MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Konrad-Zuse-Strasse 1 · D-18057 Rostock · Germany · Tel +49 (0) 3 81 20 81 - 0 · Fax +49 (0) 3 81 20 81 - 202 · www.demogr.mpg.de

MPIDR Working Paper WP 2023-020 l April 2023
https://doi.org/10.4054/MPIDR-WP-2023-020

A Validation Workflow for Mortality Forecasting

Ricarda Duerst l duerst@demogr.mpg.de
Jonas Schöley l schoeley@demogr.mpg.de
Christina Bohk-Ewald

This working paper has been approved for release by: Silvia Loi (loi@demogr.mpg.de),
Deputy Head of the Laboratory of Population Health.

A Validation Workflow for Mortality Forecasting

Ricarda Duerst[1,2], Jonas Schöley[1], Christina Bohk-Ewald[3]
[1] Max Planck Institute for Demographic Research Rostock, Germany
[2] University of Helsinki, Finland
[3] varvis®, Limbus Medical Technologies GmbH

Abstract

Accurate mortality forecasts are essential for decision makers to plan for changing needs of pension and other social security systems. Researchers have developed a variety of methods with increasing methodological complexity to forecast mortality developments. We introduce a method validation workflow designed for mortality forecasts. The aim of our workflow is to assess the suitability of forecast method depending on the prevailing mortality regime in the country of interest. For our analysis, we apply our workflow to short-term Lee-Carter forecasts for 24 countries to showcase different mortality regimes. We assess Lee-Carter's forecast performance on the life expectancy and lifespan disparity at birth. We show that the mortality regime in the country of interest plays a crucial role for the performance of a forecast method. Thus, our method validation workflow helps researchers to choose an appropriate mortality forecast method.

1. Introduction

When forecasting age-specific mortality, one must choose among an ever growing body of available methods – many of which are developments of the seminal Lee-Carter approach (Lee and Carter, 1992). Mortality forecasts inform policy decisions on pension systems, health care and other social security domains. Thus, choosing a suitable method and understanding its limits of applicability is crucial as the forecasts have real-world consequences. Here, we introduce a method validation workflow specifically designed for forecasts of mortality by age. This workflow aims at assessing the suitability of a forecasting method depending on the prevailing mortality regime in the population of interest.

Since the introduction of the Lee-Carter model in 1992 (Lee and Carter, 1992), several extensions have been proposed (Booth et al., 2006; Booth and Tickle, 2008; Basellini et al., 2022). For example, considering the advancement of survival improvements to increasingly older ages (Rau et al., 2008), more recent approaches account for trends in rates of mortality improvement and in the distribution of ages at death (Haberman and Renshaw, 2012; Li et al., 2013; Ševčíková et al., 2016; Bohk-Ewald and Rau, 2017; de Beer et al., 2017; Bardoutsos et al., 2018; Basellini and Camarda, 2019; Camarda, 2019). Other methodological trends in mortality forecasting are to account for health behaviour such as smoking, obesity and alcohol consumption (Vogt et al., 2017; Janssen et al., 2013; Wang and Preston, 2009; Janssen et al., 2021) and for mortality developments in benchmark countries (Li and Lee, 2005; Hyndman et al., 2013; Raftery et al., 2013).

The range of available methodology challenges the forecaster to choose the most suitable model for the task at hand. While recent work on demographic and mortality forecasting recognizes the importance of method validation (Rizzi et al., 2021; Basellini and Camarda, 2019; Camarda, 2019; Bohk-Ewald et al., 2018, 2017; Shang, 2015, 2012; Li et al., 2013; Booth et al., 2006; Shang et al., 2011) there is no consensus on a universally optimal mortality forecasting algorithm and it has been argued that the optimal model depends on the population under consideration. Model selection algorithms have been developed to automatically choose a suitable model based on the characteristics of the time series to forecast (Poler and Mula, 2011; Hyndman et al., 2002; Venkatachalam and Sohl, 1999; Sohl and Venkatachalam, 1995).

