Chapter 8

Coherent Mortality Forecasting with Standards: Low Mortality Serves as a Guide

Heather Booth

8.1 Introduction

Mortality forecasts are an important component of population forecasting and are central to the estimation of longevity risk in actuarial practice. Planning by the state for health and aged care services and by individuals for retirement and later life depends on accurate mortality forecasts. The overall accuracy or performance of mortality forecasting has improved since Lee and Carter (1992) introduced stochastic forecasting of mortality to the demographic community, and further improvements can undoubtedly be made.

The series of new methods and method refinements contributing to improved performance include various extensions of the Lee-Carter method (e.g., Booth 2006; Booth et al. 2002, 2006; de Jong and Tickle 2006; Li and Li 2017; Li 2012; Li et al. 2013; Shang et al. 2011; Tickle and Booth 2014). The independently developed functional data approach of Hyndman and Ullah (2007) is a generalisation of Lee-Carter. Other approaches include general linear modelling (e.g., Ahmadi and Li 2014; Currie 2014; Renshaw and Haberman 2003, 2006), Bayesian methods (e.g., Cairns et al. 2011; Raftery et al. 2013), and compositional data modelling (Bergeron-Boucher et al. 2017), among others (Basellini and Camarda 2019; Booth and Tickle 2008; Camarda 2019; de Beer and Janssen 2016; Janssen 2018; Pascariu et al. 2018). However, the principal components approach, used in the Lee-Carter method, remains prominent.

A logical and fruitful development is coherent forecasting where the mortality experience of two or more populations are forecast jointly, with the expectation that forecast performance will be improved by borrowing strength from the complemen-

H. Booth (✉)

School of Demography, Australian National University, Canberra, Australia
e-mail: heather.booth@anu.edu.au

© The Author(s) 2020

S. Mazzucco, N. Keilman (eds.), *Developments in Demographic Forecasting*,
The Springer Series on Demographic Methods and Population Analysis 49,
https://doi.org/10.1007/978-3-030-42472-5_8

153

tary, or ‘other’, population(s). Li and Lee (2005) introduced this idea by forecasting the mortality of a group of populations with similar mortality experience, identified as an integral part of model estimation. This common factor approach has been further developed by others (e.g., Li 2012).

The product-ratio method of coherent forecasting was proposed by Hyndman et al. (2013) following earlier unpublished work by Booth using ratios. The two examples used to illustrate the method forecast the mortality of two or more subpopulations within a country: the two sex-specific populations in sex-coherent mortality forecasting for Sweden and the populations of the several states of Australia in state-coherent mortality forecasting. It was noted that forecast accuracy and bias, averaged over the subpopulations, was improved by using the coherent method when compared with independent forecasts for each subpopulation. Further, forecast accuracy and bias were homogenised across the subpopulations, a feature of considerable benefit in actuarial and population projection applications. The generalisability of these findings to other countries has not previously been investigated. This study evaluates sex-coherent forecasting using a wide range of populations.

The use of an external standard or reference population in forecasting mortality has been variously proposed (Basellini and Camarda 2019; Fazole Rabbi 2019; Hyndman et al. 2013; Li and Lee 2005). The choice of external standard is often somewhat arbitrary; possible criteria include language, geographic proximity, political entity, and mortality level. However derived, a standard can be used in the product-ratio method to produce standard-coherent forecasts. By choosing an appropriate standard, the borrowed strength can be expected to result in a better forecast of the population of interest. This constitutes a novel application of the product-ratio method. This standard-coherent method is evaluated in this study.

8.2 Study Design

8.2.1 *Aim, Objectives and Hypothesis*

The overall aim of this study is to determine whether taking appropriate other mortality into account (by using the product-ratio method) improves the performance of mortality forecasting, as measured by accuracy, bias and robustness. This is addressed through three successive objectives.

The first objective is to evaluate the performance, compared with independent forecasts, of sex-coherent forecasting across a wide range of populations. It is expected, based on the example of Sweden in Hyndman et al. (2013) and preliminary research by the author that male mortality forecasts are improved when female mortality is taken into account, but not vice versa. Noting that female mortality is lower than male mortality, my hypothesis is that a low-mortality standard will serve as a better guide to future mortality, given the prevailing trend of decline, than a higher-mortality standard.

Based on this hypothesis, the second objective is to use a selection of low-mortality standards to evaluate the performance of standard-coherent forecasting across the range of populations. The third objective is to compare the forecast performance of independent, sex-coherent and standard-coherent forecasts in order to determine how these three methods rank for female and male mortality.

8.2.2 Data

Data are obtained from the Human Mortality Database (“Human Mortality Database,” 2019) (HMD) for the period 1950–2014; this period was chosen to maximise as far as practicable the number of countries with available data. This resulted in a total of 21 countries (Table 8.1) being included in the analysis. The data comprise annual age-sex-specific central death rates, or mortality rates, and corresponding populations exposed to the risk of death.

