Coherent Lee-Carter method with smoothing and adjusting for lifespan disparity

Ahbab Mohammad Fazle Rabbi[1,2], Stefano Mazzuco[1], Heather Booth[3], and Vladimir Canudas-Romo[3]

[1]*Department of Statistical Sciences, University of Padua, Italy*
[2]*Department of Mathematics and Statistics, BUTEX, Bangladesh*
[3]*School of Demography, Australian National University, Australia*

Abstract

Coherent mortality forecasting methods try to capture the influence of common improvement of health, communication, science on a specific population. The commonly used coherent method, Li-Lee (LL) method, is a hierarchical form of the Lee-Carter (LC) method. The LC model assumes an invariant age component and a presumably linear time component. The LL model fits the joint mortality data of populations which are believed to have closer interaction in terms of mortality regime (relational populations) and here the time component is adjusted to match observed life expectancy at birth. Choosing the appropriate reference population remains an arbitrary process. We propose to use smoothed mortality rates obtained by *LASSO* type regularization in LL model and hence to partially adjust the time component of the LL model according to observed lifespan disparity to get the common factor of the relational populations (reference group). Time variability is also taken into consideration during obtaining the common factor. The relational populations for making coherent forecast for a particular population is chosen from the set of available populations on the basis of closest lifespan disparity over time. The proposed methodology generates less forecast errors than existing method during the out-of-sample forecast period and also produce a more optimistic forecast of life expectancy for most of the low-mortality countries. Moreover, choosing the relational populations based on closest lifespan disparity made the choice of reference populations more scientific.
Keywords: Lee-Carter method; Li-Lee method, Coherent mortality forecasting; Mortality smoothing; Lifespan disparity; LASSO

1 Introduction

Accurate forecasts of mortality and life expectancy are needed for decision making in the social, health and financial sectors of modern societies, while fundamental changes in the welfare policies for ageing populations depend on the accurate forecasting of longevity (Booth et al. 2002). Stochastic modeling of mortality forecasting is rapidly gaining recognition in this context (Hyndman and Booth 2008). Among the several approaches to date, the most prominent is the Lee-Carter (LC) method proposed in 1992 (Lee and Carter 1992). This method decomposes the differences between age-time-specific log mortality rates and their temporal averages into two parts; an age component and a time component. The mortality forecast is obtained by standard time series forecasting of the time component. Thus, this model is a simple but powerful probabilistic forecasting approach which does not suffer from over-parameterization.

Given the advantage of the straightforward interpretation of the LC model parameters, several low-mortality countries use different variants of the model in mortality forecasting. Various adjustments and variants have been proposed to modify the accuracy of the estimated age and time components of the model. In their original method, Lee and Carter (1992) incorporated adjustments to the estimate of the time component by setting the predicted annual total death count to be equal to the observed total count. Alternative adjustment procedures involved re-estimation of the time component to match the observed annual life expectancy at birth (Lee and Miller 2001) or the annual age distribution of deaths (Booth et al. 2002). These latter two approaches also consider other modifications by which to improve forecast accuracy, including choice of fitting period and method of estimation (see also Brouhns et al. 2002).

Several alternative approaches have been proposed. Among these, Renshaw and Haberman (2000) used a generalized linear model and maximum likelihood estimation of its parameters. Different stochastic models have also been introduced to integrate the cohort dimension in mortality modelling (eg. Cairns et al. 2011). Further, the LC method has been proposed in a Bayesian framework (eg. Wiśniowski et al. 2015). Due to noise in the data and the presence of outliers, estimates of the age component can be inaccurate (Girosi and King 2006; Booth and Tickle 2008). This issue led to the application of smoothing techniques prior to model fitting (Hyndman and Ullah 2007; Rabbi and Mazzuco 2021).

To simultaneously address the issues affecting both the time and age components, Rabbi and Mazzuco (2021) proposed fitting the LC model to smoothed mortality rates and adjusting the fitted time component according to the observed annual lifespan disparity (Vaupel and Canudas-Romo 2003; Vaupel et al. 2011). Rabbi and Mazzuco (2021) applied LASSO for smoothing the mortality rates, as LASSO produces less error in data-fitting (Dokumentov et al. 2018).The adjustment of the time component with respect to lifespan disparity, ($e^{\dagger}$), which measures variation in individual age at death within a population (Vaupel and Canudas-Romo 2003), is based on the finding that lifespan disparity has a strong inverse relation with life expectancy at birth (Vaupel et al. 2011; Aburto et al. 2020; Rabbi and Mazzuco 2021). Moreover, lifespan disparity may provide more information regarding a mortality profile than does the age distribution of deaths, total number of deaths, or life expectancy, particularly for ageing populations (Rabbi and Mazzuco 2021). Lifespan disparity becomes smaller as populations age as a result of declining mortality. Lifespan disparity may provide further information about the decrease or increase in mortality at different ages (Zhang and Vaupel 2009) and may better capture premature mortality than life expectancy (Aburto and van Raalte 2018). Owing to its stability over time, Bohk-Ewald et al. (2017) employed $e^{\dagger}$ to evaluate forecast performance rather than only considering the fitted mortality rates or life expectancy. The method proposed by Rabbi and Mazzuco (2021) produced more accurate forecasts than several other LC variants for many of the low mortality populations they studied.

Among the many LC variants that have been proposed, multi-population forecasting is gaining attention as it seeks to ensure that forecasts for related populations maintain certain structural relationships based on past mortality patterns and theoretical understanding. The first approach to such 'coherent' mortality forecasting was introduced by Li and Lee (2005) as an extended hierarchical interface of the LC method. Li and Lee (2005) modified their original LC model by specifying a common trend to which population-specific rates eventually converge. In this model, the 'common factor' refers to the product of the age and time components from the fitted model of the joint mortality of related populations. Other approaches to coherent forecasting have subsequently been developed. Hyndman et al. (2013) extended the nonparametric multiple component approach of Hyndman and Ullah (2007), of which LC is a special case, Ahmadi and Li (2014) used generalized linear modeling and Bergeron-Boucher et al. (2017) used compositional data analysis of the distribution of deaths.

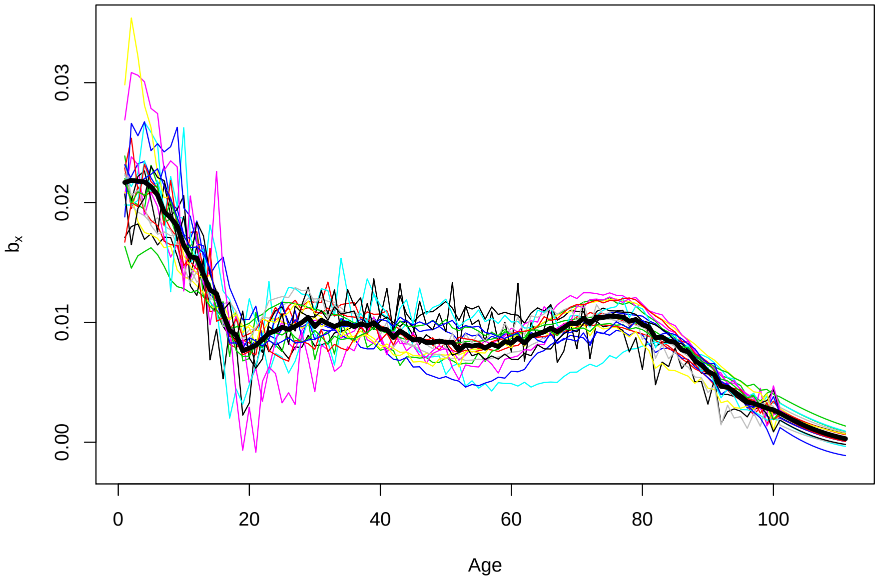
Almost all of the existing coherent forecasting methods share the core problem of choosing an appropriate reference mortality for the population of interest. The reference may be based on a single population or an appropriate group of populations, and its choice depends on several considerations. First, the male-female gap in mortality remains as a constraint while choosing a reference group. Keeping both gender together in the reference group is a topic of long debate. Second, similar economic and political frameworks among a group of countries (European Union for an example) may also be taken into account (Kjærgaard et al. 2016) for choosing appropriate reference population. Some studies also cited environmental factors and geographical location as significant impact on mortality pattern (Ahcan et al. 2014).

Third, historical events (for example war or international unions of countries) are often reflected in the demographic characteristics of a population. For example, the rapid increase in Portuguese life expectancy after the country joined the European Union (HMD 2019) is a result of the adaptation to and implementation of the health policies of the European Union (Santos et al. 2020). A similar situation occurred in the former East Germany after reunification (Vogt and Missov 2017). To the best of our knowledge, none of the coherent forecasting methods have considered such effects. Fourth, consideration should be given to the number of populations combined to form the reference group. Kjærgaard et al. (2016) found that a reference population made up of a small number of countries tends to perform better in terms of forecast accuracy than one comprising a larger group. Choosing populations with closer life expectancy to that of the population of interest was found to be a better strategy than choosing those with less similar mortality levels (Kjærgaard et al. 2016). These findings are important because choosing reference populations subjectively may introduce bias. However, merely adopting a specific number of populations for the reference group does not make selection any easier. Rather, if all countries with appropriate data were to be considered, this strategy involves enormous computational complexities. For example, with 39 possibly eligible populations, if one looks for a reference group of size 5 (including the particular population of interest) where the other reference group populations are to be chosen from the remaining 39 populations, the best reference group of size 5 would need to be identified from ${}^{39}C_4$= 82,251 possible combinations.

Further, although low-mortality convergence is taking place, each populations still has a distinct age or periodic pattern of mortality improvement. A reference group consisting populations of completely different mortality regime rather than that of population of interest may produce misleading mortality forecasts (Booth 2020). To reduce computation complexities of combining all possible groups of populations, consideration may be given to the development/specification of robust prior assumptions about coherence among population. To illustrate, the estimated relative change (b_x) in the log-mortality rate at each age x from fitted Lee-Carter model for females of 20 low-mortality countries are illustrated in Figure 1. Comparing with the mean trend of b_x (marked with thick black line), clearly, each of the populations showed different pattern of mortality improvement in different parts of life span.

In this paper, we extend previous research by utilizing the forecasting method proposed by Rabbi and Mazzuco (2021) in a coherent setting. Briefly, we propose to apply the Li and Lee (2005) methodology to smoothed mortality rates obtained by LASSO, and then to partially adjust the time component of the common factor to match observed lifespan disparity ($e^{\dagger}$). Further, the common factor of the coherent model is estimated utilizing only a subset of the available years (i.e., the best fitting period), and these same years are used to define the population-specific fitting period. The chosen group of reference populations comprises those with the closest lifespan disparity to the population of interest.

Figure 1: Estimated relative change (b_x) in the log-mortality rate at each age x from fitted Lee-Carter model for females of 20 low-mortality countries (1956-2011)



Source: HMD (2019) and authors' calculations.
Note: The bold black line is the mean trend of b_x to show the distinct mortality improvement patterns.

2 Methods

Mortality forecasting with smoothing and adjusting for lifespan disparity

The two-factor LC model is,

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

where, $m_{x,t}$ is the central death rate at age x for year t; a_x represents the average of log-mortality at age x over time; b_x is the set of age-specific constants explaining the relative speed of change at each age x; k_t represents the overall level of mortality in year t; and $\epsilon_{x,t}$ is the model residual. To obtain ordinary least square (OLS) estimates of the model parameters, a_x is estimated by averaging over observed log of $m_{x,t}$. A new matrix $Z_{x,t} = \ln(m_{x,t} - \hat{a}_x)$ is constructed after obtaining a_x. Singular value decomposition (SVD) is performed over $Z_{x,t}$ to obtain OLS estimates of b_x and k_t (Lee and Carter 1992). A maximum likelihood approach can also be considered to estimate the model parameters (Brouhns et al. 2002), however, the results from these different approaches may vary slightly (Shang 2012). Lee and Carter (1992) proposed to adjust k_t to match observed death counts, whereas later variants choose different policies to adjust k_t. Another LC variant consider to adjust k_t to match observed life expectancy at birth (Lee and Miller 2001). Later mortality forecasting is done by applying standard time series models on k_t (Lee and Carter 1992).

Rabbi and Mazzuco (2021) applied the LC method on smoothed mortality rates obtained by LASSO-type regularization and hence adjust the time component with the observed lifespan disparity. Detailed estimation procedure for LASSO is given elsewhere (Dokumentov et al. 2018; Rabbi and Mazzuco 2021). LASSO smooths the data with higher accuracy than spline based smoothing techniques, while it helps to overcome irregularities in b_x. To measure lifespan disparity, Rabbi and Mazzuco (2021) considered average number of life years lost at birth

(Vaupel and Canudas-Romo 2003; Zhang and Vaupel 2009). Symbolically,

$$e_t^\dagger = \frac{\int_0^\omega e_{x,t} d_{x,t} \ dx}{l_{0,t}}, \tag{2}$$

where, ω is the maximum attainable age; $d_{x,t}$ is the distribution of death in year t. Thus estimation of $e_t^\dagger$ is simple and straightforward.

Rabbi and Mazzuco (2021) obtained the initial model parameters following the same procedure of LC model (Lee and Carter 1992). However, after obtaining the b_x and k_t, Rabbi and Mazzuco (2021) adjusted the estimated k_t by solving

$$e^\dagger_{0;t\ observed} = \sum_0^\omega \exp(\hat{a}_x + \hat{b}_x.k_{t\ adj}) e_x l_x / l_0. \tag{3}$$

The e_x and l_x in equation (3) are obtained from the life table estimated from the fitted $\hat{m}_x$. An ARIMA(0,1,0) with drift is then fitted for the adjusted $\hat{k}_t$ for mortality forecasting.

Existing method for coherent mortality forecasting

Li and Lee (2005) modified the standard Lee and Carter (1992) model to forecast mortality for several countries by taking into account their membership in a group, rather than forecasting individually. To extend the standard LC model to a coherent setting, Li and Lee (2005) first estimated the average mortality trend for the reference group (containing the populations of interest and other "related"populations) and addressed it as common factor. Hence they added the historical particularities of particular population (unexplained by the common factor) in a hierarchical way to the standard LC model. Therefore, in the short term, inter-country mortality differences in trends may be preserved, but ultimately age-specific death rates within the group of countries are constrained to maintain a constant ratio to one another (Li and Lee 2005). This extended model can be formulated as,

$$\ln m_{x,t,i} = a_{x,i} + B_x K_t + b_{x,i} k_{t,i} + \epsilon_{x,t,i}, \tag{4}$$

where i denotes the specific country in the group, $a_{x,i}$ is country-specific average log death rates. The term B_x is the relative speed of change in mortality at each age x and K_t is the mortality index capturing the main time trend for the group. Li and Lee (2005) or LL for short called the term $B_x K_t$ the common factor as this quantity is common for all countries of the group. To obtain the country-specific estimates of the LL model, first the LC model is fitted to the joint mortality data (combining all populations in the group including the population of interest) from which the common factor is estimated for use in equation (4). At this stage, K_t is adjusted for observed life expectancy of the joint mortality data. A second singular value decomposition (SVD) is performed on the matrix of differences of country-specific mortality rates and $a_{x,i}$ and the common factor to estimate country-specific $b_{x,i}$ and $k_{t,i}$. The forecast is obtained by applying standard time series models to both of the time components. In the current study, a random walk with drift is used for forecasting both for K_t and $k_{t,i}$.

Proposed method for coherent mortality forecasting

In the proposed methodology for coherent forecasting, we apply the LL method on mortality rates smoothed by *LASSO* type regularization (Dokumentov et al. 2018) and to adjust the fitted time component of the common factor using the observed lifespan disparity, $e_0^\dagger$ (Vaupel and Canudas-Romo 2003; Zhang and Vaupel 2009). Lasso smooths the data in a similar manner to spline-based techniques for the mortality age pattern, but is more accurate (Dokumentov et al. 2018). Rabbi and Mazzuco (2021) applied this methodology for forecasting mortality for single populations. In addition, we employed a variable fitting period to address the period effect on coherent forecasting.

