing method calls for a full

specification of the mortality model, and any violation of the model may lead to a large estimation error for the risk measure. Instead, we plan to consider a quantile factor model directly on $a_{x,t}(n)$ to address the problem of estimation. The historical data enable use to compute the realized data for $a_{x,t}$ for the age-year window $\{x_0, x_0+1, \cdots, x_w\} \times = \{1, 2, \cdots, T-n+1\}$ without access to $a_{x,t}(n)$ from $T-n+2$ to T because they rely on future morality rates. An imputation procedure is then adopted for the values of $a_{x,t}(n)$ from year $T-n+2$ to year T:

$$\begin{array}{cccc} a^*_{x_0,T-n+2}(n), & a^*_{x_0,T-n+3}(n), & \ldots, & a^*_{x_0,T}(n) \\ a^*_{x_0+1,T-n+2}(n), & a^*_{x_0+1,T-n+3}(n), & \ldots, & a^*_{x_0+1,T}(n) \\ \vdots & \vdots & \ddots & \vdots \\ a^*_{x_w,T-n+2}(n), & a^*_{x_w,T-n+3}(n), & \ldots, & a^*_{x_w,T}(n) \end{array} \tag{6.2}$$

Once the imputation finishes, we apply a quantile-based factor model to the whole set of data, including both the realized data and the imputed data of $a_{x,t}(n)$. Comparing the estimate of VaR using the conventional simulation method and the newly proposed quantile-based method, the simulation method should work reasonably well when there are no misspecifications for the underlining mortality model. Under scenarios when misspecifications appear, the proposed quantile-based method is expected to provide a more robust estimate for future Value at Risk. The design of the specific form of the quantile model remains an interesting topic, and we expect that the proposed method would throw light on the further usage of quantile-based methods in mortality prediction and risk management.

References

David Atance, Ana Debón, and Eliseo Navarro. A comparison of forecasting mortality models using resampling methods. *Mathematics*, 8(9):1550, 2020.

Jushan Bai and Serena Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221, 2002.

Magali Barbieri, John R Wilmoth, Vladimir M Shkolnikov, Dana Glei, Domantas Jasilionis, Dmitri Jdanov, Carl Boe, Timothy Riffe, Pavel Grigoriev, and Celeste Winant. Data resource profile: the human mortality database (HMD). *International Journal of Epidemiology*, 44(5):1549–1556, 2015.

Christoph Bergmeir and José M Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

Christoph Bergmeir, Mauro Costantini, and José M Benítez. On the usefulness of cross-validation for directional forecast evaluation. *Computational Statistics & Data Analysis*, 76:132–143, 2014.

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

Heather Booth, John Maindonald, and Len Smith. Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56(3):325–336, 2002.

George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2015.

Elizabeth Brainerd and David M Cutler. Autopsy on an empire: understanding mortality in Russia and the former Soviet Union. *Journal of Economic Perspectives*, 19(1):107–130, 2005.

Natacha Brouhns, Michel Denuit, and Jeroen K Vermunt. A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3):373–393, 2002a.

Natacha Brouhns, Michel Denuit, Jeroen K Vermunt, et al. Measuring the longevity risk in mortality projections. *Bulletin of the Swiss Association of Actuaries*, 2(1):105–30, 2002b.

Natacha Brouhns, Michel Denuit*, and Ingrid Van Keilegom. Bootstrapping the Poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal*, 2005(3): 212–224, 2005.

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

Andrew JG Cairns, David Blake, and Kevin Dowd. Modelling and management of mortality risk: a review. *Scandinavian Actuarial Journal*, 2008(2-3):79–113, 2008.

Andrew JG Cairns, David Blake, Kevin Dowd, Guy D Coughlan, David Epstein, Alen Ong, and Igor Balevich. 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, 2009.

Andrew JG Cairns, David Blake, Kevin Dowd, Guy D Coughlan, David Epstein, and Marwa Khalaf-Allah. Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, 48(3):355–367, 2011a.

Andrew JG Cairns, David Blake, Kevin Dowd, Guy D Coughlan, and Marwa Khalaf-Allah. Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin*, 41(1): 29–59, 2011b.

Le Chang and Yanlin Shi. Age-coherent mortality modeling and forecasting using a constrained sparse vector-autoregressive model. *North American Actuarial Journal*, pages 1–19, 2022.

Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. Nbclust: an R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(1):1–36, 2014.

Liang Chen, Juan J Dolado, and Jesús Gonzalo. Quantile factor models. *Econometrica*, 89 (2):875–910, 2021.

Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2008.

Edviges Coelho and Luis C Nunes. Forecasting mortality in the event of a structural change. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(3):713–736, 2011.

Michael AA Cox and Trevor F Cox. Multidimensional scaling. In *Handbook of Data Visualization*, pages 315–347. Springer, 2008.

Iain D Currie, Maria Durban, and Paul HC Eilers. Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279–298, 2004.

Claudia Czado, Antoine Delwarde, and Michel Denuit. Bayesian Poisson log-bilinear mortality projections. *Insurance: Mathematics and Economics*, 36(3):260–284, 2005.

Ivan Luciano Danesi, Steven Haberman, and Pietro Millossovich. Forecasting mortality in subpopulations using Lee–Carter type models: A comparison. *Insurance: Mathematics and Economics*, 62:151–161, 2015.

Piet De Jong and Leonie Tickle. Extending Lee–Carter mortality forecasting. *Mathematical Population Studies*, 13(1):1–18, 2006.

Philippe Deprez, Pavel V Shevchenko, and Mario V Wüthrich. Machine learning techniques for mortality modeling. *European Actuarial Journal*, 7(2):337–352, 2017.

Liqun Diao, Yechao Meng, and Chengguo Weng. A DSA algorithm for mortality forecasting. *North American Actuarial Journal*, 25(3):438–458, 2021.

Francis X Diebold and Roberto S Mariano. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3):134–144, 1995.

Kevin Dowd, Andrew JG Cairns, David Blake, Guy D Coughlan, and Marwa Khalaf-Allah. A gravity model of mortality rates for two related populations. *North American Actuarial Journal*, 15(2):334–356, 2011.

Vasil Enchev, Torsten Kleinow, and Andrew JG Cairns. Multi-population mortality models: fitting, forecasting and comparisons. *Scandinavian Actuarial Journal*, 2017(4):319–342, 2017.

Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

Man Chung Fung, Gareth W Peters, and Pavel V Shevchenko. A unified approach to mortality modelling using state-space framework: characterisation, identification, estimation and forecasting. *Annals of Actuarial Science*, 11(2):343–389, 2017.

Man Chung Fung, Gareth W Peters, and Pavel V Shevchenko. Cohort effects in mortality modelling: a Bayesian state-space approach. *Annals of Actuarial Science*, 13(1):109–144, 2019.

Guangyuan Gao and Yanlin Shi. Age-coherent extensions of the Lee–Carter model. *Scandinavian Actuarial Journal*, 2021(10):998–1016, 2021.

Benjamin Gompertz. XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. in a letter to Francis Baily, Esq. FRS &c. *Philosophical Transactions of the Royal Society of London*, (115):513–583, 1825.

Quentin Guibert, Olivier Lopez, and Pierrick Piette. Forecasting mortality rate improvements with a high-dimensional VAR. *Insurance: Mathematics and Economics*, 88:255–272, 2019.

Steven Haberman and Arthur Renshaw. On age-period-cohort parametric mortality rate projections. *Insurance: Mathematics and Economics*, 45(2):255–270, 2009.

Steven Haberman and Arthur Renshaw. A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48(1):35–55, 2011.

Bruce E Hansen and Jeffrey S Racine. Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46, 2012.

Peter R Hansen, Asger Lunde, and James M Nason. The model confidence set. *Econometrica*, 79(2):453–497, 2011.

David Harvey, Stephen Leybourne, and Paul Newbold. Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, 13(2):281–291, 1997.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, volume 2. Springer, 2009.

Petros Hatzopoulos and Steven Haberman. A parameterized approach to modeling and forecasting mortality. *Insurance: Mathematics and Economics*, 44(1):103–123, 2009.

Petros Hatzopoulos and Steven Haberman. A dynamic parameterization modeling for the age–period–cohort mortality. *Insurance: Mathematics and Economics*, 49(2):155–174, 2011.

Petros Hatzopoulos and Steven Haberman. Common mortality modeling and coherent forecasts. an empirical analysis of worldwide mortality data. *Insurance: Mathematics and Economics*, 52(2):320–337, 2013.