The available data are for single years of age 0–109 and for the open-ended interval 110+. Initial evaluation of the mortality rates showed that, for all countries, observed rates at the oldest ages were lower in the earlier years of observation than in more recent years; it is assumed that this is the result of improved age at death

Table 8.1 Ranking of countries by sex-specific life expectancy in 2014

Country	Female		Male	
	e(0)	Rank	e(0)	Rank
Japan	86.81	1	80.51	5
Spain	85.60	2	80.08	7
France	85.43	3	79.27	11
Italy	85.15	4	80.55	4
Switzerland	85.11	5	80.93	2
Australia	84.56	6	80.59	3
Iceland	84.12	7	81.12	1
Portugal	84.12	7	77.92	18
Norway	84.10	9	80.03	8
Sweden	84.05	10	80.36	6
Canada	83.98	11	79.81	10
Finland	83.85	12	78.13	17
Austria	83.73	13	78.92	14
Belgium	83.51	14	78.57	15
Netherlands	83.29	15	79.88	9
Ireland	83.21	16	79.16	13
UK	82.99	17	79.25	12
Denmark	82.67	18	78.57	15
Czechia	81.73	19	75.72	20
USA	81.45	20	76.64	19
Hungary	79.24	21	72.26	21

reporting over time and selection effects rather than a real increase in mortality. In order to avoid erroneously modelling increasing mortality at the oldest ages, the data for ages 95 and older were combined into a revised open-ended interval. In other circumstances, it would be desirable to model mortality rates at the oldest ages (Buettner 2002). Here, however, there is little, if any, gain in doing so because the objective is to compare the performance of forecasting methods and because modelled rates would follow the same pattern in the standard as in the population of interest.

8.2.3 *Choice of Standard*

The evaluation of standard-coherent forecasting will obviously depend on the choice of standard. In line with the hypothesised role of a low-mortality standard, four leaders of the global mortality decline, measured in terms of life expectancy in 2014, were identified for use as standards. Table 8.1 shows 2014 life expectancy by sex for the 21 countries in the study. Countries with a total population size of less than one million were discounted in this process in order to avoid excessive fluctuation in the standard; this size criterion applies only to Iceland which, in fact, recorded the highest ranking male life expectancy in 2014 (Table 8.1). The sex-specific standards employed are Japan and Spain (1st and 2nd respectively for female life expectancy), and Switzerland and Australia (2nd and 3rd respectively for male life expectancy). These four countries were excluded from the analytical group of 17 countries on which the comparative analysis is based so as to maintain comparability of results.

8.2.4 *Rolling Fitting Period*

Any forecast is dependent on the particular fitting period used. Forecast error also depends on the particular year in the forecast period combined with forecast horizon. For evaluative purposes, it is important to take these influences into account as far as possible. This is done by appropriate averaging over forecasts. A rolling forecast origin is commonly used in the calculation of average error, so as to reduce the effect of fluctuations and abrupt changes in annual mortality rates in relation to the fitting period and the forecast period. In previous work, the rolling aspect has been restricted to the last year of the fitting period, or jump-off year, on the basis that time series methods give little weight to earlier data (Hyndman et al. 2013).

In this analysis, rather than fixing the first year of the fitting period, the length of the fitting period is fixed; the first and last years of the fitting period are simultaneously rolling. This is considered more robust, as a fixed first year of the fitting period could in some circumstances lead to systematic bias. Given 65 years of data and setting the maximum forecast horizon at 23 years (to obtain reliable results for up to 20 years), the fitting period length is fixed at 42 years. As the fitting period

Note: Horizons 21 to 23 are not used because of low frequencies.

Fig. 8.1 Rolling fitting period of length 42 years, calendar years in forecast period (years 43–65) and forecast horizons (1–23 years)

is rolled forward in time, the forecast horizon is correspondingly reduced. Figure 8.1 illustrates how this procedure produces forecasts for horizons, h , of 1–23 years with diminishing frequency, there being 23 forecasts of $h=1$, 22 forecasts of $h=2$, . . . , 2 forecasts of $h=22$ and 1 forecast of $h=23$. Forecasts based on three or fewer values are excluded from the evaluation; these are for horizons 21–23. Thus, the reported mean results cover horizons of 1–20 years, with greater confidence in means for shorter horizons deriving from larger numbers of observations.

8.2.5 Measures Used in Evaluation

The forecasts are evaluated using several measures, based on forecast error in the mortality rates at age x and time t for country c , $m(x, t, c)$. First, the accuracy of the point forecast is measured by the mean absolute relative error, *MARE*, in age-specific mortality rates, averaged over age and fitting period. For country c , the *MARE* for horizon h is defined as

$$MARE(h, c) = \frac{1}{(24-h) \times 96} \sum_{t=42}^{65-h} \sum_{x=0}^{95} \frac{|m(x, t+h, c) - \hat{m}(x, t+h, c)|}{m(x, t+h, c)}$$

where $\hat{m}(x, t + h, c)$ is the forecast rate for country c at age x and t is an index of year. For all horizons, the fitting period is 42 years of data starting in year $t=1, 2, \dots, 23$ and, correspondingly, the forecast period starts in year $t=43, 44, \dots, 65$ and ends in year $t=65$.

Second, the mean relative error, *MRE*, is used to assess bias. In demographic forecasting, it is often of primary interest to know whether the point forecast is biased and in which direction. For country *c*, the *MRE* for horizon *h* is defined as

$$MRE(h, c) = \frac{1}{(24 - h) \times 96} \sum_{t=42}^{65-h} \sum_{x=0}^{95} \frac{m(x, t + h, c) - \hat{m}(x, t + h, c)}{m(x, t + h, c)}$$

The use of relative errors gives equal weight across ages, regardless of the size of the rate, thus removing the effect of different levels and different age patterns of mortality in the comparative assessments. (Note that relative weights are conceptually independent of size of rate). Country comparisons are thus valid, and each country has equal weight in overall averages. Sex comparisons are similarly valid (all results are sex-specific). The use of relative errors also permits direct comparison of errors across horizons, and facilitates interpretation of averages and variability over horizons.