2.1 Choice of the reference group

Unlike different subjective approaches or geographic proximity for choosing reference group (see Ahcan et al. 2014, for example), demographic perspectives like similar pattern of life expectancies seemed more realistic in terms of forecast accuracy (Kjærgaard et al. 2016). In our proposed methodology, the reference group for a particular population is chosen on the basis of closest lifespan disparity over time. There are several benefits for considering lifespan disparity in the context of mortality forecasting. For ageing societies where survival is highly concentrated around older ages, the difference between the age at death and the expected remaining years decreases. As a result, lifespan disparity gets smaller over time for ageing societies, and it may better capture premature mortality unlike life expectancy (Aburto and van Raalte 2018). Lifespan disparity may provide further information about the spread of mortality over ages (Zhang and Vaupel 2009). $e_t^\dagger$ can be estimated using different approximation, however, all of them are provides almost similar outcome (Vaupel and Canudas-Romo 2003; Aburto and van Raalte 2018; Aburto et al. 2020). A sensitivity analysis is attached in the appendix. We base the choice of reference group populations on the observed trend of $e_t^\dagger$ with two assumptions: (i) a smaller number of populations is preferred for the sake of parsimony and (ii) future convergence of mortality among target populations. The second assumption is important in proposed methodology as we form reference group without distinguishing between male and female mortality; rather we choose them on the basis of closer lifespan disparity. For a particular population we choose the populations giving the smallest difference in observed $e^\dagger$. For a particular population i, another population j will be in the reference group if

$$\left|\bar{e}_{t,i}^\dagger - \bar{e}_{t,j}^\dagger\right| < min, \tag{5}$$

compared to other available populations. Therefore only those population will be in the reference group for i-th population for which this difference is lowest. We choose this number of member in the best reference group on the basis of forecast accuracy (details are given in results). To implement this, we estimate $e^\dagger$ for all populations under consideration over the common fitting period and average over time to obtain the population specific $\bar{e}_t^\dagger$. For m years of available mortality rates and p populations the estimates can be presented in the following matrix notation:

$$\begin{array}{c} \text{Year} \\ t_1 \\ t_2 \\ \vdots \\ \vdots \\ t_{m-1} \\ t_m \end{array} \begin{array}{c} \begin{array}{cccc} \text{Population}_1 & \text{Population}_2 & \dots & \text{Population}_p \end{array} \\ \begin{pmatrix} e_{t_1,1}^\dagger & e_{t_1,2}^\dagger & \dots & e_{t_1,p}^\dagger \\ e_{t_2,1}^\dagger & e_{t_2,2}^\dagger & \dots & e_{t_2,p}^\dagger \\ \vdots & \vdots & & \vdots \\ & & & \\ e_{t_{m-1},1}^\dagger & e_{t_{m-1},2}^\dagger & \dots & e_{t_{m-1},p}^\dagger \\ e_{t_m,1}^\dagger & e_{t_m,2}^\dagger & \dots & e_{t_m,p}^\dagger \end{pmatrix} \end{array}$$

For a particular population i, we sorted all $\bar{e}^{\dagger}_{t,j}$ in ascending order. Populations j for which $\bar{e}^{\dagger}_{t,j}$ is closest to $\bar{e}^{\dagger}_{t,i}$ is selected into the reference group for population i. The pattern of ranked, temporal $\bar{e}^{\dagger}_{i,j}$ may show inconsistent pattern depending on the study populations' mortality trend. For the low-mortality populations considered in this study, we obtained a consistent pattern of sorted $\bar{e}^{\dagger}_{i,j}$ starting from the period of 1982 as the sorted $\bar{e}^{\dagger}_{i,j}$ have almost same pattern since then. Clearly, the number of populations and required gap between $\bar{e}^{\dagger}_{t,i}$ and $\bar{e}^{\dagger}_{t,j}$ may vary among reference groups as not all populations have symmetric distance of $\bar{e}^{\dagger}$ between each other. Following the findings of Kjærgaard et al. (2016), we did not pre-fix the number of populations in best reference group. Instead, we choose a best reference group consists of less number of populations and providing greatest forecast accuracies during out-of-sample evaluation period. We did not consider or compare different strategies for choosing reference population. Nevertheless, based on the strong relation between $e^{\dagger}$ and e_0, similar relational populations may be obtained for a temporal average of life expectancies as well (Vaupel et al. 2011; Rabbi and Mazzuco 2021).

2.2 Coherent forecast of mortality rates

After identifying the populations to be used as the reference group, we smoothed the mortality rates of each population using LASSO (Dokumentov et al. 2018). The rationale and advantages for using LASSO in mortality forecasting are illustrated by Rabbi and Mazzuco (2021). The common factor obtained from the smoothed mortality rates with a smoothed B_x and a more regular k_t ultimately produced a more smoothed surface than that observed for the LC models (Girosi and King 2006). Rabbi and Mazzuco (2021) obtained higher forecast accuracy from LASSO compared to the spline based techniques. In the next step, we combined respective populations in the reference group and followed the standard LC methodology to obtain the initial estimates of the age and time component. In this stage, we applied same weight on all populations to overcome problem of combining larger exposure to smaller exposure (as done by Li and Lee (2005)),

$$m_{x,t} = \frac{1}{p}\sum_{i=1}^{p} m_{x,i,t}. \tag{6}$$

These $m_{x,i,t}$ are without any influence of population size. In this stage, fitted two-factor LC model over the joint mortality rates (reference group) will be,

$$\ln(\hat{m}_{x,t}) = \hat{a}_x + \hat{B}_x\hat{K}_t; \tag{7}$$

which is known as common factor model (Li and Lee 2005). Following Lee and Miller (2001), LL method makes a second stage estimate of $\hat{K}_t$ by finding the value of $\hat{K}_t$ which produces exactly the observed life expectancy for the fitting period of the model. Following Rabbi and Mazzuco (2018), we adjust the estimated $\hat{K}_t$ by solving the equation that produces the observed lifespan disparity, as

$$e^{\dagger}_{t\ obs} = \sum_{0}^{\omega} \exp(\hat{a}_x + \hat{B}_x.\hat{K}_{t\ adj})e_{x,t}l_{x,t}/l_{0,t}. \tag{8}$$

Here $l_{x,t}$ is the number of people alive at age x in year t ($l_{0,t}$ is the life table radix for year t). Following Rabbi and Mazzuco (2021), all the terms of the right hand side of the equation 8 are obtained from life table estimated from the fitted common factor model (equation 7) except for $\hat{K}_{t\ adj}$. After obtaining the adjusted $\hat{K}_t$, we identified the most appropriate period for which this common factor should be considered. Most of the LC variants obtained linear trend of k_t for vast majority of the populations (Lee 2000). We considered this findings for the K_t to obtain the best fitting period by choosing the fitting period which maximize the linearity (Booth et al. 2002). It is also based on the idea that some more distant past history may not be relevant for the future. In order to get the best fitting period, we obtain the close approximation of deviance (t) which is equal to $\chi^2(t)$ statistic of the lack of fit in observed distribution of deaths $D_{x,t}$,

$$\chi^2(t) = \sum_x \frac{\left[D_{x,t} - D'_{x,t}\right]^2}{D'_{x,t}}; \tag{9}$$

where $D'_{x,t}$ are fitted deaths which can be obtained from observed exposure $N_{x,t}$ as follow:

$$D'_{x,t} = N_{x,t}\left[\exp(\hat{a}_x + \hat{B}_x.K_{t\ adj})\right]. \tag{10}$$

Following Booth et al. (2002), the total lack of fit to the log-linear model derives from two sources: the base lack of fit from the log-additive model or LC model with $\hat{K}_{t\ adj}$ and the additional lack of fit from the imposition of the ARIMA(0,1,0) model on $\hat{K}_{t\ adj}$. The base lack of fit for the period S years prior to last year of the fitting period is measured by

$$\chi^2_{\text{logadd}}(S) = \sum_t \chi^2_{\text{logadd}}(t);$$

where the $D'_{x,t}$ are derived from $\hat{K}_{t\ adj}$. For the log-linear model,

$$\chi^2_{\text{loglin}}(S) = \sum_t \chi^2_{\text{loglin}}(t);$$

here the $D'_{x,t}$ are derived from the linear fit of $\hat{K}_{t\ adj}$. This total lack of fit will be greater than or equal to the base lack of fit. According to Booth et al. (2002), to compare $\chi^2_{\text{loglin}}(S)$ and $\chi^2_{\text{logadd}}(S)$ they are divided by the corresponding degrees of freedom to produce mean-χ^2 statistic. For n age categories and m years in the fitting period, the df for $\chi^2_{\text{loglin}}(S)$ is $n(m-2)$ and df for $\chi^2_{\text{logadd}}(S)$ is $(n-1)(m-1)$. The length of S is determined by the extent of the additional lack of fit relative to the total lack of fit. The additional lack of fit will be small for a good fit of the ARIMA(0,1,0) model. Booth et al. (2002) formulated the total to base lack of fit as,

$$R(S) = \frac{\chi^2_{\text{loglin}}(S)/[n(m-2)]}{\chi^2_{\text{logadd}}(S)/[(n-1)(m-2)]}. \tag{11}$$

The next step is to obtain the marginal effect of including one more year in S which can be obtained from the ratio of the differences in total and base mean-χ^2 statistics for S and $S+1$,

$$RD(S) = \frac{\left[\chi^2_{\text{loglin}}(S) - \chi^2_{\text{loglin}}(S+1)\right]/n}{\left[\chi^2_{\text{logadd}}(S) - \chi^2_{\text{logadd}}(S+1)\right]/(n-1)}. \tag{12}$$

Small values of $R(S)$ and $RD(S)$ indicate that the additional lack of fit is relatively small. The best fitting period is identified by the value of S for which $R(S)$ and $RD(S)$ are substantially smaller than corresponding statistics for preceding values of S, indicating that the inclusion of $S-1$ years (and preceding) prior to last year of fitting period in the fitting period results in a relatively large reduction in goodness of fit of the ARIMA model (Booth et al. 2002).

For country-specific coherent forecast, the basic LC model is then fitted to country-specific mortality rates without the common factor. To obtain the country-specific ordinary least square estimates of $b_{x,i}$ and $k_{t,i}$, SVD is performed on

$$Z_{x,i,t} = \ln(m_{x,i,t}) - \hat{a}_{x,i} - \hat{B}_x\hat{K}_{t\ adj}. \tag{13}$$

The estimation procedure is as before. However, at this stage the LC model is fitted without any adjustment for country-specific $k_{t,i}$ and it is fitted for the best fitting period obtained during estimation of the common factor. A random walk with drift is then fitted to both $\hat{K}_{t\ adj}$ and $\hat{k}_{t,i}$. Following Lee and Miller (2001) and Li and Lee (2005), we also used the actual data as the jump-off rates to avoid jump–off error.

2.3 Forecast accuracy and interval forecast

We consider the following two measures for checking the forecast accuracy of mortality rates:

mean absolute forecast error,

$$\text{MAE} = \frac{1}{(p+1)q} \sum_{t=1}^{q} \sum_{x=0}^{n} \left| m_{x,t} - \widehat{m}_{x,t|t-h} \right|; \tag{14}$$

mean squared forecast error,

$$\text{MSE} = \frac{1}{(p+1)q} \sum_{t=1}^{q} \sum_{x=0}^{n} \left(m_{x,t} - \widehat{m}_{x,t|t-h} \right)^2; \tag{15}$$

and for life expectancy at birth, we consider the mean error,

$$\text{ME} = \frac{1}{q} \sum_{t=1}^{q} \left(\hat{e}_{0,t} - e_{0,t} \right). \tag{16}$$

Here $m_{x,t}$ represents the observed mortality rate for age x in year t and $\widehat{m}_{x,t}$ represents the forecast; $e_{0,t}$ represents the observed life expectancy at birth in year t and $\hat{e}_{0,t}$ represents the forecast. From the available mortality data, we used the last 10 years as the period for forecasting and the previous years as the fitting period. Using the data in the fitting period, we made ten-step-ahead forecasts, and determined the forecast accuracy by comparing the forecasts with the observed data in the hold-out period. We skipped the comparison of smoothing techniques in this paper as it is well documented by Dokumentov et al. (2018) and Rabbi and Mazzuco (2021). Through out the paper we denoted the proposed method as $\text{LL}_{e^\dagger}$ and Li and Lee (2005) by LL.

To construct prediction intervals for forecast life expectancy at birth, we followed the procedure employed by Hyndman and Booth (2008). The fitted mortality rates from each forecasting technique were simulated 500 times to add disturbance to the time component of the model. Life expectancy at birth was calculated for each age-time set of simulated log-mortality rates. Prediction intervals were defined by the 80% and 95% percentiles of the ranked simulated life expectancies for each year of the forecast.

2.4 Implementation

The analysis performed in this study is implemented in R. We implemented the LL method and the new proposed method by modifying the existing functions in 'Demography' package (Hyndman et al. 2011). For LASSO we used the 'smoothAPC' package (Dokumentov and Hyndman 2017).

3 Data

The data used in this study came from Human Mortality Database (HMD 2019). The HMD has high data quality and it strives to provide mortality data for any population for which death registration and census data are virtually complete. We analyzed the mortality of 40 populations, namely the male and female populations of the following 20 low-mortality countries: Australia (AUS), Austria (AUT), Belgium (BEL), Canada (CAN), Denmark (DNK), Finland (FIN), France (FRA), Germany (DEU), Ireland (IRL), Italy (ITA), Japan (JPN), Netherlands (NLD), New Zealand (NZL), Norway (NOR), Portugal (PRT), Spain (ESP), Sweden (SWE), Switzerland (CHE), United Kingdom (UK) and USA (USA). As data for Germany are not available before 1990, we combined the data for East and West Germany. Total populations, rather than smaller subpopulations, are considered for France, New Zealand and the United Kingdom. For all of these countries, HMD covers the period 1956 to 2011, which is used as fitting period. This time period also avoid the disruptions of epidemics and WWII. The data are available for ages 0 to 110+, and we constructed life tables for those ages. Missing values at ages older than 100 were estimated using the Kannisto model (Thatcher et al. 1998), details of which appear in the Appendix.

4 Results

This section provides details of each stage of the analysis, focusing by way of illustration on the forecast of female mortality in France. We denote the new method as $\mathrm{LL}_{e^\dagger}$, and also provide comparisons with the LL method.

4.1 Ranking of reference populations

To identify populations that are close to each other in terms of lifespan disparity, we calculated $\bar{e}^\dagger$ for all 40 populations and sorted them from smallest to largest for each period starting in years 1956 through 2011 and ending in 2011. The ranking of populations remained stable for periods starting in years 1982 and later. The sorted $\bar{e}^\dagger$ were found to be ranked naturally by sex with the lowest values being for females although some overlap with males occurs in the middle of the range. Table 1 shows the sorted values of $\bar{e}^\dagger$ for the period 1982 to 2011.

Table 1: Sorted values of $\bar{e}^\dagger$ for the period 1982-2011 for 40 sex-specific low-mortality populations

Population	$\mathrm{ESP_F}$	$\mathrm{CHE_F}$	$\mathrm{JPN_F}$	$\mathrm{SWE_F}$	$\mathrm{FIN_F}$	$\mathrm{ITA_F}$	$\mathrm{AUT_F}$	$\mathrm{NOR_F}$
$\bar{e}^\dagger$	9.888	9.922	9.933	9.969	10.011	10.043	10.101	10.150
Population	$\mathrm{DEU_F}$	$\mathrm{NLD_F}$	$\mathrm{FRA_F}$	$\mathrm{BEL_F}$	$\mathrm{AUS_F}$	$\mathrm{IRL_F}$	$\mathrm{PRT_F}$	$\mathrm{UK_F}$
$\bar{e}^\dagger$	10.246	10.326	10.378	10.454	10.467	10.493	10.618	10.791
Population	$\mathrm{CAN_F}$	$\mathrm{SWE_M}$	$\mathrm{NLD_M}$	$\mathrm{NZL_F}$	$\mathrm{DNK_F}$	$\mathrm{JPN_M}$	$\mathrm{NOR_M}$	$\mathrm{IRL_M}$
$\bar{e}^\dagger$	10.881	10.911	11.039	11.100	11.103	11.310	11.340	11.341
Population	$\mathrm{CHE_M}$	$\mathrm{UK_M}$	$\mathrm{ITA_M}$	$\mathrm{AUS_M}$	$\mathrm{DEU_M}$	$\mathrm{BEL_M}$	$\mathrm{USA_F}$	$\mathrm{DNK_M}$
$\bar{e}^\dagger$	11.464	11.495	11.547	11.648	11.717	11.800	11.802	11.831
Population	$\mathrm{CAN_M}$	$\mathrm{AUT_M}$	$\mathrm{ESP_M}$	$\mathrm{NZL_M}$	$\mathrm{FIN_M}$	$\mathrm{FRA_M}$	$\mathrm{PRT_M}$	$\mathrm{USA_M}$
$\bar{e}^\dagger$	11.857	11.949	11.990	11.999	12.154	12.516	12.808	13.127

Note: Country codes are listed in section 3. M and F indicate males and females respectively. $\bar{e}^\dagger$ are calculated using unsmoothed data; ranking unchanged for smoothed data.