Lingyu He, Fei Huang, Jianjie Shi, and Yanrong Yang. Mortality forecasting using factor models: Time-varying or time-invariant factor loadings? *Insurance: Mathematics and Economics*, 98:14–34, 2021.

Andrew Hunt and David Blake. On the structure and classification of mortality models. *North American Actuarial Journal*, pages 1–20, 2020.

Andrew Hunt and David Blake. On the structure and classification of mortality models. *North American Actuarial Journal*, 25(sup1):S215–S234, 2021.

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

Maintainer Rob J Hyndman, Heather Booth, Leonie Tickle, and John Maindonald. Package 'demography', 2019.

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

Rob J Hyndman, Yeasmin Khandakar, et al. *Automatic time series for forecasting: the forecast package for R.* Number 6/07. Monash University, Department of Econometrics and Business Statistics, 2007.

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

Fanny Janssen, Leo JG van Wissen, and Anton E Kunst. Including the smoking epidemic in internationally coherent mortality projections. *Demography*, 50(4):1341–1362, 2013.

Salvatory R Kessy, Michael Sherris, Andrés M Villegas, and Jonathan Ziveyi. Mortality forecasting using stacked regression ensembles. *Scandinavian Actuarial Journal*, pages 1–36, 2021.

Torsten Kleinow. A common age effect model for the mortality of multiple populations. *Insurance: Mathematics and Economics*, 63:147–152, 2015.

Alois Kneip and Joachim Engel. Model estimation in nonlinear regression under shape invariance. *The Annals of Statistics*, 23(2):551–570, 1995.

Roger Koenker and Beum J Park. An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1-2):265–283, 1996.

Atsuyuki Kogure and Yoshiyuki Kurachi. A Bayesian approach to pricing longevity risk based on risk-neutral predictive distributions. *Insurance: Mathematics and Economics*, 46(1):162–172, 2010.

Marie-Claire Koissi, Arnold F Shapiro, and Göran Högnäs. Evaluating and extending the Lee–Carter model for mortality forecasting: Bootstrap confidence interval. *Insurance: Mathematics and Economics*, 38(1):1–20, 2006.

Vasilis Kontis, James E Bennett, Colin D Mathers, Guangquan Li, Kyle Foreman, and Majid Ezzati. Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble. *The Lancet*, 389(10076):1323–1335, 2017.

Ka Kin Lam and Bo Wang. Robust non-parametric mortality and fertility modelling and forecasting: Gaussian process regression approaches. *Forecasting*, 3(1):207–227, 2021.

WH Lawton, EA Sylvestre, and MS Maggio. Self modeling nonlinear regression. *Technometrics*, 14(3):513–532, 1972.

Ronald Lee and Timothy Miller. Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography*, 38(4):537–549, 2001.

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

Susanna Levantesi, Andrea Nigri, and Gabriella Piscopo. Clustering-based simultaneous forecasting of life expectancy time series through long-short term memory neural networks. *International Journal of Approximate Reasoning*, 140:282–297, 2022.

Han Li and Colin O'Hare. Semi-parametric extensions of the Cairns–Blake–Dowd model: A one-dimensional kernel smoothing approach. *Insurance: Mathematics and Economics*, 77:166–176, 2017.

Han Li, Colin O'Hare, and Xibin Zhang. A semiparametric panel approach to mortality modeling. *Insurance: Mathematics and Economics*, 61:264–270, 2015a.

Han Li, Colin O'hare, and Farshid Vahid. Two-dimensional kernel smoothing of mortality surface: An evaluation of cohort strength. *Journal of Forecasting*, 35(6):553–563, 2016.

Hong Li and Yang Lu. Coherent forecasting of mortality rates: A sparse vector-autoregression approach. *ASTIN Bulletin*, 47(2):563–600, 2017.

Hong Li and Yanlin Shi. Forecasting mortality with international linkages: A global vector-autoregression approach. *Insurance: Mathematics and Economics*, 100:59–75, 2021.

Jackie Li. A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. *Population Studies*, 67(1):111–126, 2013.

Jackie Li. An application of MCMC simulation in mortality projection for populations with limited data. *Demographic Research*, 30:1–48, 2014.