The units of analysis for evaluation and comparison are *MARE(h, c)* and *MRE(h, c)*. Horizon-specific mean accuracy and bias, *MARE(h)* and *MRE(h)*, are averages over countries; these describe the average ‘horizon effect’ in accuracy and bias, or degree to which forecast performance declines over time. Country-specific mean accuracy and bias, *MARE(c)* and *MRE(c)*, are averages over horizons; these measure the degree of difficulty in forecasting mortality for each population. Overall mean accuracy and bias, *MARE* and *MRE*, are averages over countries and horizons:

$$MARE = \frac{1}{17} \sum_{c=1}^{17} MARE(c) = \frac{1}{20} \sum_{h=1}^{20} MARE(h) = \frac{1}{17 \times 20} \sum_{c=1}^{17} \sum_{h=1}^{20} MARE(h, c)$$

$$MRE = \frac{1}{17} \sum_{c=1}^{17} MRE(c) = \frac{1}{20} \sum_{h=1}^{20} MRE(h) = \frac{1}{17 \times 20} \sum_{c=1}^{17} \sum_{h=1}^{20} MRE(h, c)$$

It should be noted that *MRE* is a measure of net bias. The values of *MRE(h, c)* are net across ages and across fitting periods. Additionally, *MRE(c)* is net across horizons, *MRE(h)* is net across countries, and the overall mean is net across horizons and countries. Absolute bias is used in some comparisons.

Third, the heterogeneity or standard deviations of accuracy and bias are used to assess method robustness. (Note these are not based on forecast variance as used in the estimation of the interval forecast; the interval forecast is not within the scope of the study.) Two measures of heterogeneity are used in parallel with the average measures. The first is the standard deviation across countries for each horizon:

$$SD_h(MARE) = \sqrt{\frac{1}{17} \sum_{c=1}^{17} (MARE(h, c) - MARE(h))^2}$$

and similarly for $SD_h(MRE)$. This measure shows the degree of country variation in the horizon effect. A low value is preferable as it indicates that the method is robust to different mortality conditions.

The second measure of heterogeneity is the standard deviation across horizons for each country:

$$SD_c(MARE) = \sqrt{\frac{1}{20} \sum_{h=1}^{20} (MARE(h, c) - MARE(c))^2}$$

and similarly for $SD_c(MRE)$. This shows the degree of variability over horizon in accuracy and bias for country c , due to the horizon effect, and a low value is preferable. The average of $SD_c(MARE)$ and $SD_c(MRE)$ over countries provides an overall measure of the degree of heterogeneity across horizons, which is used in comparing methods.

The study includes discussion of the sex-differences in accuracy and bias averaged over countries, $MARE_M(h) - MARE_F(h)$ and $MRE_M(h) - MRE_F(h)$, and of sex-differences in accuracy and bias averaged over horizons, $MARE_M(c) - MARE_F(c)$ and $MRE_M(c) - MRE_F(c)$. Note that these are not the accuracy and bias of the sex-difference in mortality.

8.3 Forecasting Methods

8.3.1 Functional Data Forecasting

The forecasting methods employed in this research draw on the functional forecasting approach of Hyndman and Ullah (2007). The Hyndman-Ullah functional data method (FDM) is a generalisation of the well-known Lee-Carter method (Lee and Carter 1992), and models and forecasts the natural logarithm of period age-specific mortality rates for a particular population or country (in this section, c is dropped from formulae). The functional data model is

$$\ln(m(x, t)) = a(x) + \sum_j b_j(x)k_j(t) + e(x, t) + \sigma(x, t) \varepsilon(x, t)$$

where $a(x)$ is the temporal average pattern of the logarithm of mortality by age and, for $j = 1, \dots, J$ components, $b_j(x)$ is a ‘basis function’ and $k_j(t)$ is a time series coefficient. Broadly, the $k_j(t)$ represent annual rates of mortality decline averaged over age, while the $b_j(x)$ describe the age pattern of decline averaged over time. The parameters of the model are estimated after smoothing the data over age. Thus, the $a(x)$ and $b_j(x)$ are smooth functions of age. The pairs $(b_j(x), k_j(t))$ for $j = 1, \dots, J$ are estimated using principal component decomposition. The error term $\sigma(x, t) \varepsilon(x, t)$

accounts for age-varying observational error; this is the difference between the observed rates and the smoothed rates. The error term $e(x, t)$ is modelling error, or the difference between the smoothed rates and the fitted rates from the model.

The FDM differs from the Lee-Carter method in several ways. First, as already noted, the $\ln(m(x, t))$ are smoothed over age prior to modelling. This is done using nonparametric smoothing methods and assuming monotonic increase at ages 65 and older. Each year of data is smoothed by applying weighted penalized regression splines where the weights are equal to the approximate inverse variance of the rate, i.e., $m(x, t) E(x, t)$, where $E(x, t)$ is population exposed to risk, and deaths are assumed to follow a Poisson distribution (Booth et al. 2014).

Second, the FDM uses functional principal components and, unlike Lee-Carter, employs more than one component of the decomposition. Following previous research (Hyndman and Booth 2008; Hyndman et al. 2013), six components are used for all data sets in this study. The remaining J–6 components form the error term, $e(x, t)$. Third, there is no adjustment of the time coefficients (as was the case in the original Lee-Carter method). Fourth, rather than routinely employing the random walk with drift model for forecasting the time coefficients; the most appropriate autoregressive integrated moving average (ARIMA) models are selected based on statistical criteria (Shumway and Stoffer 2006).

8.3.2 Coherent Forecasting

Coherent forecasting takes the experience of two or more populations into account and ensures that the resulting forecasts for each population are ‘non-divergent’, which encompasses the conditions that they do not converge (and cross over) in the short term nor diverge in the long term (Li and Lee 2005). The product-ratio method for coherent forecasting (Hyndman et al. 2013) uses the FDM in jointly forecasting mortality for two or more populations.