In contrast to the existing model selection literature, we are proposing a validation design for model selection, specifically created for age-specific mortality forecasts. In our case, the characteristics of the time series translate into a specific mortality regime. Thus, we hypothesise that different mortality forecast models perform differently, depending on the mortality regime of the research subject of interest.

To answer the suitability question, we develop an extensive validation workflow. We validate a method's forecast performance on two common measures of mortality: life expectancy at birth and lifespan disparity at birth (van Raalte et al., 2018; Vaupel and Romo, 2003). Further, we distinguish three analytical settings to analyse how much a model's forecast performance is influenced by

High vs. low mortality regimes via the average level of life expectancy and lifespan disparity in forecast years (as they represent mortality settings with distinct patterns of mortality over age that may (or may not) be difficult to capture by a forecast method).

Dynamic vs. static mortality regimes via the annual rate of change in life expectancy and lifespan disparity in forecast years (as they represent mortality settings with different levels of continuous mortality improvement that may (or may not) be difficult to capture by a forecast method).

Monotonic vs. non-monotonic mortality regimes via the trend change in life expectancy and lifespan disparity between observed years and forecast years (as they represent mortality settings with abrupt changes in mortality improvement from observed to forecast years that are difficult to capture for each forecast method).

In the remainder of this paper, we present our validation workflow and demonstrate its application exemplary for one forecast method on a large data basis. We show that the Lee-Carter method (Lee and Carter, 1992) is suitable to forecast the mortality of many high-income, low-mortality countries, during the most recent years. However, there have been mortality regimes in the past during which Lee-Carter's forecasts performance was unsatisfactory. This applies to the late nineteenth century to mid twentieth century, when discontinuous mortality developments have been prevalent in many of those countries. Further, we discuss limitations and validate our workflow using extended data sources and methods.

2. Methods and Data

2.1 Conceptual Validation Workflow

Our validation workflow for increasing confidence in using (or not using) a mortality forecast method for a country of interest consists of four aspects: the validation design, the mortality measure, the mortality regime, and the forecast performance. Based on these aspects, we will describe our conceptual validation workflow and further explore the requirements for its successful application.

2.1.1 Validation Design

We adopt a rolling out-of-sample validation design, also called rolling cross-validation, to assess the suitability of a forecast method. In an out-of-sample validation design, the available data is divided into a training data set and a testing data set. In our design, the data from the training set is the input for a mortality forecast. We call the calendar years of mortality data used as input the base period. The resulting mortality forecast stretches over a period that we call the forecast horizon and that covers the same years as the testing data set. Therefore, we can compare the forecast with the observed mortality in the forecast horizon. The last calendar year of the base period is called the jump-off year after which the forecast horizon starts. Depending on the coverage of the available data, and the chosen length of base period and forecast horizon, we then roll the observation window across the calendar years. This results in several mortality forecasts with different jump-off years that can be compared to the respective observed mortality data. For example, given a country with available mortality data for 70 subsequent calendar years ($t_1, ..., t_{70}$), a chosen base period of 30 years, and a forecast horizon of 20 years, we would generate a total of 22 forecasts starting after the jump-off years $t_{29}, t_{30}, ..., t_{50}$.

2.1.2 Mortality Measures

In our workflow, the validation of a mortality forecast takes place on two measures of mortality: life expectancy at birth (e_0) and life span disparity at birth ($e_0^\dagger$). Bohk-Ewald et al. (2017) argue that mortality forecast methods not only need to capture the change in life expectancy, but also the variation in the life span distribution e.g., the compression, shifting, and expansion of mortality. Both mortality measures are calculated using period life tables (see Preston et al. (2001, p. 49) for formulas of the life table and Zhang and Vaupel (2009) for life span disparity) and are based on the forecast age-specific mortality rates (m_x).