Johnny Siu-Hang Li and Mary R Hardy. Measuring basis risk in longevity hedges. *North American Actuarial Journal*, 15(2):177–200, 2011.

Johnny Siu-Hang Li, Mary R Hardy, and Ken Seng Tan. Uncertainty in mortality forecasting: an extension to the classical Lee-Carter approach. *ASTIN Bulletin*, 39(1):137–164, 2009.

Johnny Siu-Hang Li, Wai-Sum Chan, and Siu-Hung Cheung. Structural changes in the Lee-Carter mortality indexes: detection and implications. *North American Actuarial Journal*, 15(1):13–31, 2011.

Johnny Siu-Hang Li, Rui Zhou, and Mary Hardy. A step-by-step guide to building two-population stochastic mortality models. *Insurance: Mathematics and Economics*, 63: 121–134, 2015b.

Johnny Siu-Hang Li, Wai-Sum Chan, and Rui Zhou. Semicoherent multipopulation mortality modeling: the impact on longevity risk securitization. *Journal of Risk and Insurance*, 84(3):1025–1065, 2017.

Nan Li and Ronald Lee. Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42(3):575–594, 2005.

Nan Li, Ronald Lee, and Patrick Gerland. Extending the Lee-Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography*, 50 (6):2037–2051, 2013.

Tzuling Lin and Cary Chi-Liang Tsai. Hierarchical Bayesian modeling of multi-country mortality rates. *Scandinavian Actuarial Journal*, 2022(5):375–398, 2022.

Andreas Milidonis and Maria Efthymiou. Mortality leads and lags. *Journal of Risk and Insurance*, 84(S1):495–514, 2017.

Andreas Milidonis, Yijia Lin, and Samuel H Cox. Mortality regimes and pricing. *North American Actuarial Journal*, 15(2):266–289, 2011.

Annette M Molinaro, Karen Lostritto, and Mark Van Der Laan. partDSA: deletion/substitution/addition algorithm for partitioning the covariate space in prediction. *Bioinformatics*, 26(10):1357–1363, 2010.

Andrea Nigri, Susanna Levantesi, Mario Marino, Salvatore Scognamiglio, and Francesca Perla. A deep learning integrated Lee–Carter model. *Risks*, 7(1):33, 2019.

Colin O'Hare and Youwei Li. Explaining young mortality. *Insurance: Mathematics and Economics*, 50(1):12–25, 2012.

Colin O'Hare and Youwei Li. Identifying structural breaks in stochastic mortality models. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg*, 1(2), 2015.

Claudia Pedroza. A Bayesian forecasting model: predicting US male mortality. *Biostatistics*, 7(4):530–550, 2006.

Francesca Perla, Ronald Richman, Salvatore Scognamiglio, and Mario V Wüthrich. Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, 2021(7):572–598, 2021.

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

Adrian E Raftery, David Madigan, and Jennifer A Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437): 179–191, 1997.

Arthur Renshaw and Steven Haberman. Lee–Carter mortality forecasting: A parallel generalized linear modelling approach for England and Wales mortality projections. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):119–137, 2003a.

Arthur E Renshaw and Steven Haberman. Lee–Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33(2):255–272, 2003b.

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

Ronald Richman and Mario V Wüthrich. A neural network extension of the Lee–Carter model to multiple populations. *Annals of Actuarial Science*, 15(2):346–366, 2021.

Maria Russolillo, Giuseppe Giordano, and Steven Haberman. Extending the Lee-Carter model: a three-way decomposition. *Scandinavian Actuarial Journal*, 2011(2):96–117, 2011.

Miguel Santolino. The Lee-Carter quantile mortality model. *Scandinavian Actuarial Journal*, 2020(7):614–633, 2020.

Simon Schnürch, Torsten Kleinow, and Ralf Korn. Clustering-based extensions of the common age effect multi-population mortality model. *Risks*, 9(3):45, 2021.

Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317, 2016. URL https://doi.org/10.32614/RJ-2016-021.

Syazreen Shair, Sachi Purcal, and Nick Parr. Evaluating extensions to coherent mortality forecasting models. *Risks*, 5(1):16, 2017.

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

Han Lin Shang. Mortality and life expectancy forecasting for a group of populations in developed countries: a multilevel functional data method. *The Annals of Applied Statistics*, 10(3):1639–1672, 2016.