For sex-coherent forecasting, the product function is the geometric mean of sex-specific rates, $p(x, t) = \sqrt{m_F(x, t) m_M(x, t)}$, where F denotes female and M denotes male. The ratio function is the square root of the ratio of sex-specific rates, $r(x, t) = \sqrt{m_M(x, t) / m_F(x, t)}$. Because of the symmetry in the two-population case, the inverse ratio is not needed. These two functions are independently forecast using the FDM. Coherence is achieved by restricting the forecast of the ratio to converge very slowly to its temporal average; in other words, the forecast of each time coefficient converges to stationarity. For further details, including the case of three or more populations, see Hyndman et al. (2013).

The forecasts of the product and ratio functions are combined to produce forecast mortality rates. Forecast male mortality at future t is:

$$\sqrt{\widehat{m_F(x, t)} \widehat{m_M(x, t)}} \cdot \sqrt{\widehat{m_M(x, t) / m_F(x, t)}} = \sqrt{\widehat{m_M(x, t)}^2} = \widehat{m_M(x, t)} \quad (8.1)$$

and forecast female mortality at future t is:

$$\sqrt{\widehat{m_F(x, t) m_M(x, t)}} / \sqrt{\widehat{m_M(x, t) / m_M(x, t)}} = \sqrt{\widehat{m_F(x, t)^2}} = \widehat{m_F(x, t)} \quad (8.2)$$

The product-ratio coherent method makes use of the fact that the product and ratio will behave roughly independently of each other, as long as the two populations have approximately equal mortality variances (Hyndman et al. 2013). The method is directly applicable to the mortality of any two populations for which the coherence of their future mortality is postulated. Thus, the method is appropriate for standard-coherent forecasting where standard mortality is taken into account in forecasting the mortality of the population of interest. In the above equations, this is achieved by replacing F by S to denote standard (for example Japan), and by replacing M by the country of interest (for example, France). The forecast for the country of interest is then obtained by Eq. 8.1. Note that Eq. 8.2 is not used as, under the hypothesis that a low-mortality standard will serve as a better guide to future mortality, the forecast for the standard should not be obtained by reference to a population with higher mortality. In applying the standard-coherent method, sex-specific mortality rates are used.

8.4 Evidence: A Comparison of Methods

In line with the objectives of this research, sex-coherent and standard-coherent forecasts are evaluated in terms of accuracy, bias and robustness, against independent forecasts and against each other. The basic units of analysis, sex-specific accuracy and bias measures by horizon and country, $MARE(h, c)$ and $MRE(h, c)$, are illustrated in Fig. 8.2 for independent forecasts of female mortality, each graph representing 340 data points. Typical of forecasts in general, accuracy declines ($MARE(h, c)$ increases) and absolute bias increases with forecast horizon. Given relative measures of accuracy and bias, the increases observed are entirely attributable to the horizon effect. While forecasts for most countries exhibit relatively modest increases in forecast error with horizon, a handful exhibit substantial increases. Similar patterns are found in the basic units of analysis for all three methods, for accuracy and bias, and for each sex (Fig. 8.8).

8.4.1 Sex-Coherent Forecasts

The comparison of sex-coherent forecasts with independent forecasts is summarised in Fig. 8.3 using ratios of sex-coherent to independent measures, or relative

Fig. 8.2 Accuracy and bias by horizon and country, independent forecasts for female mortality

performance; see also Figs. 8.5 and 8.6, to be discussed later. The upper quadrants show country-specific relative accuracy and relative absolute bias, or ratios of averages over horizons, $MARE(c)$ and $|MRE(c)|$, for female and male mortality forecasts. These results show that the sex-coherent method is advantageous for forecasting male mortality but disadvantageous for forecasting female mortality. For male mortality, taking account of female mortality improved forecast accuracy and bias for 13–14 of the 17 countries, with an overall improvement across countries of 11% in accuracy and 12% in bias. However, taking account of male mortality in forecasting female mortality improved accuracy and bias for only 3–4 of the 17 countries, resulting in an overall reduction of 11% in accuracy and an overall increase of 32% in bias.

Similar patterns occur in relative heterogeneity across horizons, seen in the lower quadrants of Fig. 8.3. For male mortality, sex-coherent forecasting reduced the standard deviations of accuracy and bias, $SD_c(MARE)$ and $SD_c(MRE)$, for 15 of the 17 countries, with an overall reduction of 21% for both measures compared with independent forecasts. For female mortality, however, sex-coherent forecasting produced increased standard deviations for all but 3–4 countries, with overall increases of 24% for accuracy and 43% for bias.

Together, these findings generally confirm that forecast performance is improved for male mortality but reduced for female mortality when comparing the sex-coherent forecasts with independent forecasts. The hypothesis that low mortality serves as a good guide to future mortality is therefore supported in the context of sex-coherent forecasting.

Fig. 8.3 Accuracy and bias of sex-coherent forecasts relative to independent forecasts by country and sex. Note: For bias, USA female is an outlier. For bias heterogeneity, Sweden female is an outlier and Norway female and male are equal (For the interpretation of the references to colors in this Figure, please refer to the electronic version of the chapter available online)

8.4.2 *Standard-Coherent Forecasts*

The second objective of the study involves evaluation of the efficacy of several low-mortality standards in improving the performance of mortality forecasts. The third objective is to rank forecasts produced by the three methods (independent, sex-coherent and standard-coherent). These objectives are addressed in this section. Results are presented for the standard-coherent method using the four low-mortality standards described earlier, with comparable results for independent and sex-coherent forecasts. The case of Japan as standard, chosen for its leadership in life expectancy, is considered in detail; the results presented in Figs. 8.4, 8.5 and 8.6 are accuracy and bias means and standard deviations. For the remaining three low-mortality standards, only summary results are shown.