For a particular population of interest, the populations closest in terms of $\bar{e}^\dagger$ can be easily identified from Table 1. The rank order of closest (above or below) populations is important for the $\mathrm{LL}_{e^\dagger}$ method of coherent forecasting because populations are added to the reference group following this order. This process did not involve predetermining the number of populations in the reference group, as this is based on out-of-sample forecast accuracy for successively larger reference groups. To illustrate, in the case of the female population of France, the closest population in terms of $e^\dagger$ is the female population of the Netherlands, followed by the female population of Belgium, and so on; this ordering is more clearly seen in Table 2 which shows ranked absolute differences in life span disparities between French Females and all other populations. As expected, given the sex gap in mortality, the closest populations are female populations.

Table 2: Reference population for French Females based on closest absolute difference in $\bar{e}_0^\dagger$ during 1982-2011

Population	FRA_F	NLD_F	BEL_F	AUS_F	IRL_F	DEU_F	NOR_F	PRT_F
$\left\|\bar{e}_{0,FRA_F}^\dagger - \bar{e}_j^\dagger\right\|$	-	0.051	0.076	0.088	0.115	0.131	0.227	0.240
Population	AUT_F	ITA_F	FIN_F	SWE_F	UK_F	JPN_F	CHE_F	ESP_F
$\left\|\bar{e}_{0,FRA_F}^\dagger - \bar{e}_j^\dagger\right\|$	0.276	0.334	0.367	0.408	0.412	0.444	0.455	0.489
Population	CAN_F	SWE_M	NLD_M	NZL_F	DNK_F	JPN_M	NOR_M	IRL_M
$\left\|\bar{e}_{0,FRA_F}^\dagger - \bar{e}_j^\dagger\right\|$	0.502	0.533	0.660	0.721	0.725	0.932	0.962	0.963
Population	CHE_M	UK_M	ITA_M	AUS_M	DEU_M	BEL_M	USA_F	DNK_M
$\left\|\bar{e}_{0,FRA_F}^\dagger - \bar{e}_j^\dagger\right\|$	1.086	1.117	1.169	1.270	1.339	1.421	1.424	1.453
Population	CAN_M	AUT_M	ESP_M	NZL_M	FIN_M	FRA_M	PRT_M	USA_M
$\left\|\bar{e}_{0,FRA_F}^\dagger - \bar{e}_j^\dagger\right\|$	1.478	1.571	1.612	1.620	1.776	2.137	2.430	2.749

Note: Country codes are listed in section 3. M and F indicate males and females respectively.

The full series of forecast accuracies are plotted in Figure 2, where it is clearly seen that for all three measures and both methods there is an early improvement in accuracy as populations are added. However, when using the $LL_{e^\dagger}$ method the addition of the tenth population, Italian Females (Table 2), reverses this trend for all three accuracy measures. Thereafter, the accuracy level remains relatively steady as more populations are added. Thus, all measures indicate the same best reference group for French Females, namely the group with nine populations. The ninth population also identifies the threshold level of $\left|\bar{e}_i^\dagger - \bar{e}_j^\dagger\right|$ corresponding to the best reference group for French Females, as shown in Figure 2.

For LL, a slightly larger best reference group is indicated, and there is less agreement among accuracy measures. Different numbers of populations as best reference group from different accuracy measures were found for many other populations of interest for both methods. In several cases, we found that $ME(e_0)$ diminishes after adding many populations even though those populations are not so close to the population of interest in terms of observed mortality. Adding such populations does not have a large impact on MAE or MSE (Figure 2). The best reference populations obtained for French Females distance of $\bar{e}^\dagger$ with other populations are shown in the Appendix.

For French Females, the $ME(e_0)$ showed a sharp rise after adding Italian Females to the reference group. The $\bar{e}^\dagger$ for French Females was closer to that for populations in the reference group prior to the addition of Italian Females (see Table 3 for details). The observed e_0 of French and Italian Females along with forecast of e_0 from both LL and $LL_{e^\dagger}$ are plotted in Figure 3. Both for LL and $LL_{e^\dagger}$, the reference group consists of 9 populations which includes up to Austrian Females (see Table 3 for details). The observed life expectancies for both of the populations were close to each other until 2004.is visible afterwards and French e_0 is closer to the LL forecast than to the $LL_{e^\dagger}$ forecast. Further insight into the reference populations for French females can be obtained from the trend in $e^\dagger$ for some of the populations in the possible best reference group (from Table 2), seen in Figure 4.

We did not fix the number of populations in best reference group prior to model fitting. To determine the best reference group, we analyzed the forecast accuracy both for Li and Lee (2005) and the proposed method ($LL_{e^\dagger}$). As reference group, a combination of countries which produces lowest forecast error during out-of-sample evaluation is chosen. Details of forecast accuracies for proposed methods are discussed in the next section; here we discuss only the findings for French Females for illustration of identification of the best reference group. The forecast accuracy for French Females is given in Table 3. We added populations one at a time into the reference group with previous combination for both $LL_{e^\dagger}$ and LL and used same weight for both methods. The forecast accuracy for French Female during hold-out period is plotted in Figure 2.

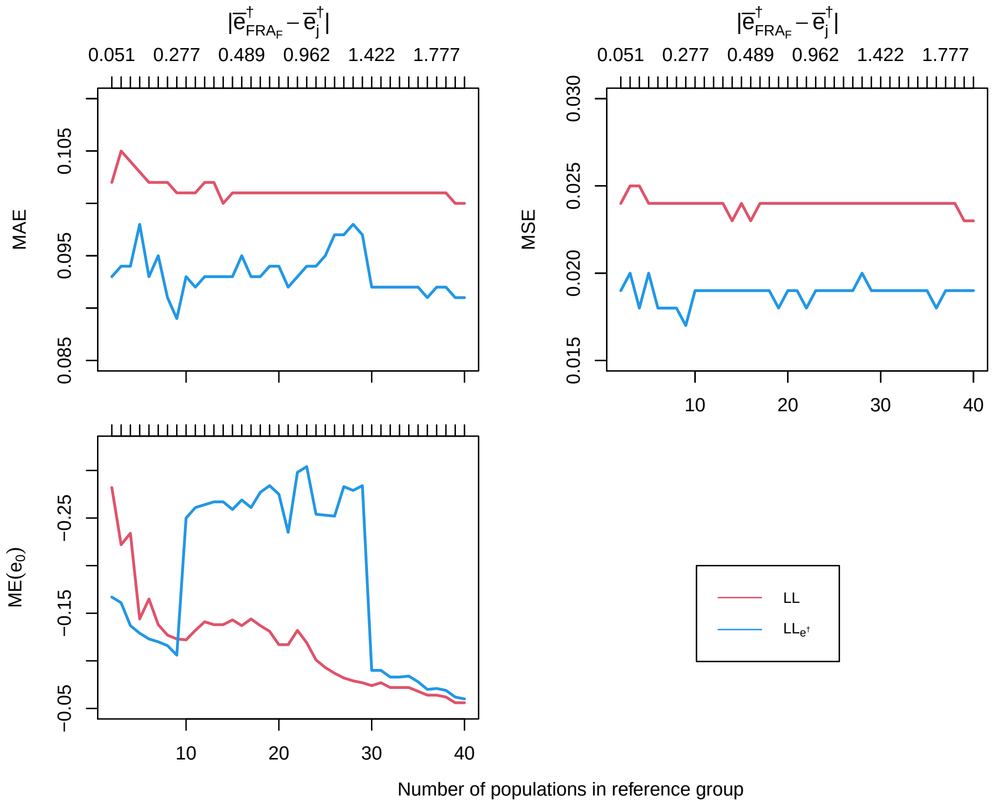
Table 3: Comparison of forecast accuracy for French female during out of sample evaluation period (2002-2011)

Reference population	MAE LL	MAE $LL_{e^\dagger}$	MSE LL	MSE $LL_{e^\dagger}$	$ME(e_0)$ LL	$ME(e_0)$ $LL_{e^\dagger}$	Best fitting period ($LL_{e^\dagger}$)
FRA_F+NLD_F	0.103	0.093	0.024	0.019	-0.282	-0.167	1956-2001
Above+BEL_F	0.105	0.094	0.025	0.020	-0.222	-0.161	1956-2001
Above+AUS_F	0.105	0.094	0.025	0.018	-0.235	-0.137	1979:2001
Above+IRL_F	0.103	0.098	0.024	0.020	-0.144	-0.129	1979:2001
Above+DEU_F	0.102	0.093	0.024	0.018	-0.165	-0.123	1977:2001
Above+NOR_F	0.102	0.095	0.024	0.018	-0.138	-0.120	1979:2001
Above+PRT_F	0.101	0.091	0.024	0.018	-0.127	-0.116	1974:2001
Above+AUT_F	0.101	**0.089**	0.023	**0.017**	-0.124	**-0.105**	1974:2001
Above+ITA_F	0.101	0.093	0.024	0.019	*-0.123*	-0.250	1964:2001
Above+FIN_F	0.101	0.092	0.024	0.019	-0.133	-0.261	1960:2001
Above+SWE_F	0.102	0.093	0.024	0.019	-0.141	-0.264	1960:2001
Above+UK_F	0.102	0.093	0.024	0.019	-0.139	-0.267	1960:2001
Above+JPN_F	*0.100*	0.093	*0.023*	0.019	-0.138	-0.212	1957:2001
Above+CHE_F	0.101	0.093	0.024	0.019	-0.144	-0.208	1957:2001
Above+ESP_F	0.101	0.095	0.024	0.019	-0.138	-0.201	1957:2001
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Note: Bold texts are used for showing the lowest errors obtained by$LL_{e^\dagger}$, while italic texts are used for showing lowest error obtained by LL.

In Figure 2, for all three measures of forecast accuracy there occurs reduction (increment) in accuracy (error) level until adding certain countries for both methods. We also showed the level of $\left|\bar{e}_i^\dagger - \bar{e}_j^\dagger\right|$ for French Females in Figure 2. For both of the methods, after adding some certain number of populations in reference group, the accuracy (error) level becomes steady. A last drop after which the error level becomes steady also plays the role of a threshold level of $\left|\bar{e}_i^\dagger - \bar{e}_j^\dagger\right|$ to choose the best reference group for a particular population i. This threshold in level of forecast accuracy is important, because it clearly shows which countries belongs to the best reference group for a population i. All three measures of forecast accuracy indicated the same best reference group for French females in the case of $LL_{e^\dagger}$. For LL, MAE and MSE indicate the same best reference group, whereas different best reference group was indicated by $ME(e_0)$. Different combinations of populations as best reference group from different measures of forecast accuracies are also observed for many other populations. For several other populations, we observed that $ME(e_0)$ diminishes after adding many populations even though those populations are not so close to that of interest in terms of observed mortality level. Adding these populations with huge gap in level of population-specific $\bar{e}^\dagger$ does not have high impact on

Figure 2: Forecast accuracies for French Females by number of populations in reference group by forecast method

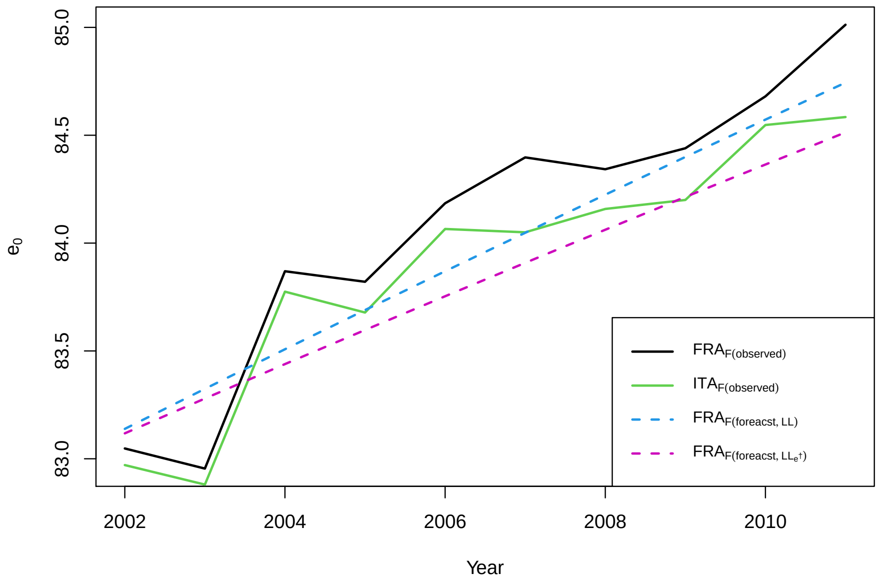


MAE or MSE (Figure 2). The best reference populations obtained for French Females according to distance of $\bar{e}^{\dagger}$ with other population is attached in the appendix.

For French Females, the ME(e_0) showed a sharp rise after adding Italian Females into the reference group. The $\bar{e}^{\dagger}$ for French Females was closer to populations added in reference group prior to Italian Females (see Table 3 for details). The observed e_0 of French and Italian Females along with forecast of e_0 from both LL and $\text{LL}_{e^{\dagger}}$ are plotted in Figure 3. Both for LL and $\text{LL}_{e^{\dagger}}$, the reference group consists of 10 populations which includes up to Austrian Females (see Table 3 for details). The observed life expectancies for both of the populations were close to each other until 2004. Irregular divergence is visible afterwards and French e_0 is closer to the LL forecast than to the $\text{LL}_{e^{\dagger}}$ forecast. Further insight into the reference populations for French females can be obtained from the trend in $e^{\dagger}$ for some of the populations in the possible best reference group (from Table 2), seen in Figure 4.

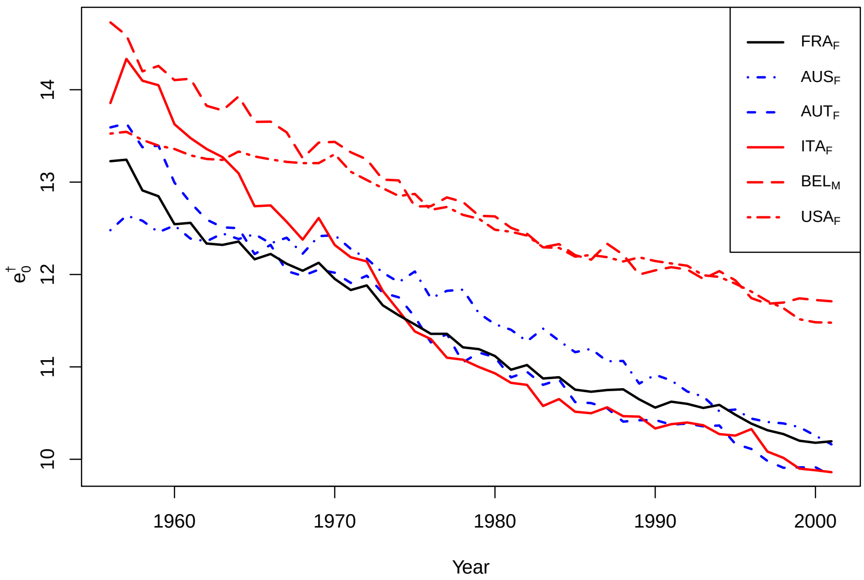
For Italian Females, observed trend of $e_0^{\dagger}$ during the hold-out period is close to other populations in the best reference group for French Females, but the past trend was quite different. This explains the change in forecast accuracy after adding many populations. For instance, Belgian Males are ranked as 29-th closest population to French Females in terms of lifespan disparity with a much higher level of mortality (Table 2). Combining many such divergent populations in the reference group finally increases accuracy because of using the same weight in the common factor model, but, this attained without any similarities in mortality patterns. Although adding a large number of countries increases the level of accuracy for life expectancy, it does not change MAE or MSE substantially from the values obtained for a smaller number of populations in the reference group.