Han Lin Shang and Heather Booth. Synergy in fertility forecasting: improving forecast accuracy through model averaging. *Genus*, 76(1):1–23, 2020.

Han Lin Shang and Steven Haberman. Model confidence sets and forecast combination: an application to age-specific mortality. *Genus*, 74(1):1–23, 2018.

Han Lin Shang and Steven Haberman. Retiree mortality forecasting: A partial age-range or a full age-range model? *Risks*, 8(3):69, 2020.

Han Lin Shang and Rob J Hyndman. Grouped functional time series forecasting: an application to age-specific mortality rates. *Journal of Computational and Graphical Statistics*, 26(2):330–343, 2017.

Yanlin Shi. Forecasting mortality rates with the adaptive spatial temporal autoregressive model. *Journal of Forecasting*, 40(3):528–546, 2021.

Sandra E Sinisi and Mark J van der Laan. Deletion/substitution/addition algorithm in learning with applications in genomics. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–38, 2004.

Dilan SriDaran, Michael Sherris, Andrés M Villegas, and Jonathan Ziveyi. A group regularisation approach for constructing generalised age-period-cohort mortality projection models. *ASTIN Bulletin*, 52(1):247–289, 2022.

Cary Chi-Liang Tsai and Echo Sihan Cheng. Incorporating statistical clustering methods into mortality models to improve forecasting performances. *Insurance: Mathematics and Economics*, 99:42–62, 2021.

John W Tukey et al. *Exploratory Data Analysis*, volume 2. Reading, MA, 1977.

Shripad Tuljapurkar. Future mortality: A bumpy road to Shangri-La?, 2005.

Frank Van Berkum, Katrien Antonio, and Michel Vellekoop. The impact of multiple structural changes on mortality predictions. *Scandinavian Actuarial Journal*, 2016(7):581–603, 2016.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer science & business media, 1999a.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999b.

Andrés Villegas, Vladimir K Kaishev, and Pietro Millossovich. StMoMo: An R package for stochastic mortality modelling. In *7th Australasian Actuarial Education and Research Symposium*, 2015.

Andrés M Villegas, Steven Haberman, Vladimir K Kaishev, and Pietro Millossovich. A comparative study of two-population models for the assessment of basis risk in longevity hedges. *ASTIN Bulletin*, 47(3):631–679, 2017.

Chou-Wen Wang and Sharon S Yang. Pricing survivor derivatives with cohort mortality dependence under the Lee–Carter framework. *Journal of Risk and Insurance*, 80(4):1027–1056, 2013.

Chou-Wen Wang, Hong-Chih Huang, I-Chien Liu, et al. A quantitative comparison of the Lee-Carter model under different types of non-Gaussian innovations. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 36(4):675–696, 2011.

Chou-Wen Wang, Hong-Chih Huang, and I-Chien Liu. Mortality modeling with non-Gaussian innovations and applications to the valuation of longevity swaps. *Journal of Risk and Insurance*, 80(3):775–798, 2013.

Chou-Wen Wang, Jinggong Zhang, and Wenjun Zhu. Neighbouring prediction for mortality. *ASTIN Bulletin*, 51(3):689–718, 2021.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

John R Wilmoth. Computational methods for fitting and extrapolating the Lee-Carter model of mortality change. Technical report, Technical Report, Department of Demography, University of California, Berkeley, 1993.

Arkadiusz Wiśniowski, Peter WF Smith, Jakub Bijak, James Raymer, and Jonathan J Forster. Bayesian population forecasting: extending the Lee-Carter method. *Demography*, 52(3):1035–1059, 2015.

Jackie ST Wong, Jonathan J Forster, and Peter WF Smith. Bayesian mortality forecasting with overdispersion. *Insurance: Mathematics and Economics*, 83:206–221, 2018.

Bowen Yang, Jackie Li, and Uditha Balasooriya. Cohort extensions of the Poisson common factor model for modelling both genders jointly. *Scandinavian Actuarial Journal*, 2016 (2):93–112, 2016.

Sharon S Yang and Chou-Wen Wang. Pricing and securitization of multi-country longevity risk with mortality dependence. *Insurance: Mathematics and Economics*, 52(2):157–169, 2013.