8.4.2.1 *Japan as Standard*

The evaluation focusses first on the horizon effect. Forecast accuracy and bias are averaged across countries. The upper quadrants of Fig. 8.4 show horizon-specific average accuracy and bias, $MARE(h)$ and $MRE(h)$, for the three methods by sex. Comparing methods, the standard-coherent forecast is the most accurate at all horizons for both sexes. For male mortality, the sex-coherent forecast is more accurate than the independent forecast, but for female mortality the reverse is found, as previously noted. Similar patterns among methods occur for bias, revealing a systematic tendency in the forecasts (except standard-coherent forecasts for female mortality) to underestimate the extent of future mortality decline (see also Fig. 8.8). These findings also show that the horizon effect is stable on average: mean accuracy and bias worsen steadily over forecast horizon, with an increasing advantage of standard-coherent forecasting.

The corresponding standard deviations, $SD_h(MARE)$ and $SD_h(MRE)$, are compared in the lower quadrants of Fig. 8.4. Heterogeneity among countries is relatively low at shorter horizons, particularly for accuracy, but increases rapidly at longer horizons, a result of substantial increases for some countries but not others (Fig. 8.8). This heterogeneity is significantly reduced by standard-coherent forecasting, while being selectively modified by sex-coherent forecasting as previously noted.

Focussing now on countries, forecast accuracy and bias are averaged across horizons. Figures 8.5 and 8.6 (upper quadrants) show, for females and males respectively, country-specific average accuracy and bias, $MARE(c)$ and $MRE(c)$, by method. For many countries (17 for male mortality and 9–10 for female mortality), the standard-coherent forecast is the best among the three methods in terms of both accuracy and bias, and this is reflected in the overall means (shown top right) which are averages across countries. Again, the sex-coherent method performs less well than the independent method for female mortality (Fig. 8.5) but performs better for male mortality (Fig. 8.6). These rankings among methods are also found for the

Fig. 8.4 Accuracy and bias by method (Japan as standard): country averages by horizon (For the interpretation of the references to colors in this Figure, please refer to the electronic version of the chapter available online)

Fig. 8.5 Accuracy and bias by method (Japan as standard) by country, female mortality

Fig. 8.6 Accuracy and bias by method (Japan as standard) by country, male mortality

standard deviations, $SD_c(MARE)$ and $SD_c(MRE)$, shown in the lower quadrants of Figs. 8.5 and 8.6.

This analysis identifies three countries, namely Czechia, Denmark and Hungary, for which forecasting errors are systematically largest when using the independent method, possibly due to their irregular patterns of mortality decline. Both female and male mortality in these countries gain substantially in performance from standard-coherent forecasting (Figs. 8.5 and 8.6). For Portugal, large gains also occur for male mortality, but losses in performance occur for female mortality. Small losses also occur for female mortality in populations for which forecast errors are low when using the independent method (Fig. 8.5). Overall, standard-coherent forecasting improves accuracy by 17% for female mortality and 41% for male mortality, while bias is reduced by 99% and 63% respectively. These results are generally consistent with the hypothesis that a low-mortality standard serves as a good guide to future mortality decline.

8.4.2.2 Other Standards

The efficacy of standard-coherent forecasting in improving forecast accuracy and bias clearly depends on the choice of standard. In this section, the three additional standards are considered; these are Spain, Switzerland and Australia. Summary results are shown in Table 8.2, comprising overall means and standard deviations, relative to independent forecasts, of accuracy and bias for the three methods and the four standards. (Note that as mean bias is a net measure, its size depends partly on the degree of counterbalancing of positive and negative biases; this explains the very low value for overall mean bias for female mortality when using Japan as standard, and also influences other values for bias.) For female mortality, the results obtained when using Spain, Switzerland and Australia as standard are similar to those for Japan as standard: the standard-coherent method improves accuracy and bias, and reduces across-country average heterogeneity across horizons. For male mortality, however, the effects are less consistent; when using Spain or Switzerland as standard, performance is reduced or only marginally improved.

8.5 Discussion

This analysis has evaluated the performance of two methods of coherent mortality forecasting in terms of the means and standard deviations of forecast accuracy and bias in female and male mortality in 17 low-mortality countries. The purpose of the evaluation was to test the hypothesis that low mortality serves as a good guide to future mortality when used in coherent forecasting, and high mortality does not. The findings support this hypothesis to a large extent but, for male mortality in particular, exceptions occur.

Table 8.2 Relative mean accuracy and bias by method and standard for female and male mortality, and relative sex-difference in accuracy and bias

	Independent	Sex-coherent	Japan-coherent	Spain-coherent	Switzerland-coherent	Australia-coherent
Female mortality						
Overall mean accuracy	1.000	1.108	0.834	0.861	0.894	0.791
Overall mean bias	1.000	1.318	0.005	0.453	0.791	0.576
Average SD(accuracy)	1.000	1.238	0.810	0.901	0.907	0.646
Average SD(bias)	1.000	1.430	0.803	0.812	0.919	0.649
Male mortality						
Overall mean accuracy	1.000	0.892	0.594	1.051	0.920	0.757
Overall mean bias	1.000	0.880	0.371	1.038	0.898	0.605
Average SD(accuracy)	1.000	0.793	0.439	1.080	0.970	0.685
Average SD(bias)	1.000	0.787	0.357	1.025	1.025	0.659
Sex-difference						
Overall mean accuracy	1.000	0.496	0.153	1.398	0.967	0.695
Overall mean bias	1.000	0.520	0.672	1.517	0.986	0.629
Average SD(accuracy)	1.000	0.339	0.061	1.263	1.034	0.725
Average SD(bias)	1.000	0.354	0.056	1.168	1.097	0.665

Note: Overall mean refers to mean accuracy and bias averaged over horizons and countries, *MARE* and *MRE* respectively. Average SD refers to the country-specific standard deviations across horizons, *SD_c(MARE)* and *SD_c(MRE)*, averaged over countries. The sex-difference refers to the difference in accuracy and bias measures (and not to the difference in mortality)

8.5.1 Support for the Low-Mortality Hypothesis

The results show that sex-coherent forecasting improves forecast performance, relative to independent forecasting, for male mortality but not for female mortality. Average gains in performance for male mortality forecasting range from 11% to 21%, while average losses in performance for female mortality forecasting amount to 11–43% (Table 8.2). Given lower female mortality than male mortality in all countries in the study (Table 8.1), both results support the hypothesis.