Figure 3: Observed and forecast of e_0 for French Females (2002-2011)



Note: Observed Italian Females e_0 are also added for comparison.

Figure 4: Trend of $e^{\dagger}$ for French Females and some other populations (1956-2001)

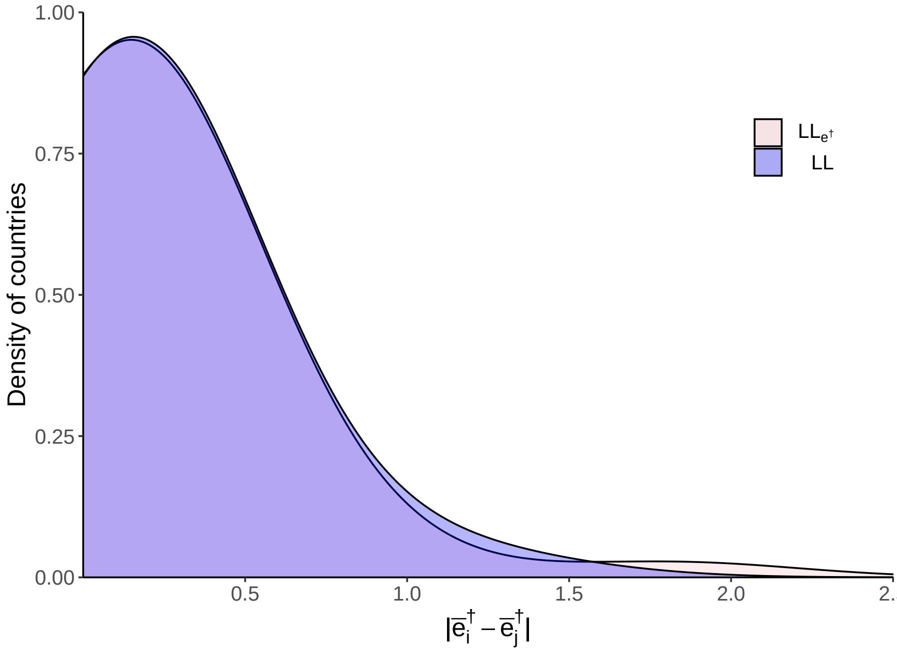


Note: The populations in blue lines are later found to be in best reference group for French Females and red are for those populations which are not in best reference group for $LL_{e^{\dagger}}$.

4.2 Size of best reference group

Unlike previous approaches we did not restrict the number of populations in the best reference group; rather we choose to observe the number of populations in the reference group for which the forecast accuracy is highest. From empirical analysis, we obtained different results for LL and $LL_{e^\dagger}$. The best reference groups obtained from LL and $LL_{e^\dagger}$ using different measures of forecast accuracy for all 40 appear in the Appendix. The distribution of number of populations in the best reference group according to $\left|\bar{e}_i^\dagger - \bar{e}_j^\dagger\right|$ is plotted in Figure 5, based on the three measures of forecast accuracy together as different measures of forecast accuracy indicate different populations as the best reference group.

Figure 5: Distribution of countries obtaining reference population according to $\left|\bar{e}_i^\dagger - \bar{e}_j^\dagger\right|$.



Note: We plotted a kernel density instead of histogram as it shows the skewness properly.

In the case of same accuracy level for two or more combinations, the combination having smaller number of countries is chosen as best reference group for a parsimonious model. The optimal number of populations in reference group and corresponding differences in $\left|\bar{e}_i^\dagger - \bar{e}_j^\dagger\right|$ considering all forecast errors are summarized in Table 4. The errors are considered separately for males, females and both sexes together.

Table 4: Summary statistics for best reference group

	Summary statistics	Male	$LL_{e^\dagger}$ Female	All	Male	LL Female	All
Number of populations	Median	5	5	5	4	4	4
	IQR	6	4	5	4	4	4
Difference in $\left\|\bar{e}_i^\dagger - \bar{e}_j^\dagger\right\|$	Median	0.15	0.14	0.15	0.15	0.09	0.12
	IQR	0.26	0.20	0.24	0.26	0.20	0.25

Note: IQR stands for interquartile range. We choose it over other measure of dispersion as it gives more idea about spread of the distribution in this context.

4.3 Forecast accuracy

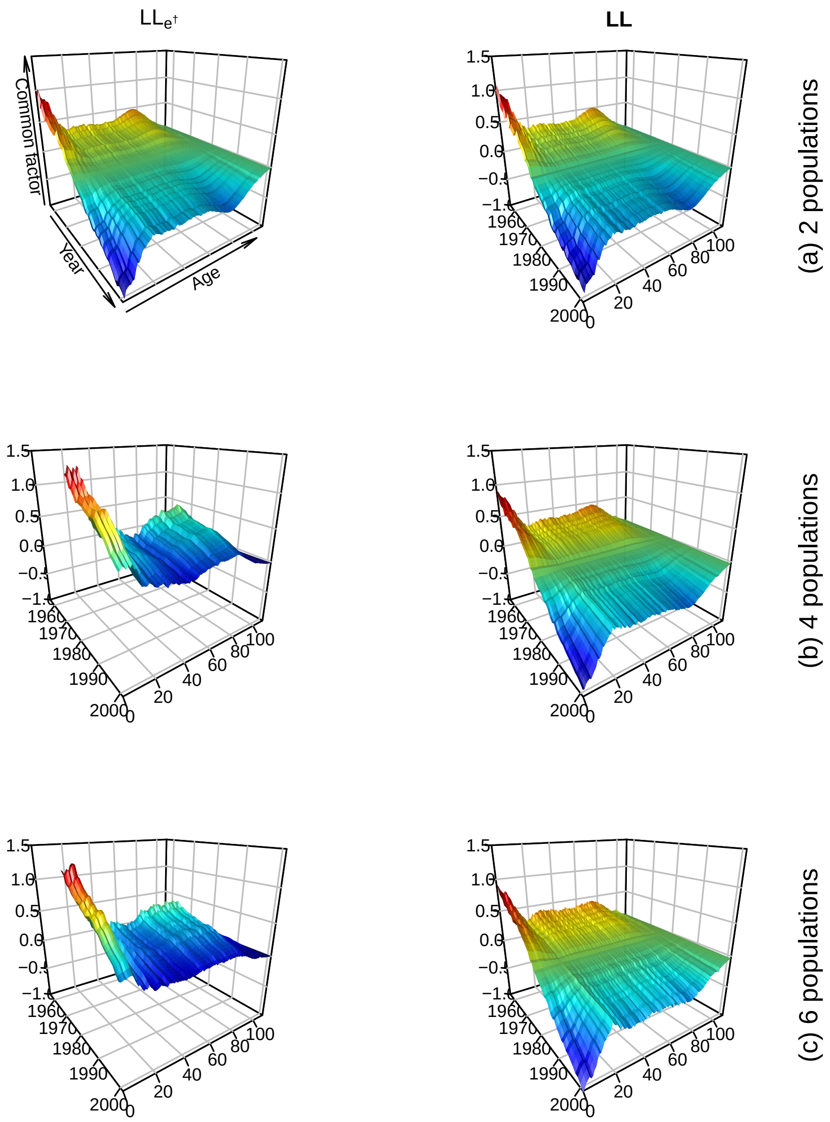
The comparison of different measures of forecast accuracy during hold-out period is given in Table 5. For both methods, we obtained forecast accuracy through out-of-sample evaluation, and from the three possible best reference groups we use the one with the smallest number of populations for our coherent forecast. $\mathrm{LL}_{e^\dagger}$ returns lower MAE and MSE than LL for all populations except for US Females. It has been already mentioned before that different measures of forecast accuracy often leads to different best reference groups. The reason of better accuracy of $\mathrm{LL}_{e^\dagger}$ is already explained in Figure 3 and 4; inclusion of population(s) with close lifespan disparity produce higher forecast accuracy for a population of interest. The proposed modifications produce different common factor for $\mathrm{LL}_{e^\dagger}$ than that of LL. The common factor obtained from $\mathrm{LL}_{e^\dagger}$ and LL for French Females are plotted in Figure 6 to get more insight of the proposed modifications. Here we considered 3 different reference groups consisting 2, 4 and 6 populations respectively. The reference groups are constructed according to closest lifespan disparity as before.

For small number of populations in reference group, the effect of adjusting K_t according to $e^\dagger$ is highly visible (Figure 6a). However, for group consisting large number of populations (10 or more) clearly the best fitting period is responsible for higher forecast accuracy in the case of $\mathrm{LL}_{e^\dagger}$ (Figure 6b,c). However, with adding more populations to the reference group, this best fitting period slowly shifts to full fitting period eventually. The reference group consisting of all 40 low-mortality populations considers full observational period (1956-2001) as best fitting period in $\mathrm{LL}_{e^\dagger}$.

Table 5: Comparison of minimum values of different measures of forecast accuracy for female populations of 20 low-mortality countries during hold-out period (2002-2011)

	MAE		MSE		ME(e_0)	
Country	LL	$\mathrm{LL}_{e^\dagger}$	LL	$\mathrm{LL}_{e^\dagger}$	LL	$\mathrm{LL}_{e^\dagger}$
Australia	0.121	*0.096*	0.038	*0.022*	-0.011	*-0.004*
Austria	0.193	*0.156*	0.100	*0.066*	0.007	*0.003*
Belgium	0.153	*0.135*	0.070	*0.053*	0.002	-0.009
Canada	0.080	*0.068*	0.016	*0.013*	-0.042	*-0.041*
Denmark	0.231	*0.200*	0.129	*0.102*	-0.672	*-0.542*
Finland	0.223	*0.199*	0.125	*0.080*	-0.017	-0.072
France	0.100	*0.089*	0.023	*0.017*	-0.123	*-0.106*
Germany	0.087	*0.082*	0.017	*0.014*	0.049	0.082
Ireland	0.251	*0.242*	0.143	*0.130*	-0.998	-1.063
Italy	0.085	*0.078*	0.019	*0.015*	-0.005	-0.032
Japan	0.120	*0.116*	0.034	*0.031*	0.431	0.465
Netherlands	0.149	*0.138*	0.068	*0.056*	-0.516	*0.514*
New Zealand	0.224	*0.184*	0.118	*0.838*	-0.298	-0.301
Norway	0.230	*0.188*	0.150	*0.100*	-0.036	-0.217
Portugal	0.174	*0.140*	0.074	*0.047*	-0.597	*-0.281*
Spain	0.100	*0.096*	0.023	*0.021*	0.008	0.011
Sweden	0.169	*0.149*	0.079	*0.063*	-0.037	*-0.012*
Switzerland	0.222	*0.165*	0.161	*0.074*	0.105	0.105
United Kingdom	0.079	*0.069*	0.014	*0.010*	-0.343	*-0.331*
USA	0.054	0.054	0.005	0.005	-0.156	-0.221

Figure 6: Common factor of the fitted models with different sizes of reference group

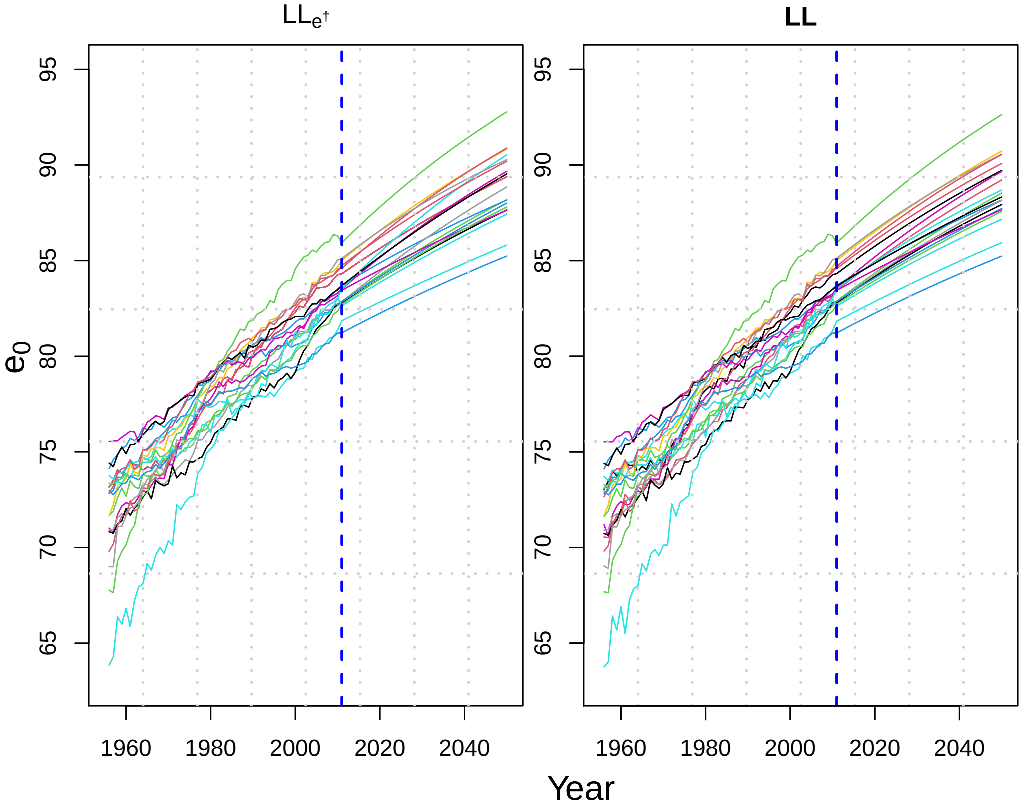


Since LL$_{e^\dagger}$ produces lower error than LL for all combinations, the population-specific lowest forecast error for both LL$_{e^\dagger}$ and LL are shown in Table 5. LL$_{e^\dagger}$ and LL produced same level of MAE and MSE for US Females. For ME(e_0), LL$_{e^\dagger}$ performed better with 9 populations. Unusual rise in ME(e_0) is observed in the case of LL$_{e^\dagger}$ for several other countries, for all of them we noticed same pattern as observed for French Females (Figure 4). One important feature of the proposed LL$_{e^\dagger}$ is the concept of best fitting period. We did not find best fitting period for all of the combinations for French Females (Table 3). The best reference group (bold texts in Table 3) obtained for French Females consists of 9 populations and it has the best fitting period for 1974 to 2001. Best fitting period corresponding to best reference groups from all these measures of forecast accuracies are shown in the appendix (appendix A3).

4.4 Forecast of life expectancy

The forecasts of female life expectancy at birth for all 20 low-mortality countries are plotted in Figure 7. For both of the methods, we checked forecast accuracy during out-of-sample evaluation and from three possible best reference groups we choose the one having lower number of populations to make coherent forecast.