2.1.3 Mortality Regimes

We consider three mortality regimes in which the evaluation of the forecast performance takes place. For the high vs. low regimes, we analyse the forecast performance in relation to the mean value of our mortality measures e_0 and $e_0^\dagger$ in the forecast horizon. Our goal is to see whether a forecast method can capture different levels of mortality and their associated age patterns. For the dynamic vs. static regimes, we assess whether the trend of the mortality measures over time plays a role in the forecast performance, e.g. how well can a forecast method capture moderate to strong shifts in the level and age pattern of mortality in the forecast years? To do so, we calculate the annual rate of change of the mortality measures in the forecast horizon and analyse it in relation to the forecast performance. Lastly, we compare monotonic vs. non-monotonic regimes by comparing forecast performance in relation to the trend change of e_0 and $e_0^\dagger$. We define the trend change as the difference between the annual rate of change in the forecast horizon and the annual rate of change in the base period. We perform all analyses separately for females and males.

2.1.4 Forecast performance

We use the point estimates of a mortality forecast method to quantify the forecast performance. First, we calculate the forecast error Δ_t which is defined as the difference between the forecast value F_t and the observed value Y_t in year t,

$$\Delta_t = F_t - Y_t. \tag{1}$$

From there, we calculate several error measures that quantify two aspects of the forecast performance: bias and accuracy. Regarding forecast bias, the percentage error is a measure that indicates overestimation of the observed mortality measures for values larger than zero and underestimation, conversely:

$$\mathrm{PE}_t = 100 \times \Delta_t / Y_t \tag{2}$$

In terms of accuracy, Shcherbakov et al. (2013) recommend the use of more than one measure. To quantify the forecast accuracy, we use the absolute percentage error (APE)

$$\mathrm{APE} = |\mathrm{PE}_t| \tag{3}$$

and two summarizing error measures based on the forecast error: the mean square error

$$\mathrm{MSE} = \mathrm{mean}(\Delta_t^2) \tag{4}$$

and the root mean square error

$$\mathrm{RMSE} = \sqrt{\mathrm{mean}(\Delta_t^2)}. \tag{5}$$

2.2 Data Sources

Based on the data requirements for a successful application, we choose data from the open-access Human Mortality Database (**?**) for the application of our validation workflow. The HMD provides high-quality population and mortality data for 41 industrialised countries. For women and men, we extracted mortality rates (m) by single ages (x) from zero to 110 and above, from life table data. However, data coverage across calendar years varies by country, ranging from the longest time span of 267 years (1751 to 2017) for Sweden to 62 years (1956 to 2017) for Germany.

To validate on other mortality regimes and patterns beside the low-mortality countries found in the HMD, we conduct additional validation on data from the latest edition of the United Nations World Population Prospects (**?**). The UN provides mortality and population data on 209 countries and regions of the world (for an overview of world regions, subregions and countries see Table 3 in the Appendix) which allows for a comprehensive application of our validation workflow.

However, the structure of the UN data forces us to perform additional data preparation steps prior to the application of our validation workflow. Deaths and population counts from the UN WPP 2019 are classified in age groups and groups of calendar years from 1950 to the present. Further, the provided data are estimates of the demographic changes, based on numerous empirical data sources (for detailed methodology, see **?**). The process of un-grouping and preparing the data is documented in the Appendix.

3. Results

3.1 Application of Validation Workflow

To showcase the application of our validation workflow, we assess the suitability of the Lee-Carter method (Lee and Carter, 1992) to forecast female mortality of 24 highly developed countries 20 years ahead. The Human Mortality Database (HMD) provides at least 50 consecutive years of data for these countries that we require to calculate forecast errors 20 years ahead using a base period of 30 years (see Table 2 for list of countries and data availability).