At the same time, standard-coherent forecasting with each of the four low-mortality standards improves performance for female mortality, with gains of 8–99%. Again, these results support the hypothesis that low mortality serves as a good guide to future mortality, given that all four countries used as standards have low mortality relative to almost all other countries in the study (Table 8.1). For male mortality, however, standard-coherent forecasting with these low-mortality standards is not always advantageous. While using Japan or Australia as standard improves performance by 24–63%, using Switzerland or Spain as standard produces small gains or losses in performance. (The results for Spain as standard (not shown) indicate that poor performance cannot be attributable to high or similar male mortality in Spain compared with six of the populations considered (Table 8.1).) Thus, in the case of male mortality, the hypothesis is only partially supported by standard-coherent forecasting.

Further, for both female and male mortality, the lowest-mortality standard of the same sex does not produce the greatest gains in performance. The best performing standard for female mortality is Australia, chosen on the basis of male mortality, while the best performing standard for male mortality is Japan, chosen on the basis of female mortality. However, Japan and Australia serve as the two best guides for both female and male future mortality. These findings point to choice of low-mortality standard as an important consideration (Kjærgaard et al. 2016; Stoeldraijer 2019).

8.5.2 Ranking of Methods

Considering only Japan or Australia as standard, the ranking of methods by performance varies as a result of the differential effect of sex-coherent forecasting. For female mortality, the best method is standard-coherent forecasting, followed by independent forecasting, with sex-coherent forecasting in third place. For male mortality, standard-coherent forecasting is again best, followed by sex-coherent forecasting and then independent forecasting. These rankings hold for accuracy and bias, and for means and standard deviations. In most cases, these rankings also hold over forecast horizons.

8.5.3 *Benefits of a Low-Mortality Standard*

In the case of Japan as standard, the average trajectories of mean accuracy and bias change steadily over horizon (Fig. 8.4) and similar patterns are found for most individual countries. The horizon effects for accuracy and bias are considerably reduced by standard-coherent forecasting and heterogeneity across countries is also reduced. Thus confidence in standard-coherent forecasts is considerably greater than in independent forecasts which systematically overestimate future mortality rates and underestimate future life expectancy. Standard-coherent forecasting is also advantageous in reducing forecast error due to particular mortality conditions. The latter may be partially manifest in jump-off error indicated by error at $h = 1$. Jump-off error is greater on average for male mortality than for female mortality and, like the horizon effect, is reduced by standard-coherent forecasting (Fig. 8.4). Additionally, heterogeneity among countries with respect to forecast accuracy and bias is substantially reduced by standard-coherent forecasting; this is seen by horizon in the lower quadrants of Fig. 8.4, and is also evident in the upper quadrants of Figs. 8.5 and 8.6.

8.5.4 *Homogenisation of Accuracy and Bias by Sex*

One of the features of sex-coherent forecasting noted by Hyndman et al. (2013) is the homogenisation of forecast accuracy and bias for female and male mortality by horizon. Because forecast errors are generally smaller for female mortality than for male mortality, the opposing effects of sex-coherent forecasting result in smaller sex-differences in accuracy and bias. Figure 8.7 (upper quadrants) shows that sex-coherent forecasting substantially reduces the sex-difference in forecast accuracy and bias at longer horizons, compared with independent forecasting. This is the case for 14 of the 17 countries (Fig. 8.7 lower quadrants) and on average the sex-difference is reduced by 50% for accuracy and 48% for bias (Table 8.2).

Homogenisation by sex of forecast accuracy and bias is also an outcome of standard-coherent forecasting with Japan as standard, as also seen in Fig. 8.7. Compared with independent forecasting, Fig. 8.7 shows that standard-coherent forecasting substantially reduces the sex-difference in both forecast accuracy and bias, and in 16 of the 17 countries, resulting in overall reductions of 85% for accuracy and 33% for bias. Thus for accuracy, homogeneity by sex is greatest for the standard-coherent method (with Japan as standard) while for bias, homogeneity is greatest for the sex-coherent method. In both cases, independent forecasts are least homogeneous.

Greater homogeneity by sex of accuracy and bias is a significant advantage for forecasting practice, as it reduces the likelihood of unbalanced forecasts of female and male mortality. Increased confidence in the internal consistency of mortality forecasts is of direct benefit in actuarial applications and in forecasting the age-sex structure of populations.

Fig. 8.7 Sex-difference (male – female) in accuracy and bias by method (Japan as standard)

Ratios of sex-differences in the overall means and standard deviations of accuracy and bias are shown for all methods and standards in Table 8.2. Sex-coherent forecasting reduces the sex-difference in the standard deviations of accuracy and bias by two-thirds. For standard-coherent forecasting, using Japan as standard very substantially reduces the sex-difference in performance while using Australia as standard reduces it by about one third. However, using Spain as standard is consistently disadvantageous for male mortality and hence for sex-differences, while using Switzerland as standard has little effect.