Figure 7: Observed (1956-2011) and forecast (2012:2050) of female life expectancy at birth for 20 low-mortality countries



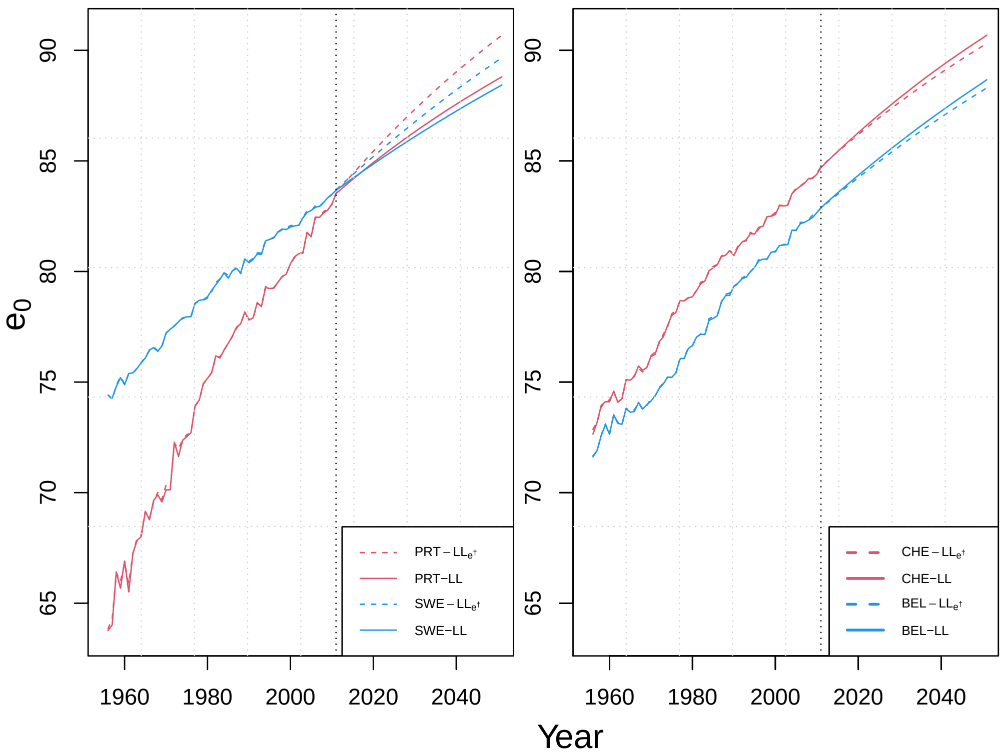
The coherent forecast of life expectancy at birth in 2050 for all of these populations are given in Table 6. Except for Australia, Belgium, Denmark, Ireland, Norway, Spain and Switzerland, the forecast of life expectancy were higher for $LL_{e^\dagger}$ than that of LL.

Table 6: Comparison of coherent forecast of life expectancy at birth in 2050 for female populations of 20 low-mortality countries

Country	$LL_{e^\dagger}$	LL	Country	$LL_{e^\dagger}$	LL
Australia	89.382	89.719	Japan	92.777	92.636
Austria	89.382	89.216	Netherlands	88.009	87.721
Belgium	88.184	88.538	New Zealand	87.432	87.159
Canada	88.166	88.166	Norway	87.636	87.662
Denmark	85.813	85.943	Portugal	90.537	88.692
Finland	89.668	89.665	Spain	90.278	90.567
France	90.824	90.730	Sweden	89.535	88.327
Germany	88.851	88.154	Switzerland	90.194	90.557
Ireland	87.649	87.934	United Kingdom	87.802	87.571
Italy	90.887	90.082	USA	85.237	85.235

To gain greater insight of the obtained results we further identified the populations for which $LL_{e^\dagger}$ have highest difference in forecast of life expectancy compare to LL, both from positive and negative point of view. For Portugal and Sweden the forecast obtained from $LL_{e^\dagger}$ have highest positive difference with that of LL whereas the highest negative difference were observed for Belgium and Switzerland. The forecast of these four countries are plotted in Figure 8.

Figure 8: Observed (1956-2010) and forecast (2012-2050) female life expectancy at birth for Portugal, Sweden, Switzerland and Belgium by coherent forecasting methods



Note: Among 20 low-mortality countries, differences by forecasting method ($LL_{e^\dagger}$ - LL) were greatest for Portugal and Sweden and smallest for Switzerland and Belgium.

Among these four countries, Portugal is mentioned before for remarkable improvement in health status and rapid increase of life expectancy (Van Oyen et al. 2013). Between 2000 and 2015 the female life expectancy increased by almost four years for Portugal, almost 5 years for males (HMD 2019). However, these improvements have not occurred at the same pace for different income groups and disparities exist for other important dimensions of health. Cardiovascular diseases and cancer are the largest contributors to mortality(Van Oyen et al. 2013). On the other hand, females of Sweden and Switzerland have steady pattern of mortality improvement since long (HMD 2019). However, this scenario is is not the case for Danish Females. During the past decades, the life expectancy of Danish women has lagged behind that of women in neighboring Western European countries (Jacobsen et al. 2002). Among various causes-of-deaths, ischaemic heart diseases followed by lung cancer are responsible for lower life expectancy of Danish Females.

4.5 Interval forecast

Following the results of forecast of life expectancy to 2050, the prediction interval of e_0 is plotted in Figure 9 for Females in Portugal, Sweden, Switzerland and Belgium. For Sweden and Belgium, the prediction interval of the $\text{LL}_{e^\dagger}$ is slightly wider than that of LL, whereas prediction interval from LL are slightly wider for Portugal and Switzerland. Interval forecasts for all female populations are shown in the Appendix Interval forecast for other female populations are added in Appendix (appendix A2).

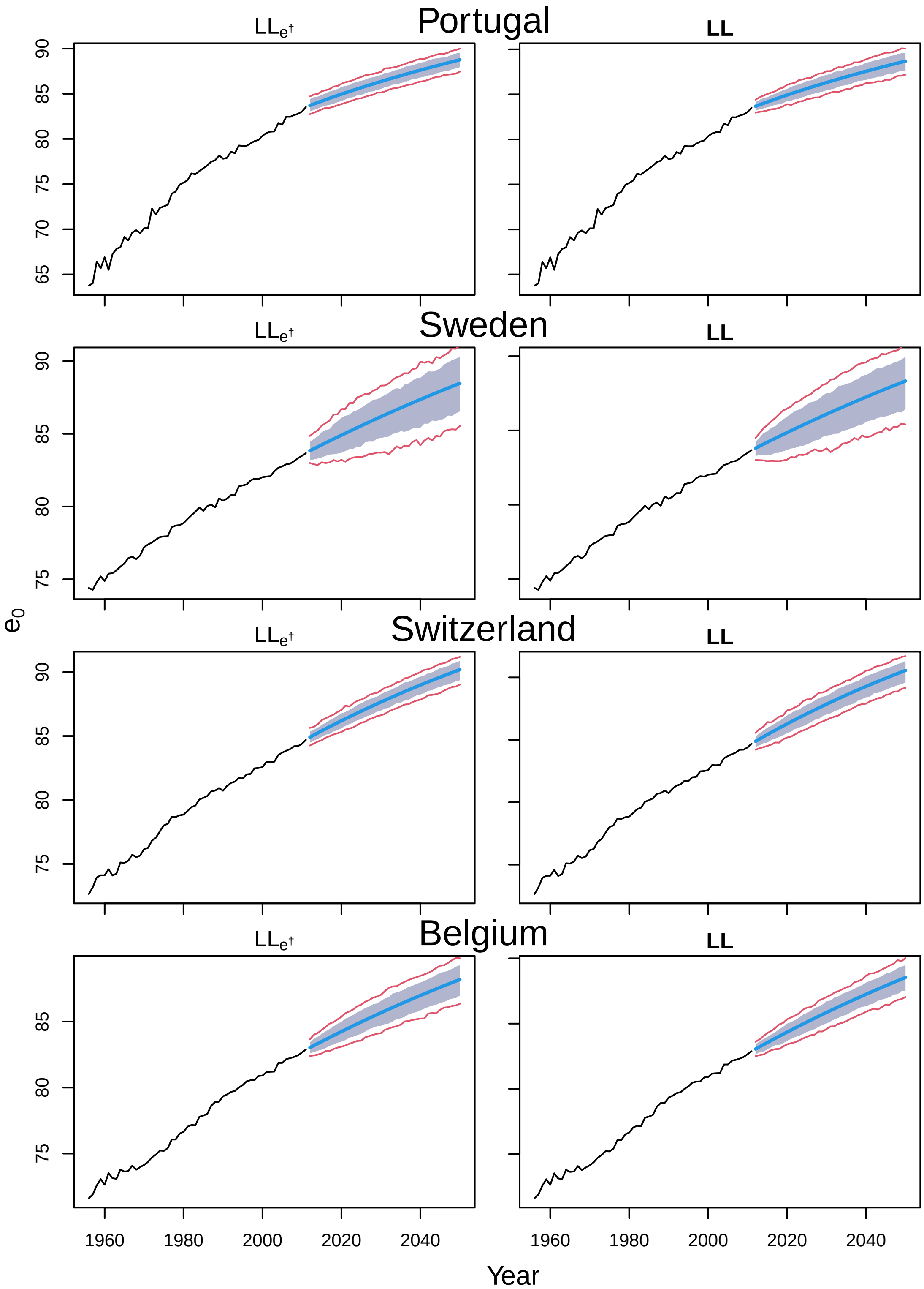
5 Discussion and Conclusion

In this line of research on Lee-Carter framework, we introduced a new methodology for coherent forecasting. Choosing the appropriate reference group is an old puzzle for coherent forecasting and different reference populations bring about quite different results (Kjærgaard et al. 2016). We addressed this problem by proposing a robust definition of reference population based on closest lifespan disparity among populations. This definition is found to be applicable for existing coherent forecasting technique as well. Following Rabbi and Mazzuco (2021), our proposed coherent method incorporates lifespan disparity during parameter estimation and application of LASSO-type smoothing prior to fitting the model. In identifying the reference group of populations on the basis of closest lifespan disparity and forecast accuracy, our method also determines the best fitting period for this reference group. Compared with the LL method, the proposed method produced more accurate forecasts during out-of-sample evaluation and produced more optimistic forecasts of life expectancy at birth.

The novel contribution of the proposed method is the objective determination of the reference group of populations that jointly comprise the reference population. In defining the reference group for a population of interest, we considered populations of both sexes equally. Consideration of male and female populations together for coherent forecasting in same reference group is a topic of debate due to different age patterns of mortality. There are two responses regarding this issue. First, during sorting of populations with respect to $\bar{e}_0^\dagger$ to identify the closest, male and female populations were separated naturally with very little overlap of values from the value of Including a different-sex population in the reference group occurred for only a few populations. Second, most of these countries implement the same policy for both sexes regarding old-age health care, as the issues of aging are common to both. Thus, in practice the procedure is not problematic.

In coherent forecasting, we applied equal weight on mortality rates of all populations to construct joint mortality matrix (both for $\text{LL}_{e^\dagger}$ and LL). This resolved the problem of mixing population with large exposure with smaller one, however, using equal weight also has consequences. Different pattern of age-specific mortality rates among different populations are result of both exposure size and different distribution of causes-of-deaths. Defining an appropriate population-specific weight to adjust the problem of different distribution of causes-of-deaths will be complicated because different data sources will be needed to obtain harmonized data for causes-of-death. Although the low-mortality countries are converging in terms of longevity, still, each of these populations are distinct in terms of age-specific mortality rates (HMD 2019). Equal weight also reduces the benefit of lasso in case of reference group consists of larger number of populations. Another limitation of the proposed method is that the interval forecast of life expectancy at birth is narrow for several of the populations (for both methods). Although $\text{LL}_{e^\dagger}$ produced slightly wider prediction interval than LL in many cases, still, this is an old criticism regarding Lee-Carter variants (Hyndman and Ullah 2007). In the proposed model it happened

Figure 9: Observed (1956-2010) and forecast (2012-2050) female life expectancy at birth with prediction intervals for Portugal, Sweden, Belgium and Switzerland by coherent forecasting method



Note: The blue area represents 80% prediction interval and red lines are for 95% prediction interval.

due to application of smoothing and new adjustment technique. This adjustment makes the time component more linear for most of the populations and consequently reduces the variance of the ARIMA model. In addition, variance of the model is also lower in the proposed method due to smoothing.

As most of the populations do not have data after age 100 for several years, we extrapolated the mortality rates of age 100:110+ using Kannisto model (Thatcher et al. 1998). Although we obtained missing mortality rates keeping the fitted rates at the closest distance to observed data, still it reduced the variability across the populations in centenarian mortality (HMD 2019). Also, it is clear that smoothing can create significant differences in mortality forecasting (Figure 6), particularly for smaller reference groups. For coherent forecasting model, we separately smoothed the populations using Lasso and then combined them in common factor model. Instead of smoothing prior to combine the mortality rates, applying Lasso after combining populations may produce different results. However, smoothing after combining populations brings about a higher computational complexity because in this way Lasso should be run every time a new population is added in reference group to check the forecast accuracy.

Based on the findings of the current research, we might point out some future scope of research on mortality forecasting. The first issue will be to overcome to problem of invariant b_x in Lee-carter framework. One possible solution to do that is to adopt Bayesian approaches on parameter estimation. Secondly, the incorporation of a cohort effect might provide more insight into the mortality of a population. This has been done already in the LC method, but still nothing has been done in coherent settings. Thirdly, we introduced a new systematic approach to obtain best reference group for a population. Although $e_0^\dagger$ better reflects the distribution of death for a population, there are still many other demographic indicators are subject to analyze to obtain the best reference group. Fourthly, although Lasso is found to be more effective smoothing technique for our data, it is a slightly time consuming method. Faster algorithm for getting optimal results for Lasso will be helpful. Fifthly, it is wise for coherent forecasting to consider life table of longer time series and longer life span (say 120+ years). Future research on mortality forecasting may consider to make a more generalized definition for length of lifespan (in terms of upper limit of lifespan) to be considered for coherent setup.

References

Aburto, J. M. and van Raalte, A. (2018). Lifespan dispersion in times of life expectancy fluctuation: the case of central and eastern europe. *Demography*, 55(6):2071–2096.

Aburto, J. M., Villavicencio, F., Basellini, U., Kjærgaard, S., and Vaupel, J. W. (2020). Dynamics of life expectancy and life span equality. *Proceedings of the National Academy of Sciences*, 117(10):5250–5259.

Ahcan, A., Medved, D., Olivieri, A., and Pitacco, E. (2014). Forecasting mortality for small populations by mixing mortality data. *Insurance: Mathematics and Economics*, 54:12–27.

Ahmadi, S. S. and Li, J. S.-H. (2014). Coherent mortality forecasting with generalized linear models: A modified time-transformation approach. *Insurance: Mathematics and Economics*, 59:194–221.

Bergeron-Boucher, M.-P., Canudas-Romo, V., Oeppen, J., and Vaupel, J. W. (2017). Coherent forecasts of mortality with compositional data analysis. *Demographic Research*, 37:527–566.

Bohk-Ewald, C., Ebeling, M., and Rau, R. (2017). Lifespan disparity as an additional indicator for evaluating mortality forecasts. *Demography*, 54(4):1559–1577.

Booth, H. (2020). Coherent mortality forecasting with standards: Low mortality serves as a guide. In *Developments in Demographic Forecasting*, pages 153–178. Springer, Cham.

Booth, H., Maindonald, J., and Smith, L. (2002). Applying lee-carter under conditions of variable mortality decline. *Population studies*, 56(3):325–336.

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

Brouhns, N., Denuit, M., and Vermunt, J. K. (2002). A poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and economics*, 31(3):373–393.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., and Khalaf-Allah, M. (2011). Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin: The Journal of the IAA*, 41(1):29–59.

Dokumentov, A. and Hyndman, R. (2017). Smoothing of two-dimensional demographic data, optionally taking into account period and cohort effects.

Dokumentov, A., Hyndman, R. J., and Tickle, L. (2018). Bivariate smoothing of mortality surfaces with cohort and period ridges. *Stat*, 7(1):e199.

Girosi, F. and King, G. (2006). *Demographic Forecasting*. Cambridge University Press, Cambridge.

HMD (2019). *Human Mortality Database*. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).

Hyndman, R. J. and Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, 24(3):323–342.

Hyndman, R. J., Booth, H., Tickle, L., and Maindonald, J. (2011). Demography: Forecasting mortality, fertility, migration and population data. *R package version*, 1.

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

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

Jacobsen, R., Keiding, N., and Lynge, E. (2002). Long term mortality trends behind low life expectancy of danish women. *Journal of Epidemiology & Community Health*, 56(3):205–208.

Kjærgaard, S., Canudas-Romo, V., and Vaupel, J. W. (2016). The importance of the reference populations for coherent mortality forecasting models. In *European Population Conference*, Mainz, Germany.

Lee, R. (2000). The lee-carter method for forecasting mortality, with various extensions and applications. *North American actuarial journal*, 4(1):80–91.

Lee, R. and Miller, T. (2001). Evaluating the performance of the lee-carter method for forecasting mortality. *Demography*, 38(4):537–549.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671.