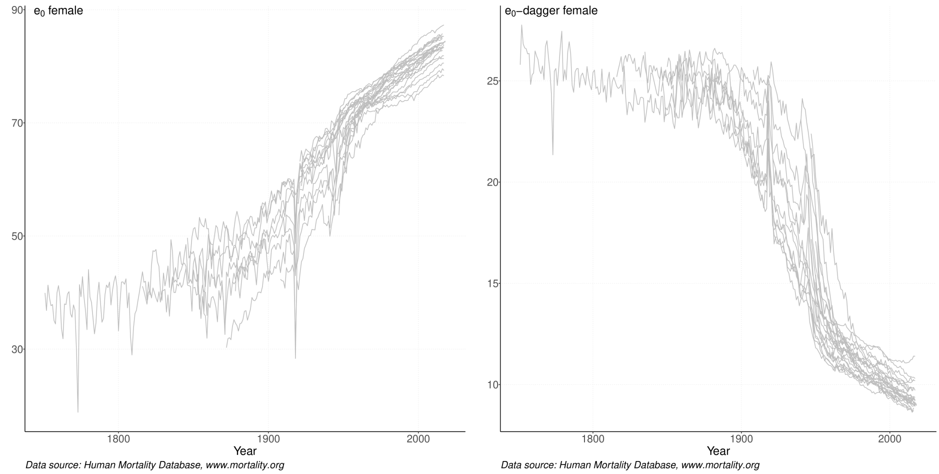
We selected the countries based on the data demands of our validation design. We need 50 consecutive years of data for the exemplary analysis of percentage errors (PE) for 20-years forecasts using a base period (bp) of 30 years.

The general trend of mortality in the chosen high-income, low-mortality countries over the available calendar years is characterized by increasing life expectancy at birth (see Figure 1, a) and decreasing lifespan disparity at birth (b). The mortality development of these countries follows a similar trajectory: after strong fluctuations in the nineteenth century, life expectancy at birth has increased rapidly in the twentieth century, interrupted by country-specific events like the Spanish flu from 1918 to 1920. Since the mid-nineteen-hundreds, increases in life expectancy at birth are steady and smooth compared to earlier centuries. In the year 2016, e_0 ranged from 78.47 (Bulgaria) to 87.17 (Japan) years of life for females. As life expectancy has improved, lifespan disparity at birth has decreased in a reversed similar pattern: since the early twentieth century, $e_0^\dagger$ has decreased rapidly, interrupted by few major mortality shocks. Since the late twentieth century we find a rather smooth and moderate decline in lifespan disparity. For females, $e_0^\dagger$ ranged from 11.41 (USA) to 8.73 (Spain) life years lost in 2016.

The characteristics of the past mortality development will affect the performance of the Lee-Carter forecasts. The Lee-Carter model is an extrapolative method that assumes that past trends will continue in the future. Therefore, we expect higher forecast errors for forecasts with jump-off years (JOY) lying in between periods of mortality development with different slopes. E.g., a Lee-Carter forecast of life span disparity for Sweden using data from 1920 to 1949 and trying to forecast life span disparity in 1970 will likely overestimate the mortality improvement and therefore underestimate the actual observed life span disparity, resulting in a higher, negative percentage error.

Figure 2 shows the visual analysis of the three mortality regimes for Lee-Carter forecasts 20 years ahead according to our validation workflow for females. The colour gradient from black over purple and orange to yellow depicts the range of JOY from 1780 to 1997. Only a few countries with the longest periods of available data (Sweden, Denmark, Norway, Netherlands) are represented in the earliest JOY while all 24 countries are represented in the more recent JOY. In all six individual panels of the figure, we display the forecast percentage error (PE) on the vertical axis. A PE larger than zero indicates overestimation and a value smaller than zero shows underestimation. We show the PE for life expectancy at birth (e_0) in the left column and for lifespan disparity ($e_0^\dagger$) in the right column.

Figure 1: Female mortality development of 24 industrialised countries from 1751 to 2019. Data source: Human Mortality Database (HMD).