Rui Zhou, Yujiao Wang, Kai Kaufhold, Johnny Siu-Hang Li, and Ken Seng Tan. Modeling period effects in multi-population mortality models: Applications to Solvency II. *North American Actuarial Journal*, 18(1):150–167, 2014.

Rui Zhou, Guangyu Xing, and Min Ji. Changes of relation in multi-population mortality dependence: An application of threshold VECM. *Risks*, 7(1):14, 2019.

Appendices

Appendix A

Appendix of Chapter 2

A.1 Geographic Grouping

The following table illustrates the grouping of populations using geographic information and classifies the thirty populations into eight geographic groups:

Target Country	**Geographic Group**	**Target Country**	**Geographic Group**
Australia	Oceania	Netherlands	West Europe
Austria	West Europe	New Zealand	Oceania
Belarus	East Europe	Norway	Scandinavia
Belgium	West Europe	Poland	East Europe
Bulgaria	East Europe	Portugal	South Europe
Canada	North America	Russia	East Europe
Czech Republic	East Europe	Slovakia	East Europe
Denmark	Scandinavia	Spain	South Europe
Finland	Scandinavia	Sweden	Scandinavia
France*	West Europe	Switzerland	West Europe
Hungary	East Europe	Taiwan	Asia
Italy	South Europe	England & Wales	Great Britain
Japan	Asia	Scotland	Great Britain
Latvia	East Europe	U.S.A.	North America
Lithuania	East Europe	Ukraine	East Europe

Table A.1: Geographic grouping information.

- Oceania (2 members): Australia, New Zealand
- North America (2 members): Canada, U.S.A.
- Great Britain (2 members): England & Wales, Scotland
- Asia (2 members): Japan, Taiwan;
- Scandinavia (4 members): Denmark, Finland, Norway, Sweden
- West Europe (5 members): Austria, Belgium, France, Switzerland, Netherlands
- East Europe (10 members): Belarus, Poland, Bulgaria, Russia, Czech Republic, Slovakia, Hungary, Latvia, Lithuania, Ukraine
- South Europe (3 members): Spain, Italy, Portugal

A.2 Population-specific Results

Table A.2: Population-specific test MSEs comparing the prediction performance of DSA·MSE model versus benchmark ACF·GeoInfo model.

	ACF·GeoInfo	DSA·MSE	Change%
Female Population			
Australia	0.0274	0.0218	-20.36%
Austria	0.0937	0.0675	-27.97%
Belarus	0.0694	0.0720	3.73%
Belgium	0.0499	0.0518	3.88%
Bulgaria	0.0961	0.0481	-49.90%
Canada	0.0198	0.0168	-15.13%
Czech Republic	0.0899	0.0623	-30.71%
Denmark	0.1196	0.0949	-20.69%
Finland	0.0972	0.0924	-4.96%
France	0.0208	0.0150	-27.71%
Hungary	0.0853	0.0529	-38.02%
Italy	0.0259	0.0216	-16.47%
Japan	0.0146	0.0173	18.19%
Latvia	0.1981	0.1256	-36.60%
Lithuania	0.0870	0.0775	-10.93%
Netherlands	0.0496	0.0500	0.84%
New Zealand	0.0769	0.0760	-1.15%
Norway	0.0802	0.0867	8.01%
Poland	0.0375	0.0146	-61.05%
Portugal	0.0571	0.0549	-3.90%
Russia	0.0620	0.0344	-44.57%
Slovakia	0.0904	0.0749	-17.11%
Spain	0.0324	0.0286	-11.56%
Sweden	0.0708	0.0689	-2.70%
Switzerland	0.0865	0.0948	9.64%
Taiwan	0.0237	0.0213	-10.23%
England & Wales	0.0152	0.0137	-9.80%
Scotland	0.0683	0.0649	-4.94%
U.S.A.	0.0129	0.0131	1.81%
Ukraine	0.0709	0.0287	-59.49%
MEAN	**0.0643**	**0.0521**	**-15.59%**
MEDIAN	**0.0689**	**0.0523**	**-10.58%**

Appendix B

Appendix of Chapter 3

B.1 Calibration of the ACF-ts Model

Below are the specific steps for the calibration and extrapolation of the ACF-ts model with Δt fixed:

1. Shift data of the auxiliary population (i.e., population 2) Δt years backward along the timeline if $\Delta t \geq 0$ or Δt years forward if $\Delta t < 0$. We then aggregate the data by using age-time-specific population size as the weighting variable. Specifically, if we let $E_j(x,t)$ to denote the exposure number for age x and time t from population j, $j = 1, 2$, then the aggregate death rates are computed as

$$\frac{E_1(x,t)m_1(x,t) + E_2(x,t)m_2(x,t)}{E_1(x,t) + E_2(x,t)}.$$

   The resulting aggregated data span over the time horizon $[-\Delta t, T]$ if $\Delta t \geq 0$ or $[0, T - \Delta t]$ if $\Delta t < 0$; that is,

   - when $\Delta t \geq 0$:
     - over $[-\Delta t, 0)$, it only contains data from the auxiliary ($j = 2$);
     - over $[0, T - \Delta t]$, it contains the weighted average of the data from both the target ($j = 1$) and the auxiliary ($j = 2$);
     - over $(T - \Delta t, T]$, it only contains data from the target ($j = 1$).
   - when $\Delta t < 0$:
     - over $[0, -\Delta t)$, it only contains data from the target ($j = 1$);

    - over $[-\Delta t,\, T]$, it contains the weighted average of the data from both the target ($j = 1$) and the auxiliary ($j = 2$);
    - over $(T,\, T - \Delta t]$, it only contains data from the auxiliary ($j = 2$).

2. Based on the aggregated data from Step 1, we calibrate $B(x)$ and $K(t)$ as the first left and right singular vectors from a singular value decomposition (SVD) procedure. The length of calibrated sequence $K(t)$ depends on the value of Δt. The time index of realized $K(t)$ is $\{-\Delta t, -\Delta t + 1, \ldots, T\}$ for $\Delta t \geq 0$ and $\{0, 1, \ldots, T - \Delta t\}$ for $\Delta t < 0$.

3. Denote the residuals $R_1(x,t) = [\log m_1(x,t) - B(x)K(t)]$ and $R_2(x,t) = [\log m_2(x,t) - B(x)K(t-\Delta t)]$. Then, we calibrate $a_j(x)$ as an average of the residual rates over the modeling period for each population and age:

$$a_j(x) = \frac{1}{T+1}\sum_{t=0}^{T} R_j(x,t), \quad j = 1, 2, \text{and } \; x = x_1, \ldots, x_n.$$

4. For each $j = 1, 2$, we apply a SVD procedure further to $R_j(x,t) - a_j(x)$ to get $b_j(x)$ and $k_j(t)$.

5. Fit the sequence $K(t)$ by a RWD, and each sequence $k_j(t)$ with an AutoRegressive Integrated Moving Average (ARIMA) model using the `auto.arima` function from the `R` package `forecast`.

6. We follow the conventional extrapolation paradigm for mortality forecasting. We obtain projections of $K(t)$ and $k_1(t)$ into future years t and then use Equation (3.1) to obtain forecasts by taking the white noise terms as zero. The prediction of $K(t)$ is obtained either from the calibration step or by an extrapolating procedure, depending on the value of Δt. If $\Delta t \geq 0$, the aggregated data from Step 1 span over the time horizon $[-\Delta t, T]$ and in this case, the prediction of $K(t)$ for any future year $t > T$ should be obtained via extrapolating the established time series model. In contrast, if $\Delta t < 0$, the aggregated data from Step 1 span over the time horizon $[0, T - \Delta t]$, where we note $T - \Delta t > T$. In this case, we directly input the calibrate values of $K(t)$ obtained in Step 2 into equation (3.1) for mortality forecasts over the period $[T,\, T - \Delta t]$, and apply the extrapolation procedure for prediction beyond time $T - \Delta t$.

B.2 Calibration of the CBD-ts Model

The model extrapolation of the CBD-ts model follows the same Steps 5 and 6 for the ACF-ts model described in Section B.1. Below are the specific steps in our calibration for the

CBD-ts model with a fixed Δt:

1. Assume the number of deaths at age x in year t for jth population, denoted by $D_j(x,t)$, follows a Binomial distribution with an exposure number $E_j(x,t)$ and one-year death probability $q_j(x,t)$, i.e., $D_j(x,t) \sim Bin\big[E_j(x,t), q_j(x,t)\big]$.