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

Rabbi, A. M. F. and Mazzuco, S. (2018). Mortality and life expectancy forecast for (comparatively) high mortality countries. *Genus*, 74(1):18.

Rabbi, A. M. F. and Mazzuco, S. (2021). Mortality forecasting with the lee–carter method: Adjusting for smoothing and lifespan disparity. *European Journal of Population*, 37(1):97–120.

Renshaw, A. E. and Haberman, S. (2000). Modelling the recent time trends in uk permanent health insurance recovery, mortality and claim inception transition intensities. *Insurance: Mathematics and Economics*, 27(3):365–396.

Santos, J. V., Lobo, M., Neiva, R. M., Viana, J., Souza, J., Dias, C. C., Cylus, J., Ricciardi, W., and Freitas, A. (2020). European union state of health from 1990 to 2017: time trends and its enlargements effects. *International Journal of Public Health*, 65(2):175–186.

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

Thatcher, A. R., Kannisto, V., and Vaupel, J. W. (1998). The force of mortality at ages 80 to 120.

Van Oyen, H., Nusselder, W., Jagger, C., Kolip, P., Cambois, E., and Robine, J.-M. (2013). Gender differences in healthy life years within the eu: an exploration of the ”health–survival” paradox. *International journal of public health*, 58(1):143–155.

Vaupel, J. W. and Canudas-Romo, V. (2003). Decomposing change in life expectancy: A bouquet of formulas in honor of nathan keyfitz's 90th birthday. *Demography*, 40(2):201–216.

Vaupel, J. W., Zhang, Z., and van Raalte, A. A. (2011). Life expectancy and disparity: an international comparison of life table data. *BMJ open*, pages bmjopen–2011.

Vogt, T. C. and Missov, T. I. (2017). Estimating the contribution of mortality selection to the east–west german mortality convergence. *Population health metrics*, 15(1):33.

Wiśniowski, A., Smith, P. W., Bijak, J., Raymer, J., and Forster, J. J. (2015). Bayesian population forecasting: extending the lee-carter method. *Demography*, 52(3):1035–1059.

Zhang, Z. and Vaupel, J. W. (2009). The age separating early deaths from late deaths. *Demographic Research*, 20:721–730.

Appendix

A1. Old age mortality

Mortality rates are not available for some of the later age-groups for the whole fitting period. Considering shifting mortality of almost all the populations, we fitted Kannisto's model at later age group for coherent forecasting. For ages $x = 80, 81, \ldots 110+$, let the observed death counts are noted as D_x and exposure as E_x. We extrapolate the unavailable mortality rates for later age groups by fitting the Kannisto's model of old age mortality on observed death rates M_x to estimate the underlying hazards function μ_x as,

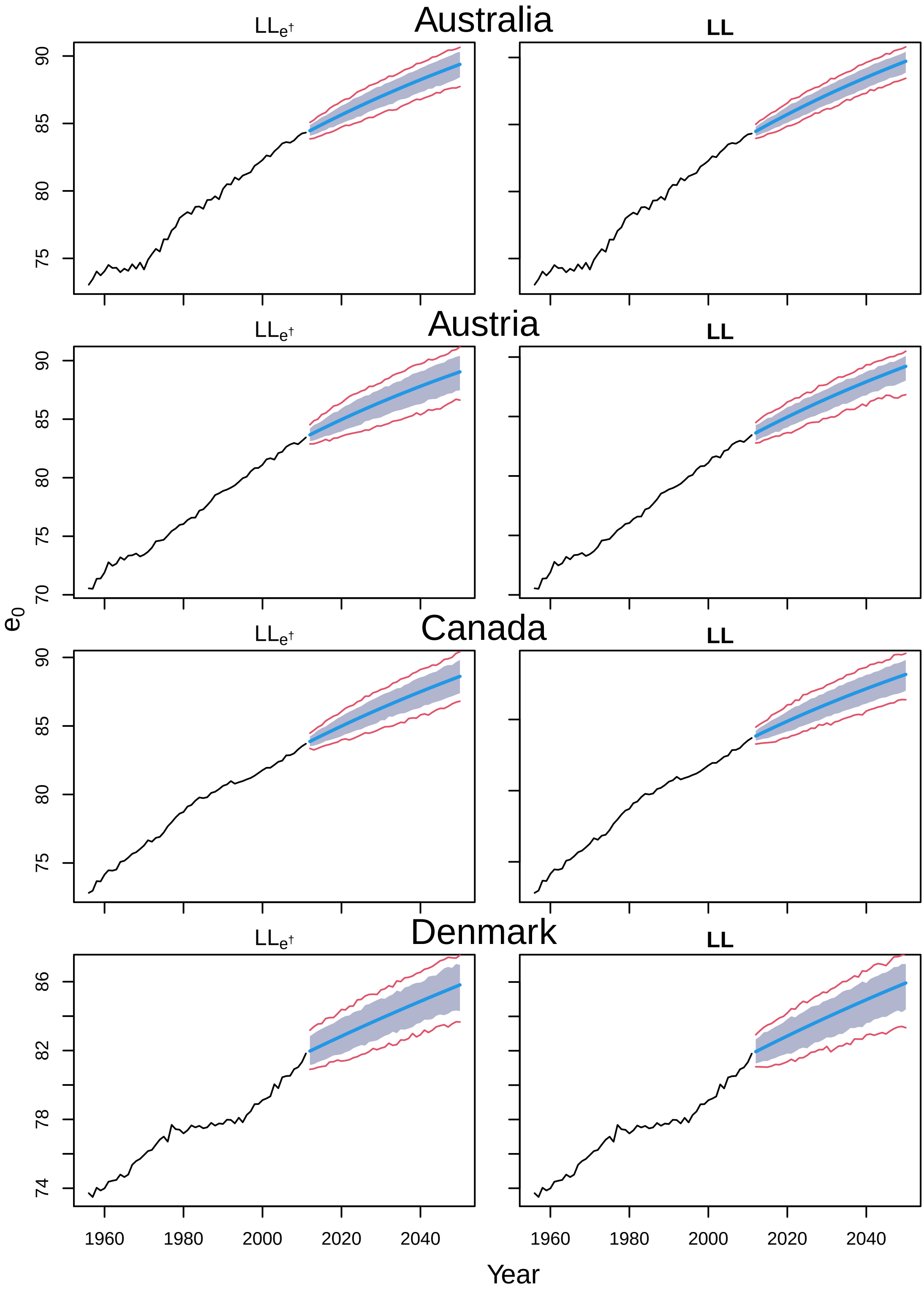
$$\mu_{x,(a,b)} = \frac{ae^{b(x-80)}}{1 + ae^{b(x-80)}}; \qquad a, b \geq 0.$$

Sensitivity of $e_0^\dagger$ due to Kannisto fitted mortality rates

We tried different combinations of fitting period and then added different combination of smoothed data with observed data. In this analysis we used the data obtained from fitting period at age 80:100 and adding the smoothed data of age 100:100+ with observed data till age 99. Among various combination we tried, this combination was the closest to real data and the difference of estimated and observed $e_0^\dagger$ during the fitting period (1956-2011) were lowest for this combination.

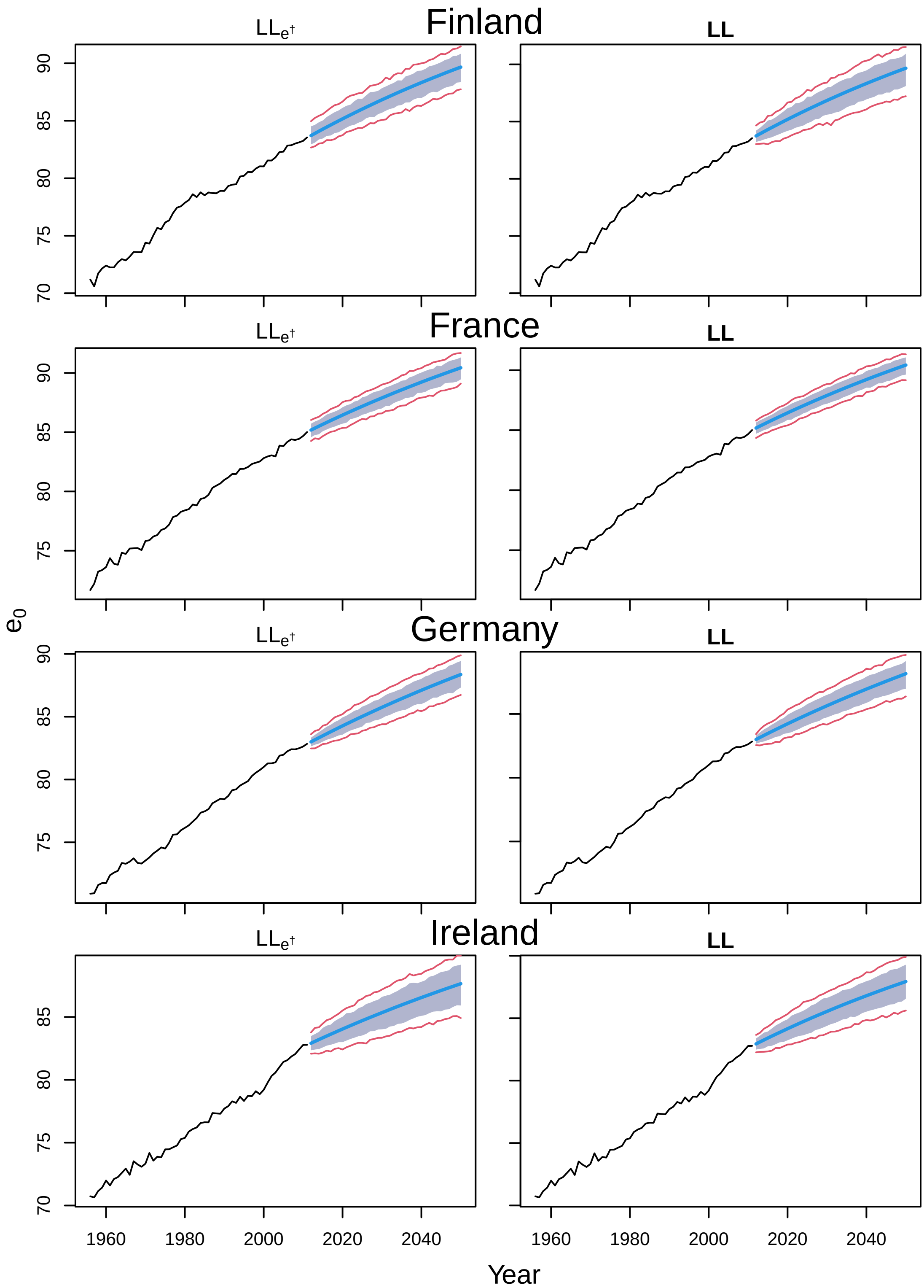
A2. Interval forecast of life expectancy

Figure 10: Prediction interval of female life expectancy at birth till 2050 for Australia, Austria, Canada and Denmark



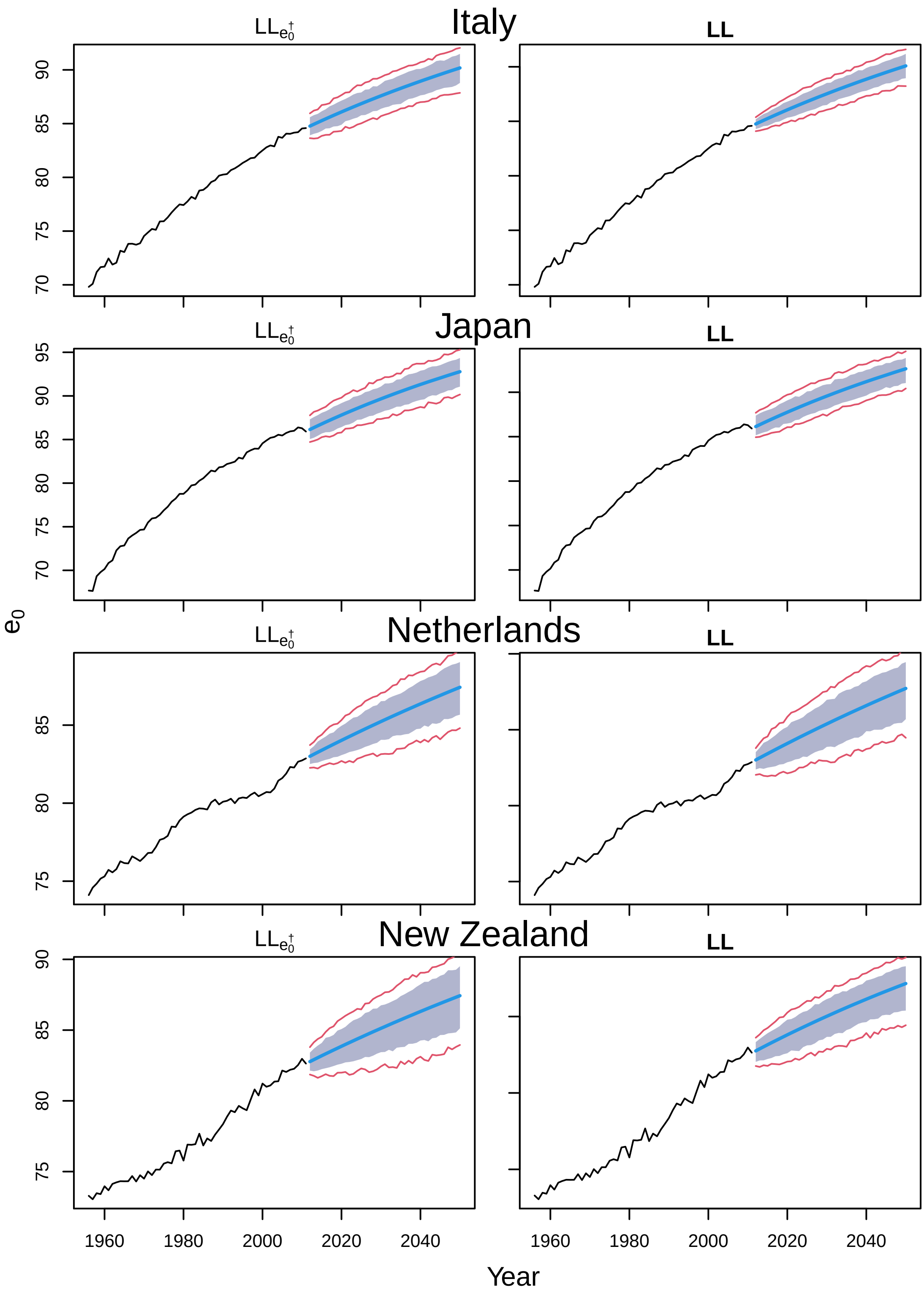
Note: The blue area represents 80% prediction interval and red lines are for 95% prediction interval.