(a) Life Expectancy at Birth **(b)** Life Span Disparity at Birth

High vs. low mortality regimes: PE vs. average level of e_0 and $e_0^\dagger$ in forecast years In the top row of Figure 2, we plot the PE in relation to the mean value of mortality in the forecast horizon of 20 years (horizontal axis) for life expectancy at birth and lifespan disparity. Overall, we find that the mean level of e_0 appears to have a smaller effect on the PE of Lee-Carter forecasts than the mean level of $e_0^\dagger$. Specifically, the overall small PE of e_0 shifts horizontally to the right from earlier to recent JOY (purple to yellow) and from low to high average levels of e_0 (35 to 85 years). At the same time, we find that the variance of the PE for e_0 decreases for later JOY: while the Lee-Carter method mostly underestimates life expectancy at birth slightly for mid-level JOY (red and orange) it performs well in terms of forecast accuracy and bias for the most recent JOY (yellow). Regarding the PE for $e_0^\dagger$, we find a backward shift, in the form of an arch, from high levels (25 years) to low levels (10 years) of life years lost from earlier to later JOY (black to yellow). We find that the Lee-Carter method cannot properly capture the strong decrease in lifespan disparity in the mid-level JOY (red and orange) and, consequently, tends to overestimate $e_0^\dagger$. Here, bias and inaccuracy are even more accentuated than for e_0 (PE up to +80% as opposed to +20%). However, for the earliest and most recent JOY (purple and yellow), the Lee-Carter method produces accurate and unbiased results for $e_0^\dagger$.

Dynamic vs. static mortality regimes: PE vs. annual rate of change of e_0 and $e_0^\dagger$ in forecast years The middle row of Figure 2 shows the relationship of the PE with the annual rate of change of e_0 and $e_0^\dagger$ in the forecast horizon (fh) 20 years ahead. If the rate is smaller than zero, e_0 resp. $e_0^\dagger$ have decreased in the forecast horizon. For values larger than zero, e_0 and $e_0^\dagger$ have increased. The larger the absolute value of the annual rate of change, the stronger is the change in e_0 and $e_0^\dagger$ (positive or negative) in the forecast years. Overall, we observe a strong effect of the annual rate of change of e_0 and $e_0^\dagger$ on accuracy and bias of the Lee-Carter forecasts.

For early and mid-level jump-off years (purple to orange), the Lee-Carter method cannot capture the strong increase in life expectancy (+0.5 to +1 years of life) in the forecast horizon. This results in forecast inaccuracy and mostly underestimation. However, the Lee-Carter method performs well for the most recent jump-off years (yellow) where there is a more moderate increase in life expectancy (+0 to +0.3 years of life) in many countries.

We see a similar relationship between the annual rate of change and PE for lifespan disparity at birth. The larger the decrease in $e_0^\dagger$ is (up to -0.5 years of life lost), the heavier is the overestimation of the Lee-Carter forecasts (PE up to 80%). Moving towards an annual rate of change of zero in the earliest and most recent jump-off years (purple and yellow), the Lee-Carter forecasts become more accurate.

Monotonic vs. non-monotonic mortality regimes: PE vs. trend change in e_0 and $e_0^\dagger$ between observed and forecast years In the bottom row of Figure 2, we show the relationship of the PE with the trend change in mortality from the base period (bp) to the forecast horizon (fh). If the annual rate of change in the forecast horizon is smaller than the rate in the base period, the trend change is smaller than zero. If the annual rate of change in the forecast horizon exceeds the rate in the base period, the trend change is larger than zero. Trend changes in mortality are especially hard to capture in forecasts as they are always unexpected. Overall, we find that trend changes appear to have a strong impact on the PE of e_0 and $e_0^\dagger$.

As we see from the figures, trend changes have happened for both life expectancy and lifespan disparity, resulting in higher forecast inaccuracy especially in the early and middle jump-off years (purple to orange). However, a trend change from smaller increases in life expectancy in the base period to larger increases in life expectancy in the forecast horizon (positive trend change) effects the PE of e_0 more strongly than a negative trend change. Further, positive trend changes in e_0 are related to underestimation (PE up to -70%) while negative trend changes rather cause overestimation (PE up to +40%). For the most recent jump-off years (yellow), the Lee-Carter method gives more accurate results as the trend change is closer to zero.