2. With the realized number of death $D_j(x,t)$ and the exposure number $E_j(x,t)$, we compute the log-likelihood function:

$$\ell = \sum_{j=1}^{2} \sum_{l=1}^{n} \sum_{t \in \mathbb{S}} \ell\big(D_j(x_l,t), q_j(x_l,t)\big) \tag{B.1}$$

   where

$$\ell\big(D_j(x,t), q_j(x,t)\big) \quad \propto \quad D_j(x,t) \log\big[q_j(x,t)\big] + \big[E_j(x,t) - d_j(x,t)\big] \log\big[1 - q_j(x,t)\big].$$

3. From equations (3.3) and (3.4), we have the following two expressions:

$$q_1(x,t) = \text{expit}\big[K(t) + (x - \bar{x})k_1(t)\big], \tag{B.2}$$
$$q_2(x,t) = \text{expit}\big[K(t - \Delta t) + (x - \bar{x})k_2(t)\big], \tag{B.3}$$

   where $\text{expit}(u) = \frac{\exp(u)}{1+\exp(u)}$.

4. Calibrate the $K(t)$ and $k_j(t)$ sequences by plugging (B.2) and (B.3) into the log-likelihood (B.1) and then maximizing it with respect to $K(t)$ and $k_j(t)$ using the Newton-Raphson iterative procedure.

B.3 Population-specific results

Table B.1: Population-specific test SSEs comparing the prediction performance of the BMBE based approaches ACF-ts·RankAvg versus benchmark ACF·GeoInfo model.

	ACF·GeoInfo	ACF-ts·RankAvg	Change%
	Male Population		
Australia	26.45	27.22	2.93%
Austria	34.94	34.54	-1.14%
Belgium	38.75	33.22	-14.29%
Canada	13.33	14.98	12.40%
Czech Republic	39.67	28.54	-28.05%
Denmark	66.33	65.67	-0.98%
Finland	73.81	58.63	-20.57%
France	15.10	16.02	6.13%
Hungary	54.41	28.80	-47.07%
Italy	36.10	28.03	-22.35%
Japan	9.36	4.72	-49.61%
Netherlands	35.08	31.13	-11.28%
New Zealand	54.01	56.62	4.84%
Norway	72.18	55.94	-22.50%
Poland	15.00	6.53	-56.47%
Portugal	73.83	68.92	-6.64%
Slovakia	56.78	38.13	-32.85%
Spain	29.67	32.10	8.22%
Sweden	64.36	49.41	-23.23%
Switzerland	70.34	66.47	-5.50%
Taiwan	27.43	27.19	-0.89%
England & Wales	15.64	12.42	-20.55%
Scotland	63.91	61.71	-3.45%
U.S.A.	7.37	6.88	-6.59%
MEAN	**41.41**	**35.58**	**-14.15%**
MEDIAN	**37.43**	**31.62**	**-8.96%**

Table B.2: Population-specific test SSEs comparing the prediction performance of the BMBE based approaches CBD-ts·RankAvg versus benchmark CBD model.

	CBD	CBD-ts·RankAvg	Change%
	Male Population		
Australia	2.32	2.88	23.94%
Austria	3.72	3.62	-2.73%
Belgium	3.81	3.67	-3.85%
Canada	2.16	1.63	-24.64%
Czech Republic	2.30	1.79	-22.06%
Denmark	3.60	1.89	-47.46%
Finland	5.16	5.62	8.91%
France	5.04	5.05	0.20%
Hungary	2.35	2.64	12.52%
Italy	1.52	1.54	1.30%
Japan	2.83	2.19	-22.65%
Netherlands	5.90	3.45	-41.49%
New Zealand	3.44	5.34	55.27%
Norway	5.01	3.36	-33.00%
Poland	1.50	1.32	-12.56%
Portugal	2.92	2.93	0.28%
Slovakia	2.37	2.60	9.99%
Spain	2.30	2.09	-8.88%
Sweden	2.43	1.67	-31.39%
Switzerland	3.00	2.93	-2.19%
Taiwan	1.25	2.34	86.93%
England & Wales	2.54	2.70	6.23%
Scotland	3.31	2.15	-34.96%
U.S.A.	3.32	3.05	-7.96%
MEAN	**3.09**	**2.85**	**-3.76%**
MEDIAN	**2.88**	**2.67**	**-3.29%**