Figure 11: Prediction interval of female life expectancy at birth till 2050 for Finland, France, Germany and Ireland



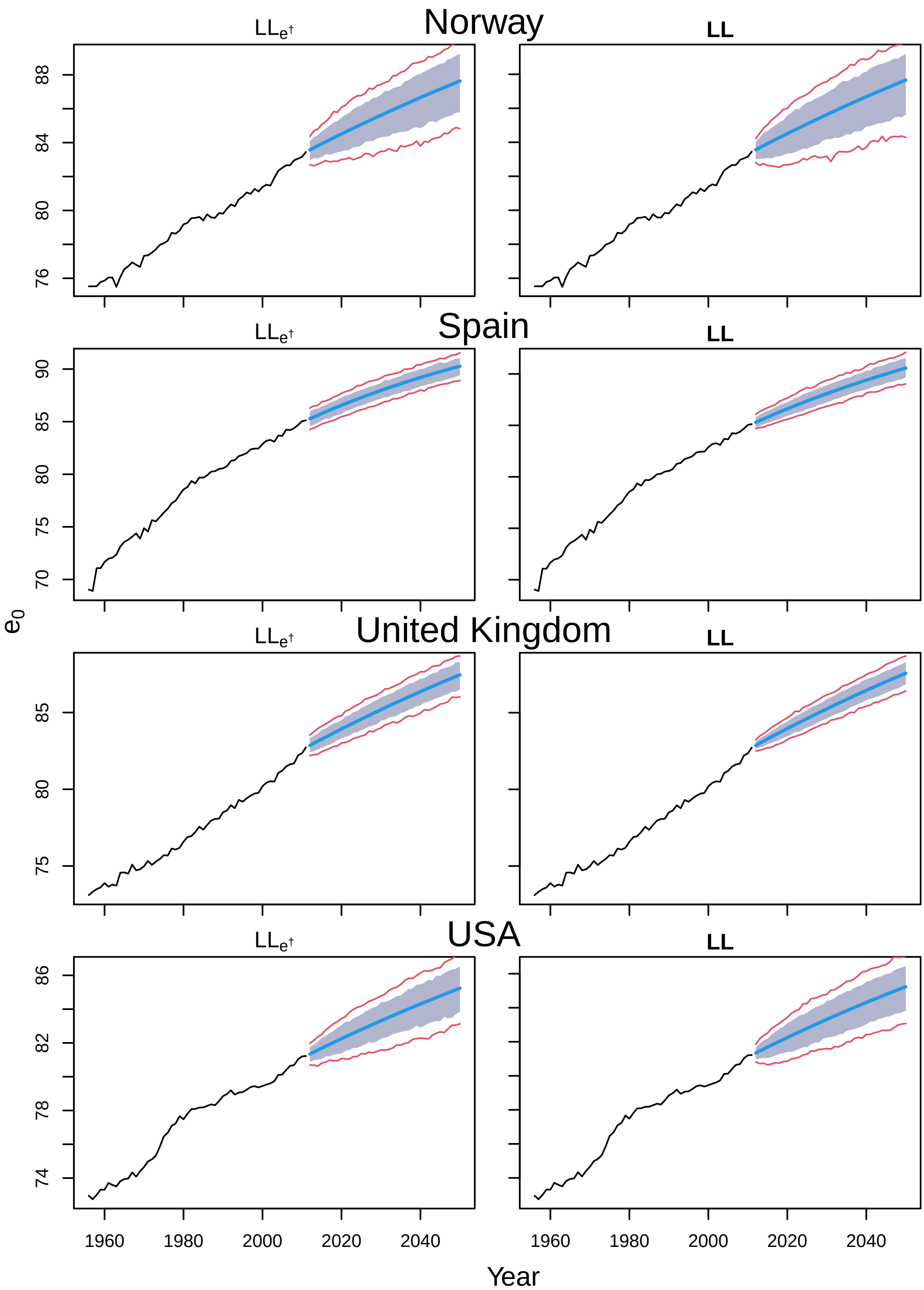
Note: The blue area represents 80% prediction interval and red lines are for 95% prediction interval.

Figure 12: Prediction interval of female life expectancy at birth till 2050 for Italy, Japan, Netherlands and New Zealand



Note: The blue area represents 80% prediction interval and red lines are for 95% prediction interval.

Figure 13: Prediction interval of female life expectancy at birth till 2050 for Norway, Spain, United Kingdom and USA



Note: The blue area represents 80% prediction interval and red lines are for 95% prediction interval.

A3. Best reference group for coherent forecasting

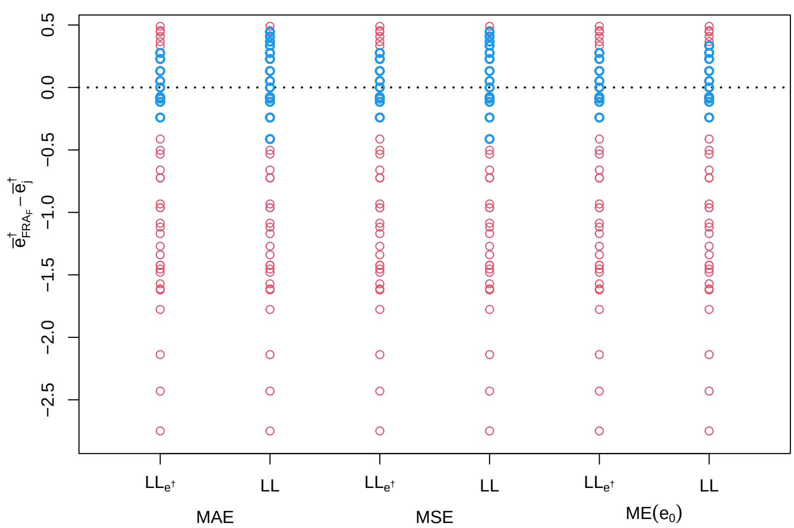
A different approximation is used in Table 1 as a sensitivity analysis of $e^{\dagger}$. Here we used the approximation mentioned by Aburto and van Raalte (2018) to estimate $e^{\dagger}$. Except for slight changes in the value, we obtained same order of the populations as we have in the Table 1 of the main paper. Figure 14 shows the concentrating nature the populations according to different measures of forecast accuracy. This figure signifies our choice of best reference group according to closest $\bar{e}^{\dagger}$.

Table 7: Sorted $\bar{e}^{\dagger}$ over the period 1982-2011 for the 20 low-mortality countries

Population	$\mathrm{ESP_F}$	$\mathrm{CHE_F}$	$\mathrm{JPN_F}$	$\mathrm{SWE_F}$	$\mathrm{FIN_F}$	$\mathrm{ITA_F}$	$\mathrm{AUT_F}$	$\mathrm{NOR_F}$
$\bar{e}^{\dagger}$	9.623	9.662	9.670	9.709	9.748	9.779	9.840	9.889
Population	$\mathrm{DEU_F}$	$\mathrm{NLD_F}$	$\mathrm{FRA_F}$	$\mathrm{BEL_F}$	$\mathrm{AUS_F}$	$\mathrm{IRL_F}$	$\mathrm{PRT_F}$	$\mathrm{UK_F}$
$\bar{e}^{\dagger}$	9.985	10.067	10.118	10.193	10.206	10.229	10.352	10.530
Population	$\mathrm{CAN_F}$	$\mathrm{SWE_M}$	$\mathrm{NLD_M}$	$\mathrm{NZL_F}$	$\mathrm{DNK_F}$	$\mathrm{JPN_M}$	$\mathrm{NOR_M}$	$\mathrm{IRL_M}$
$\bar{e}^{\dagger}$	10.621	10.653	10.774	10.839	10.847	11.052	11.076	11.080
Population	$\mathrm{CHE_M}$	$\mathrm{UK_M}$	$\mathrm{ITA_M}$	$\mathrm{AUS_M}$	$\mathrm{DEU_M}$	$\mathrm{BEL_M}$	$\mathrm{USA_F}$	$\mathrm{DNK_M}$
$\bar{e}^{\dagger}$	11.208	11.232	11.287	11.389	11.460	11.540	11.545	11.571
Population	$\mathrm{CAN_M}$	$\mathrm{AUT_M}$	$\mathrm{ESP_M}$	$\mathrm{NZL_M}$	$\mathrm{FIN_M}$	$\mathrm{FRA_M}$	$\mathrm{PRT_M}$	$\mathrm{USA_M}$
$\bar{e}^{\dagger}$	11.597	11.693	11.733	11.737	11.899	12.265	12.551	12.872

Note: Country codes are listed in section 3. M and F indicate males and females respectively.

Figure 14: Closest population to be included as reference for French Females according to different measures of forecast accuracy



Note: The horizontal black dotted line represents French Females. The Red dots are set available populations while blue dots represent the ones forming the best reference group in terms of highest forecast accuracy.

The best reference groups obtained from three different measures of forecast accuracy during out-of-sample evaluations are given in this section for LL and $LL_{e^\dagger}$. The combinations without mentioning best fitting period utilized the full available date in case of $LL_{e^\dagger}$. German male forecast accuracies were increasing indefinitely, so it is omitted for LL. For $LL_{e^\dagger}$, only the combinations with different best fitting period rather than full observed time are mentioned.

Table 8: Best reference group for males of low-mortality countries according to lowest MAE in LL

Country	Other populations in best reference group	$\bar{e}_i^\dagger - \bar{e}_j^\dagger$
Australia	DEU_M	0.06
Austria	ESP_M	0.04
Belgium	$USA_F, DNK_M, CAN_M, DEU_M, AUT_M, AUS_M$	0.15
Canada	10 populations	0.30
Denmark	CAN_M	0.025
Finland	NZL_M, ESP_M, AUT_M	0.204
France	$PRT_M, FIN_M, NZL_M, ESP_M, AUT_M$	0.56
Germany	-	-
Ireland	NOR_M	0.00048
Italy	UK_M, CHE_M, AUS_M	0.1
Japan	NOR_M	0.03
Netherlands	NZL_M, DNK_F, SWE_M	0.12
New Zealand	ESP_M, AUT_M	0.049
Norway	IRL_M	0.0004
Portugal	$FRA_M, USA_M, FIN_M, NZL_M, ESP_M, AUT_M$	0.858
Spain	NZL_M, AUT_M, CAN_M	0.13
Sweden	CAN_M	0.03
Switzerland	$UK_M, ITA_M, IRL_M, NOR_M, JPN_M, AUS_M$	0.184
UK	$CHE_M, ITA_M, AUS_M, IRL_M, NOR_M, JPN_M, DEU_M, BEL_M$	0.335
USA	PRT_M	0.31

Table 9: Best reference group for males of low-mortality countries according to lowest MSE in LL

Country	Other populations in best reference group	$\bar{e}_i^\dagger - \bar{e}_j^\dagger$
Australia	DEU_M	0.06
Austria	ESP_M	0.04
Belgium	$USA_F, DNK_M, CAN_M, DEU_M, AUT_M, AUS_M$	0.15
Canada	$DNK_M, USA_F, BEL_M, AUT_M, ESP_M, DEU_M, NZL_F, AUS_M, FIN_M$	0.29
Denmark	CAN_M	0.029
Finland	$NZL_M, ESP_M, AUT_M, CAN_M$	0.29
France	$PRT_M, FIN_M, NZL_M, ESP_M, AUT_M$	0.56
Germany	-	-
Ireland	$NOR_M, JPN_M, CHE_M, UK_M$	0.15
Italy	$UK_M, CHE_M, AUS_M, DEU_M, IRL_M$	0.20
Japan	$NOR_M, IRL_M, CHE_M, UK_M, DNK_M$	0.206
Netherlands	$NZL_M, DNK_F, SWE_M, CAN_F$	0.15
New Zealand	$ESP_M, AUT_M, CAN_M, FIN_M, DNK_M, USA_F, BEL_M$	0.199
Norway	IRL_M	0.0004
Portugal	$FRA_M, USA_M, FIN_M, NZL_M, ESP_M, AUT_M, CAN_M$	0.95
Spain	13 populations	0.52
Sweden	CAN_M	0.03
Switzerland	$UK_M, ITA_M, IRL_M, NOR_M, JPN_M, AUS_M, DEU_M$	0.25
UK	$CHE_M, ITA_M, AUS_M, IRL_M, NOR_M, JPN_M, DEU_M$	0.304
USA	PRT_M	0.31

Table 10: Best reference group for males of low-mortality countries according to lowest ME(e_0) in LL

Country	Other populations in best reference group	$\bar{e}_i^\dagger - \bar{e}_j^\dagger$
Australia	$\mathrm{DEU_M, ITA_M}$	0.10
Austria	$\mathrm{ESP_M}$	0.04
Belgium	$\mathrm{USA_F, DNK_M, CAN_M, DEU_M, AUT_M, AUS_M, ESP_M}$	0.19
Canada	10 populations	0.30
Denmark	$\mathrm{CAN_M, USA_F}$	0.029
Finland	$\mathrm{NZL_M, ESP_M, AUT_M}$	0.204
France	$\mathrm{PRT_M, FIN_M, NZL_M, ESP_M, AUT_M}$	0.56
Germany	-	-
Ireland	$\mathrm{NOR_M, JPN_M, CHE_M, UK_M}$	0.15
Italy	$\mathrm{UK_M, CHE_M, AUS_M}$	0.1
Japan	$\mathrm{NOR_M}$	0.03
Netherlands	$\mathrm{NZL_M, DNK_F, SWE_M}$	0.12
New Zealand	$\mathrm{ESP_M}$	0.008
Norway	$\mathrm{IRL_M}$	0.004
Portugal	$\mathrm{FRA_M}$	0.29
Spain	$\mathrm{NZL_M, AUT_M}$	0.04
Sweden	$\mathrm{CAN_M, UK_M, NLD_M}$	0.127
Switzerland	$\mathrm{UK_M, ITA_M, IRL_M, NOR_M, JPN_M, AUS_M}$	0.184
UK	$\mathrm{CHE_M}$	0.03
USA	$\mathrm{PRT_M, FRA_M}$	0.61

Table 11: Best reference group for females of low-mortality countries according to lowest MAE in LL

Country	Other populations in best reference group	$\bar{e}_i^\dagger - \bar{e}_j^\dagger$
Australia	$\mathrm{BEL_F, IRL_F, FRA_F}$	0.088
Austria	13 populations	0.39
Belgium	$\mathrm{AUS_F, IRL_F, NLD_F}$	0.12
Canada	$\mathrm{SWE_M, UK_F, NLD_F}$	0.15
Denmark	$\mathrm{NZL_F, NLD_M}$	0.06
Finland	$\mathrm{ITA_F, SWE_F, JPN_F, CHE_F, AUT_F, ESP_F, NOR_F, DEU_F, NLD_F, FRA_F}$	0.36
France	13 populations	0.44
Germany	-	-
Ireland	$\mathrm{AUS_F, BEL_F}$	0.03
Italy	$\mathrm{FIN_F, AUT_F, SWE_F, NOR_F, JPN_F, CHE_F, ESP_F}$	0.15
Japan	$\mathrm{CHE_F, SWE_F}$	0.01
Netherlands	$\mathrm{FRA_F}$	0.05
New Zealand	$\mathrm{DNK_F, NLD_M, , SWE_M}$	0.06
Norway	$\mathrm{AUT_F}$	0.048
Portugal	$\mathrm{IRL_F, AUS_F}$	0.151
Spain	$\mathrm{CHE_F, JPN_F, SWE_F}$	0.08
Sweden	$\mathrm{JPN_F, FIN_F, CHE_F}$	0.046
Switzerland	$\mathrm{JPN_F, ESP_F, SWE_F, FIN_F, ITA_F, AUT_F, NOR_F, DEU_F, NLD_F}$	0.403
UK	$\mathrm{CAN_F, SWE_M, PRT_F, NLD_M, IRL_F}$	0.297
USA	$\mathrm{BEL_M}$	0.002

Table 12: Best reference group for females of low-mortality countries according to lowest MSE in LL

Country	Other populations in best reference group	$\lvert\bar{e}^\dagger_i - \bar{e}^\dagger_j\rvert$
Australia	$\mathrm{BEL_F, IRL_F, FRA_F}$	0.088
Austria	13 populations	0.39
Belgium	$\mathrm{AUS_F, IRL_F, NLD_F}$	0.12
Canada	$\mathrm{SWE_M, UK_F, NLD_F, NZL_F, DNK_F}$	0.22
Denmark	$\mathrm{NZL_F, NLD_M}$	0.06
Finland	13 populations	0.44
France	13 populations	0.44
Germany	11 populations	0.31
Ireland	$\mathrm{AUS_F, BEL_F}$	0.03
Italy	$\mathrm{FIN_F, AUT_F, SWE_F, NOR_F, JPN_F}$	0.11
Japan	$\mathrm{CHE_F, SWE_F}$	0.01
Netherlands	$\mathrm{FRA_F}$	0.05
New Zealand	$\mathrm{DNK_F}$	0.003
Norway	$\mathrm{AUT_F, DEU_F}$	0.09
Portugal	$\mathrm{IRL_F, AUS_F}$	0.151
Spain	$\mathrm{CHE_F, JPN_F, SWE_F}$	0.08
Sweden	$\mathrm{JPN_F, FIN_F, CHE_F, ITA_F}$	0.07
Switzerland	$\mathrm{JPN_F, ESP_F, SWE_F, FIN_F}$	0.088
UK	$\mathrm{CAN_F, SWE_M, PRT_F, NLD_M}$	0.248
USA	$\mathrm{BEL_M}$	0.002

Table 13: Best reference group for females of low-mortality countries according to lowest ME(e_0) in LL