Trend changes in lifespan disparity take place particularly for mid-level jump-off years and range from -0.4 to 0.4 (orange). Here, the negative relationship between PE $e_0^\dagger$ and trend change is even more accentuated than for e_0. The Lee-Carter method penalizes strong negative trend changes with severe overestimation (PE up to +60%) and positive trend changes with underestimation (PE up to -20%). As the trend change for the most recent jump-off years is close to zero (yellow), Lee-Carter performs well in terms of forecast accuracy. However, we observe slight overestimation of $e_0^\dagger$.

Initial Recommendation Overall, the analyses above show that the Lee-Carter method is suitable for forecasting mortality of the 24 HMD countries in the most recent jump-off years 1980 to 1997. For these cases, judging from the forecast percentage error, the forecast results are mostly accurate and unbiased. The Lee-Carter method cannot capture drastic changes in the annual rate of change and strong trend changes, resulting in bias and loss of accuracy. These characteristics apply to the mid-level jump-off years, where a transition from high-level to low-level mortality has taken place. Therefore, we would recommend the Lee-Carter method for high or low, rather static and monotonic mortality regimes.

3.2 Sensitivity of Results

We check the sensitivity of our results for the Lee-Carter method in several ways. First, we perform the previous analysis for male mortality. Figure 3 shows the percentage errors in the three analytical settings for males and can be found in the Appendix. We observe similar patterns in the relationship of the PE with the mean level, the annual rate of change, and the trend change of e_0 and $e_0^\dagger$. However, the overall level of mortality is higher for males than for females. The variance in the PE for both e_0 and $e_0^\dagger$ is larger for males compared to females.

Second, we use additional forecast error measures to judge the forecast performance. Table 1 shows three additional forecast error measures that give an insight on the overall bias and accuracy of forecast for e_0 and $e_0^\dagger$. The mean percentage error (MPE), the mean absolute percentage error (MAPE) and the root mean square error (RMSE) were calculated for 20-years Lee-Carter forecasts over all possible JOY for each country. Then we took the arithmetic mean over the error measures for countries belonging to each world region (see Table 2 in the Appendix for classification of HMD countries into regions). In general, we find that the choice of error measure influences the results. Regarding the forecast bias, the MPE shows that the forecasts mainly underestimate e_0 and overestimate $e_0^\dagger$. Only exceptions are Japan and Northern American females for which the opposite is the case. There are differences between the accuracy measures MAPE and RMSE. The RMSE indicates higher accuracy for forecasts of $e_0^\dagger$, compared to the accuracy for e_0. However, the MAPE shows more inaccurate results for $e_0^\dagger$ than for e_0. Regarding the different world regions, the error measures indicate different world regions as the best (highlighted in *italic font*) and worst (highlighted in **bold font**) forecast performance. For example, judging from the RMSE, forecasts for Southern Europe are most inaccurate for both e_0 and $e_0^\dagger$. However, by the MPE, forecast errors for Japan are the lowest for e_0 but the highest for $e_0^\dagger$. In addition, Table 1 confirms the results from the graphical analysis of analytical setting that forecasts of male mortality are generally more inaccurate compared to forecasts of female mortality.

Third, we use data from the United Nations World Population Prospects (**?**) to assess the forecast performance of Lee-Carter for other mortality regimes. The data from the UN had to be prepared in several aspects before forecasts could be estimated and assessed. For the full procedure of data preparation and classification of countries into world regions, see the Appendix.

Figure 4 in the Appendix shows the mortality development from 1950 to 2019 by different world regions measured by e_0 and $e_0^\dagger$. In contrast to the mortality of countries from the

Table 1: Forecast accuracy and bias of 20-years Lee-Carter forecast by world region and sex over all possible Jump-off years.