8.5.5 *Strengths of the Study*

An important and purposeful feature of this study is the use of relative measures of accuracy and bias: *MARE* and *MRE*. These measures aggregate and average the proportional forecast errors in age-specific mortality rates, with equal weight to each age, and are thus comparable across mortality levels and age pattern. This means that they are also comparable across horizons, countries and sex; differences and ratios are also comparable. This is a major strength of the study. Non-proportional errors, which are typically larger for higher rates, are influenced by decreasing mortality and portray a conservative horizon effect. In this study, increases in *MARE* and *MRE* with increasing horizon are not influenced by level of mortality.

A second strength is the use of a rolling fitting period, designed to avoid systematic effects in forecast errors arising from random temporal variation in the data in the fitting or forecast periods. By averaging over fitting periods, the effects of jump-off year (jump-off error), calendar year in the forecast period and horizon are averaged (Fig. 8.1). The fixed length of the rolling fitting period has little effect on forecast error, relative to a fixed first year. Indeed, the first year of the fitting period advances from 1950 to 1972, the latter completely omitting the period of mortality stagnation in the 1960s experienced in many low-mortality countries. By using a rolling start year of fixed length, the study takes into account as broad a range of mortality situations as possible.

The comparison of the three methods is further validated by their common use of FDM with identical parameters. Thus the comparison reveals the effects of taking other mortality into account through coherent forecasting. The two coherent methods are also directly comparable: the sex-coherent method is in fact a special case of the standard-coherent method where the standard is the other sex.

It is also of note that the study has a theoretical basis. Most studies introducing new methods have focussed on technical aspects and have been largely experimental.

8.5.6 *Limitations*

A common criticism of forecasting with the aid of a standard is that the standard itself is not forecast. In this study, the standard represents low mortality. As has been shown, it would be inappropriate to use coherent methods to forecast the standard

because the other mortality would by definition be higher. In using standard-coherent forecasting, the forecast of the standard is not of interest. Rather, the standard should be forecast using the independent method, bearing in mind that such forecasts tend to overestimate future mortality (Fig. 8.8). The gains in accuracy for mortality forecasts for all other countries far outweigh this limitation. Further, it should be noted that the method does not require that the standard be forecast.

It should be borne in mind that the means and standard deviations of accuracy and bias are derived from the same forecasting errors in age-specific mortality rates. Given the nature of mortality data, large errors tend to be associated with less regular age and time patterns of change, which also produce large standard deviations across horizons and countries. Patterns across horizons and countries can therefore be expected to be similar. It should also be noted that the study uses heterogeneity in average error (the units of analysis) by horizon and country to assess robustness of methods. Given averaging over rolling fitting period, the study does not assess the accuracy of forecasts for individual calendar years in the forecast period.

The equal weight allocated to each age in the relative measures of accuracy and bias may be regarded as a limitation in situations where emphasis is required on ages where mortality rates are high. Weighting of *MARE* and *MRE* by age would address this requirement while still retaining the advantage of comparability across populations. In other circumstances, the mean absolute error and mean error may be used, but comparability across horizons, countries and sexes would be lost.

A further limitation of this study is that interval forecasts (Shang et al. 2011) are not considered. Conceptually, this follows from analysing accuracy and bias based on errors in the point forecast of age-sex-specific mortality rates, rather than on errors in the forecast distribution of these rates. Further, the rates used in calculation of these measures are net of random observational error by virtue of the smoothing procedure integral to functional data modelling, and the measures are further stabilised by averaging over age and fitting period. Thus, a significant component of the error contributing to the prediction interval of the forecast is excluded from consideration. Further research is needed to address the accuracy of prediction intervals in the framework of this analysis.

8.6 Conclusion

Coherent forecasting offers one approach to the reduction of error in mortality forecasts. Using the product-ratio method of coherent forecasting with functional data models, this study has shown that coherent forecasting with an empirical low-mortality standard can be highly advantageous in terms of forecast performance. Low-mortality-coherent forecasting has the ability to increase accuracy, reduce bias and limit the heterogeneity in these measures. Additionally, sex-differences in forecast performance are reduced, producing greater homogeneity by sex of accuracy and bias, thereby increasing confidence in forecasts by sex. These are important advantages in real world forecasting.

This study has provided clear guidance for female and male mortality forecasting. In both cases, a same-sex low-mortality-standard is optimal. For male mortality, sex-coherent forecasting is also advantageous, on average, but for female mortality sex-coherent forecasting is counterproductive. This study has identified Japan and Australia as the two standards producing the best forecast performance for both female and male mortality in the recent past, while Spain and Switzerland are much less useful as standards. Why this is so remains unclear. The hypothesis that low-mortality is a good guide to future mortality is largely supported, but the role of other features of the standard need further investigation.

Appendix

Fig. 8.8 Units of analysis: accuracy and bias by horizon and country for female and male mortality by method (Japan as standard). Note: Country average shown in black

Fig. 8.8 (continued)