Country	Other populations in best reference group	$\lvert\bar{e}^\dagger_i - \bar{e}^\dagger_j\rvert$
Australia	$\mathrm{BEL_F, IRL_F, FRA_F, NLD_F, PRT_F, DEU_F}$	0.22
Austria	$\mathrm{NOR_F, ITA_F, FIN_F, SWE_F, DEU_F, JPN_F, CHE_F}$	0.17
Belgium	$\mathrm{AUS_F, IRL_F, NLD_F, PRT_F}$	0.16
Canada	$\mathrm{SWE_M}$	0.03
Denmark	$\mathrm{NZL_F, NLD_M}$	0.06
Finland	$\mathrm{ITA_F, SWE_F, JPN_F, CHE_F}$	0.04
France	$\mathrm{NLD_F, BEL_F, AUS_F, IRL_F, DEU_F, NOR_F, PRT_F, AUT_F, ITA_F}$	0.33
Germany	$\mathrm{NLD_F}$	0.079
Ireland	$\mathrm{AUS_F}$	0.02
Italy	$\mathrm{FIN_F}$	0.03
Japan	$\mathrm{CHE_F, SWE_F}$	0.01
Netherlands	$\mathrm{FRA_F}$	0.05
New Zealand	$\mathrm{DNK_F, , NLD_F, SWE_M, JPN_M, CAN_F, NOR_M, IRL_M}$	0.241
Norway	$\mathrm{AUT_F}$	0.048
Portugal	$\mathrm{IRL_F, AUS_F, BEL_F, UK_F, FRA_F}$	0.24
Spain	$\mathrm{CHE_F, JPN_F, SWE_F, FIN_F, ITA_F, AUT_F, NOR_F, DEU_F, NLD_F}$	0.43
Sweden	$\mathrm{JPN_F, FIN_F}$	0.041
Switzerland	$\mathrm{JPN_F, ESP_F, SWE_F, FIN_F, ITA_F, AUT_F, NOR_F, DEU_F, NLD_F}$	0.403
UK	$\mathrm{CAN_F, SWE_M}$	0.12
USA	$\mathrm{BEL_M, DNK_M, CAN_F, DEU_M}$	0.084

Table 14: Best reference group for males of low-mortality countries according to lowest MAE in $\mathrm{LL}_{e^\dagger}$

Country	Other populations in best reference group (best fitting period from:)	$\bar{e}_i^\dagger - \bar{e}_j^\dagger$
Australia	$\mathrm{DEU_M}$	0.06
Austria	$\mathrm{ESP_M}$	0.04
Belgium	9 populations	0.199
Canada	10 populations	0.30
Denmark	$\mathrm{CAN_M, USA_F, BEL_M}$	0.03
Finland	$\mathrm{NZL_M, ESP_M, AUT_M}$	0.204
France	$\mathrm{PRT_M}$	0.29
Germany	12 populations	0.376
Ireland	$\mathrm{NOR_M, JPN_M, CHE_M, UK_M}$	0.15
Italy	$\mathrm{UK_M, CHE_M, AUS_M}$	0.1
Japan	$\mathrm{NOR_M}$	0.03
Netherlands	$\mathrm{NZL_M, DNK_F, SWE_M, CAN_F}$ (1974:)	0.15
New Zealand	$\mathrm{ESP_M}$	0.008
Norway	$\mathrm{IRL_M}$	0.0004
Portugal	9 populations (1965:)	1.00
Spain	$\mathrm{NZL_M, AUT_M, CAN_M, CAN_M}$	0.13
Sweden	$\mathrm{CAN_M}$	0.03
Switzerland	$\mathrm{UK_M, ITA_M}$	0.08
UK	10 populations	0.33
USA	$\mathrm{PRT_M, FRA_M, FIN_M, NZL_M, ESP_M, AUT_M}$ (1965:)	1.17

Note: For combinations without mentioned best fitting period have best fit for (1956-2001).

Table 15: Best reference group for males of low-mortality countries according to lowest MSE in $\mathrm{LL}_{e^\dagger}$

Country	Other populations in best reference group (best fitting period from:)	$\bar{e}_i^\dagger - \bar{e}_j^\dagger$
Australia	$\mathrm{DEU_M}$	0.06
Austria	$\mathrm{ESP_M}$	0.04
Belgium	$\mathrm{USA_F, DNK_M, CAN_M, DEU_M, AUT_M, AUS_M, ESP_M}$	0.19
Canada	$\mathrm{DNK_M, USA_F, BEL_M, AUT_M, ESP_M, DEU_M, NZL_F, AUS_M, FIN_M}$	0.29
Denmark	$\mathrm{CAN_M, USA_F, BEL_M, DEU_M}$	0.11
Finland	$\mathrm{NZL_M, ESP_M, AUT_M, CAN_M}$	0.29
France	$\mathrm{PRT_M}$	0.29
Germany	13 populations	0.377
Ireland	$\mathrm{NOR_M, JPN_M, CHE_M, UK_M}$	0.15
Italy	$\mathrm{UK_M, CHE_M, AUS_M, DEU_M}$	0.16
Japan	$\mathrm{NOR_M}$	0.03
Netherlands	$\mathrm{NZL_M, DNK_F, SWE_M, CAN_F}$ (1974:)	0.15
New Zealand	$\mathrm{ESP_M}$	0.008
Norway	$\mathrm{IRL_M, JPN_M, CHE_M, UK_M}$	0.15
Portugal	9 populations (1965:)	1.00
Spain	13 populations	0.52
Sweden	$\mathrm{CAN_M}$	0.03
Switzerland	$\mathrm{UK_M, ITA_M}$	0.08
UK	10 populations	0.33
USA	$\mathrm{PRT_M, FRA_M, FIN_M, NZL_M, ESP_M, AUT_M, DEU_M}$ (1965:)	1.40

Note: For combinations without mentioned best fitting period have best fit for (1956-2001).

Table 16: Best reference group for males of low-mortality countries according to lowest ME(e_0) in LL$_{e^\dagger}$

Country	Other populations in best reference group (best fitting period from:)	$\bar{e}_i^\dagger - \bar{e}_j^\dagger$
Australia	$\text{DEU}_\text{M}, \text{ITA}_\text{M}$	0.10
Austria	ESP_M	0.04
Belgium	9 populations	0.25
Canada	10 populations	0.30
Denmark	$\text{CAN}_\text{M}, \text{USA}_\text{F}, \text{BEL}_\text{M}$	0.03
Finland	$\text{NZL}_\text{M}, \text{ESP}_\text{M}, \text{AUT}_\text{M}, \text{DNK}_\text{M}, \text{USA}_\text{F}$ (1964:)	0.35
France	PRT_M	0.29
Germany	12 populations	0.376
Ireland	$\text{NOR}_\text{M}, \text{JPN}_\text{M}, \text{CHE}_\text{M}, \text{UK}_\text{M}$	0.15
Italy	$\text{UK}_\text{M}, \text{CHE}_\text{M}, \text{AUS}_\text{M}$	0.1
Japan	NOR_M	0.03
Netherlands	$\text{NZL}_\text{M}, \text{DNK}_\text{F}, \text{SWE}_\text{M}, \text{CAN}_\text{F}$ (1974:)	0.15
New Zealand	ESP_M	0.008
Norway	IRL_M	0.004
Portugal	$\text{FRA}_\text{M}, \text{USA}_\text{M}, \text{FIN}_\text{M}, \text{NZL}_\text{M}$ (1965:)	1.00
Spain	$\text{NZL}_\text{M}, \text{AUT}_\text{M}$	0.04
Sweden	CAN_M	0.03
Switzerland	$\text{UK}_\text{M}, \text{ITA}_\text{M}$	0.08
UK	10 populations	0.33
USA	$\text{PRT}_\text{M}, \text{FRA}_\text{M}, \text{FIN}_\text{M}, \text{NZL}_\text{M}$ (1965:)	1.12

Note: For combinations without mentioned best fitting period have best fit for (1956-2001).

Table 17: Best reference group for females of low-mortality countries according to lowest MAE in LL$_{e^\dagger}$

Country	Other populations in best reference group (best fitting period from:)	$\bar{e}_i^\dagger - \bar{e}_j^\dagger$
Australia	$\text{BEL}_\text{F}, \text{IRL}_\text{F}, \text{FRA}_\text{F}$	0.088
Austria	$\text{NOR}_\text{F}, \text{ITA}_\text{F}, \text{FIN}_\text{F}, \text{SWE}_\text{F}, \text{DEU}_\text{F}, \text{JPN}_\text{F}, \text{CHE}_\text{F}, \text{ESP}_\text{F}, \text{NLD}_\text{F}$ (1957:)	0.22
Belgium	$\text{AUS}_\text{F}, \text{IRL}_\text{F}, \text{NLD}_\text{F}, \text{PRT}_\text{F}, \text{DEU}_\text{F}, \text{NOR}_\text{F}$ (1974:)	0.30
Canada	$\text{SWE}_\text{M}, \text{UK}_\text{F}, \text{NLD}_\text{F}, \text{NZL}_\text{F}$	0.21
Denmark	$\text{NZL}_\text{F}, \text{NLD}_\text{M}, \text{SWE}_\text{M},$	0.19
Finland	$\text{ITA}_\text{F}, \text{SWE}_\text{F}, \text{JPN}_\text{F}, \text{CHE}_\text{F}$	0.08
France	$\text{NLD}_\text{F}, BEL_\text{F}, \text{AUS}_\text{F}, \text{IRL}_\text{F}, \text{DEU}_\text{F}, \text{NOR}_\text{F}, \text{PRT}_\text{F}, \text{AUT}_\text{F}$ (1974:)	0.33
Germany	$\text{NLD}_\text{F}, \text{NOR}_\text{F}, \text{FRA}_\text{F}, \text{AUT}_\text{F}, \text{ITA}_\text{F}$	0.20
Ireland	$\text{AUS}_\text{F}, \text{BEL}_\text{F}$	0.03
Italy	$\text{FIN}_\text{F}, \text{AUT}_\text{F}, \text{SWE}_\text{F}, \text{NOR}_\text{F}, \text{JPN}_\text{F}, \text{CHE}_\text{F}, \text{ESP}_\text{F}, \text{DEU}_\text{F}$ (1957:)	0.20
Japan	$\text{CHE}_\text{F}, \text{SWE}_\text{F}, \text{ESP}_\text{F}, \text{FIN}_\text{F}$	0.07
Netherlands	$\text{FRA}_\text{F}, \text{DEU}_\text{F}, \text{BEL}_\text{F}, \text{AUS}_\text{F}$ (1977:)	0.14
New Zealand	$\text{DNK}_\text{F}, \text{NLD}_\text{M}, \text{SWE}_\text{M}, \text{JPN}_\text{M}, \text{CAN}_\text{F}, \text{NOR}_\text{M}$ (1976:)	0.24
Norway	AUT_F	0.048
Portugal	$\text{IRL}_\text{F}, \text{AUS}_\text{F}, \text{BEL}_\text{F}, \text{UK}_\text{F}, \text{FRA}_\text{F}$ (1966:)	0.24
Spain	$\text{CHE}_\text{F}, \text{JPN}_\text{F}, \text{SWE}_\text{F}, \text{FIN}_\text{F}, \text{ITA}_\text{F}, \text{AUT}_\text{F}, \text{NOR}_\text{F}$ (1957:)	0.26
Sweden	$\text{JPN}_\text{F}, \text{FIN}_\text{F}, \text{CHE}_\text{F}$	0.04
Switzerland	13 populations (1958:)	0.57
UK	CAN_F (1972:)	0.08
USA	BEL_M	0.002

Note: For combinations without mentioned best fitting period have best fit for (1956-2001).

Table 18: Best reference group for females of low-mortality countries according to lowest MSE in LL$_{e^\dagger}$

Country	Other populations in best reference group (best fitting period from:)	$\bar{e}_i^\dagger - \bar{e}_j^\dagger$
Australia	$\mathrm{BEL_F}$	0.012
Austria	$\mathrm{NOR_F, ITA_F, FIN_F, SWE_F}$	0.13
Belgium	$\mathrm{AUS_F, IRL_F, NLD_F, PRT_F, DEU_F, NOR_F}$ (1974:)	0.30
Canada	$\mathrm{SWE_M, UK_F, NLD_F, NZL_F}$	0.21
Denmark	$\mathrm{NZL_F, NLD_M, SWE_M,}$	0.19
Finland	$IT_{\mathrm{F}}, \mathrm{SWE_F, JPN_F, CHE_F}$	0.08
France	$\mathrm{NLD_F}, NEL_{\mathrm{F}}, \mathrm{AUS_F, IRL_F, DEU_F, NOR_F, PRT_F, AUT_F}$ (1974:)	0.33
Germany	$\mathrm{NLD_F, NOR_F, FRA_F, AUT_F, ITA_F, BEL_F, AUS_F}$ (1977:)	0.22
Ireland	$\mathrm{AUS_F, BEL_F}$	0.03
Italy	$\mathrm{FIN_F, AUT_F, SWE_F, NOR_F, JPN_F, CHE_F, ESP_F, DEU_F, NLD_F}$	0.28
Japan	$\mathrm{CHE_F, SWE_F, ESP_F, FIN_F}$	0.07
Netherlands	$\mathrm{FRA_F, DEU_F, BEL_F, AUS_F}$ (1977:)	0.14
New Zealand	$\mathrm{DNK_F, NLD_M, , SWE_M, JPN_M, CAN_F, NOR_M}$ (1976:)	0.24
Norway	$\mathrm{AUT_F}$	0.04
Portugal	$\mathrm{IRL_F, AUS_F, BEL_F, UK_F, FRA_F}$ (1966:)	0.24
Spain	$\mathrm{CHE_F, JPN_F, SWE_F, FIN_F, ITA_F, AUT_F, NOR_F}$ (1957:)	0.26
Sweden	$\mathrm{JPN_F, FIN_F, CHE_F}$	0.04
Switzerland	10 populations (1957:)	0.45
UK	$\mathrm{CAN_F}$ (1972:)	0.08
USA	$\mathrm{BEL_M}$	0.002

Note: For combinations without mentioned best fitting period have best fit for (1956-2001).

Table 19: Best reference group for females of low-mortality countries according to lowest ME(e_0) in LL$_{e^\dagger}$

Country	Other populations in best reference group (best fitting period from:)	$\bar{e}_i^\dagger - \bar{e}_j^\dagger$
Australia	$\mathrm{BEL_F, IRL_F, FRA_F}$	0.088
Austria	$\mathrm{NOR_F}$	0.04
Belgium	$\mathrm{AUS_F}$	0.012
Canada	$\mathrm{SWE_M, UK_F, NLD_F, NZL_F, DNK_F}$	0.22
Denmark	$\mathrm{NZL_F, NLD_M, SWE_M,}$	0.19
Finland	$\mathrm{ITA_F, SWE_F, JPN_F, CHE_F}$	0.08
France	$\mathrm{NLD_F, BEL_F, AUS_F, IRL_F, DEU_F, NOR_F, PRT_F, AUT_F}$	0.33
Germany	$\mathrm{NLD_F, NOR_F, FRA_F, AUT_F, ITA_F}$	0.20
Ireland	$\mathrm{AUS_F, BEL_F, FRA_F}$	0.11
Italy	$\mathrm{FIN_F}$	0.03
Japan	$\mathrm{CHE_F, SWE_F, ESP_F, FIN_F}$	0.07
Netherlands	$\mathrm{FRA_F}$	0.05
New Zealand	$\mathrm{DNK_F}$	0.003
Norway	$\mathrm{AUT_F}$	0.04
Portugal	$\mathrm{IRL_F, AUS_F, BEL_F, UK_F, FRA_F, CAN_F, NLD_F, SWE_M, DEU_F}$ (1966:)	0.37
Spain	$\mathrm{CHE_F, JPN_F, SWE_F, FIN_F, ITA_F, AUT_F, NOR_F}$ (1957:)	0.26
Sweden	$\mathrm{JPN_F, FIN_F, CHE_F}$	0.04
Switzerland	13 populations	0.57
UK	$\mathrm{CAN_F}$ (1972:)	0.08
USA	$\mathrm{BEL_M}$	0.002

Note: For combinations without mentioned best fitting period have best fit for (1956-2001).