References

- Ahmadi, S. S., & Li, J. S.-H. (2014). Coherent mortality forecasting with generalized linear models: A modified time-transformation approach. *Insurance: Mathematics & Economics*, 59, 194–221. <https://doi.org/10.1016/j.insmatheco.2014.09.007>.
- Basellini, U., & Camarda, C. G. (2019). Modelling and forecasting adult age-at-death distributions. *Population Studies*, 1–20. <https://doi.org/10.1080/00324728.2018.1545918>.
- Bergeron-Boucher, M.-P., Canudas-Romo, V., Oeppen, J., & Vaupel, J. W. (2017). Coherent forecasts of mortality with compositional data analysis. *Demographic Research*, 37, 527–566.
- Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, 22(3), 547–581. <https://doi.org/10.1016/j.ijforecast.2006.04.001>.
- Booth, H., & Tickle, L. (2008). Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science*, 3(I/II), 3–43.
- Booth, H., Maindonald, J., & Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56(3), 325–336. <https://doi.org/10.1080/00324720215935>.
- Booth, H., Hyndman, R. J., Tickle, L., & De Jong, P. (2006). Lee-Carter mortality forecasting: A multi-country comparison of variants and extensions. *Demographic Research*, 15(9), 289–310.
- Booth, H., Hyndman, R. J., & Tickle, L. (2014). Prospective life tables. In *Computational actuarial science, with R* (pp. 319–344). Boca Raton: CRC Press. <https://doi.org/10.1201/b17230-11>.
- Buettner, T. (2002). Approaches and experiences in projecting mortality patterns for the oldest-old. *North American Actuarial Journal*, 6(3), 14–29. <https://doi.org/10.1080/10920277.2002.10596053>.
- Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., & Khalaf-Allah, M. (2011). Bayesian stochastic mortality modeling for two populations. *ASTIN Bulletin*, 41(1), 29–59. <https://doi.org/10.2143/AST.41.1.2084385>.
- Camarda, C. G. (2019). Smooth constrained mortality forecasting. *Demographic Research*, 41(38), 1091–1130. <https://doi.org/10.4054/DemRes.2019.41.38>.
- Currie, I. D. (2014). On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal*, (July), 1–28. <https://doi.org/10.1080/03461238.2014.928230>.
- de Beer, J., & Janssen, F. (2016). A new parametric model to assess delay and compression of mortality. *Population Health Metrics*, 14(1), 1–21. <https://doi.org/10.1186/s12963-016-0113-1>.
- de Jong, P., & Tickle, L. (2006). Extending Lee-Carter mortality forecasting. *Mathematical Population Studies*, 13, 1–18.
- Fazle Rabbi, A. M. (2019). *Modified Lee-Carter methods with LASSO type smoothing and adjusting for lifespan disparity*. Ph.D. thesis. University of Padua.
- Human Mortality Database. (2019). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Retrieved from <http://www.mortality.org> on 30 June 2019.
- Hyndman, R. J., & Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, 24(3), 323–342. <https://doi.org/10.1016/j.ijforecast.2008.02.009>.
- Hyndman, R. J., & Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis*, 51(10), 4942–4956. <https://doi.org/10.1016/j.csda.2006.07.028>.
- Hyndman, R. J., Booth, H., & Yasmeen, F. (2013). Coherent mortality forecasting: The product-ratio method with functional time series models. *Demography*, 50(1), 261–283. <https://doi.org/10.1007/s13524-012-0145-5>.
- Janssen, F. (2018, December 1). Advances in mortality forecasting: Introduction. *Genus*. Springer International Publishing. <https://doi.org/10.1186/s41118-018-0045-7>
- Kjærgaard, S., Canudas-Romo, V., & Vaupel, J. W. (2016). The importance of the reference populations for coherent mortality forecasting models. In *European population conference*, Mainz, Germany.

- Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87(419), 659–671.
- Li, J. (2012). A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. *Population Studies*, (November), 37–41. <https://doi.org/10.1080/00324728.2012.689316>.
- Li, N., & Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, 42(3), 575–594. <https://doi.org/10.1353/dem.2005.0021>.
- Li, H., & Li, J. S.-H. (2017). Optimizing the Lee-Carter approach in the presence of structural changes in time and age patterns of mortality improvements. *Demography*, 54(3), 1073–1095. <https://doi.org/10.1007/s13524-017-0579-x>.
- Li, N., Lee, R., & Gerland, P. (2013). Extending the Lee-Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography*, 50(6), 2037–2051. <https://doi.org/10.1007/s13524-013-0232-2>.
- Pascariu, M. D., Canudas-Romo, V., & Vaupel, J. W. (2018). The double-gap life expectancy forecasting model. *Insurance: Mathematics & Economics*, 78, 339–350. <https://doi.org/10.1016/j.insmatheco.2017.09.011>.
- Raftery, A. E., Chunn, J. L., Gerland, P., & Sevcíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries – Springer. *Demography*, 50, 777–801. Retrieved from http://link.springer.com/article/10.1007/s13524-012-0193-x/fulltext.html?wt_mc=alerts.TOCjournals.13524.
- Renshaw, A. E., & Haberman, S. (2003). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics & Economics*, 33(2), 255–272. [https://doi.org/10.1016/S0167-6687\(03\)00138-0](https://doi.org/10.1016/S0167-6687(03)00138-0).
- Renshaw, A. E., & Haberman, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics & Economics*, 38(3), 556–570. <https://doi.org/10.1016/j.insmatheco.2005.12.001>.
- Shang, H. L., Booth, H., & Hyndman, R. (2011). Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. *Demographic Research*, 25(5), 173–214. <https://doi.org/10.4054/DemRes.2011.25.5>.
- Shumway, R. H., & Stoffer, D. S. (2006). *Time series analysis and its applications: With R examples* (2nd ed.). New York: Springer.
- Stoeldraijer, L. (2019). An evaluation of methods to coherently forecast mortality based on both quantitative and qualitative criteria. In *Mortality forecasting in the context of non-linear past mortality trends: An evaluation*. University of Groningen. Retrieved from [https://www.rug.nl/research/portal/nl/publications/mortality-forecasting-in-the-context-of-nonlinear-past-mortality-trends-an-evaluation\(70a8dbec-628a-42b9-afc6-cff076639040\).html](https://www.rug.nl/research/portal/nl/publications/mortality-forecasting-in-the-context-of-nonlinear-past-mortality-trends-an-evaluation(70a8dbec-628a-42b9-afc6-cff076639040).html).
- Tickle, L., & Booth, H. (2014). The longevity prospects of Australian seniors: An evaluation of forecast method and outcome. *Asia-Pacific Journal of Risk and Insurance*, 8(2), 1–34. <https://doi.org/10.1515/apjri-2013-0004>.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.