World Region	**Sex**	**MPE**		**MAPE**		**RMSE**	
		e0	**ed**	**e0**	**ed**	**e0**	**ed**
Eastern Europe	female	-2,4	*1,2*	2,9	6,9	2,5	0,8
	male	**-6,5**	4,2	**7,3**	8,1	5,6	1,1
Northern Europe	female	-3,0	3,8	**4,2**	8,6	3,5	1,5
	male	-4,3	3,5	5,9	7,3	4,9	1,5
Western Europe	female	-2,5	3,1	3,1	5,4	2,9	1,0
	male	-4,3	4,7	5,8	6,7	4,9	1,3
Southern Europe	female	**-3,6**	4,7	3,6	8,4	**4,3**	**1,7**
	male	-5,4	6,9	6,0	**11,6**	**6,7**	**2,4**
Northern America	female	0,5	-2,1	*0,9*	*3,7*	*0,9*	*0,5*
	male	-2,1	*1,4*	2,2	*3,8*	1,9	*0,6*
Australia & New Zealand	female	-1,9	5,3	1,9	7,2	1,7	0,8
	male	-4,7	5,0	4,8	6,5	3,9	0,8
Japan	female	*0,1*	**-13,1**	*0,9*	**13,1**	*0,9*	1,3
	male	*1,3*	**-7,6**	*1,3*	7,6	*1,1*	1,0

Source: Own calculations with data from the HMD **?**

HMD, there are differences in the trajectories of the world regions. While the overall trend is increasing for e_0 and decreasing for $e_0^\dagger$ for all world regions, there are differences in the onset and slope of mortality improvements. E.g., we observe a stagnation of e_0 in Africa from approximately 1990 to 2000 and unmatched strong declines of $e_0^\dagger$ in Europe in the 1950s. Further, there are difference in the general level of mortality. Europe, Northern America, and Oceania have the highest life expectancy and lowest life span disparity throughout the observed period. Africa's mortality is declining the slowest, compared to the other world regions. Northern America, and Latin America and the Caribbean are catching up to the mortality levels of Oceania. However, a gap remains in the latest observed years.

We performed the graphical analysis of percentage errors in the three analytical settings in the same way as for the HMD data. Figure 5 in the Appendix shows the Percentage Error (PE) of 20-years forecasts for females in relation to a) the mean level of mortality in the forecast horizon (fh), b) the annual rate of change in mortality in the fh, and c) the trend change in mortality from base period (bp) to fh, for e_0 (left panels) and $e_0^\dagger$ (right panels). The results show that there are differences in the forecast performance of Lee-Carter depending on world region. Lee-Carter produces the most unbiased and accurate results for Jump-off years that range between 1985 and 1995 for e_0 depending on world region. Forecasts with the earliest or latest jump-off years tend to be the most biased and inaccurate, with a few exceptions. However, in general, the forecast performance for Lee-Carter is acceptable for most of the world regions with a PE ranging between -5 and 5 for e_0 and -10 and 10 for $e_0^\dagger$ (excluding the errors for Africa).

Regarding the PE vs. the mean level of mortality in the forecast horizon (Figure 5 in Appendix), we do not observe the same decrease of PE with increasing JOY as for the

HMD data for e_0, nor the arch-shaped trajectory of PE over time for $e_0^\dagger$. There is an effect of the level of e_0 and $e_0^\dagger$ on the PE visible for Africa. Compared to the other world regions, the forecasts for Africa are more inaccurate for the earliest and latest jump-off years. Further, the world regions follow different trajectories: for e_0, PE is decreasing with increasing JOY for Africa, Asia, and Europe but increasing for Latin America and the Caribbean. Oceania and Northern America have periods of increasing and decreasing PE, although for low ranges of PE. Regarding $e_0^\dagger$, there is an increasing trend of PE with increasing JOY for most of the world regions with exception of Latin America and the Caribbean, and Northern America.

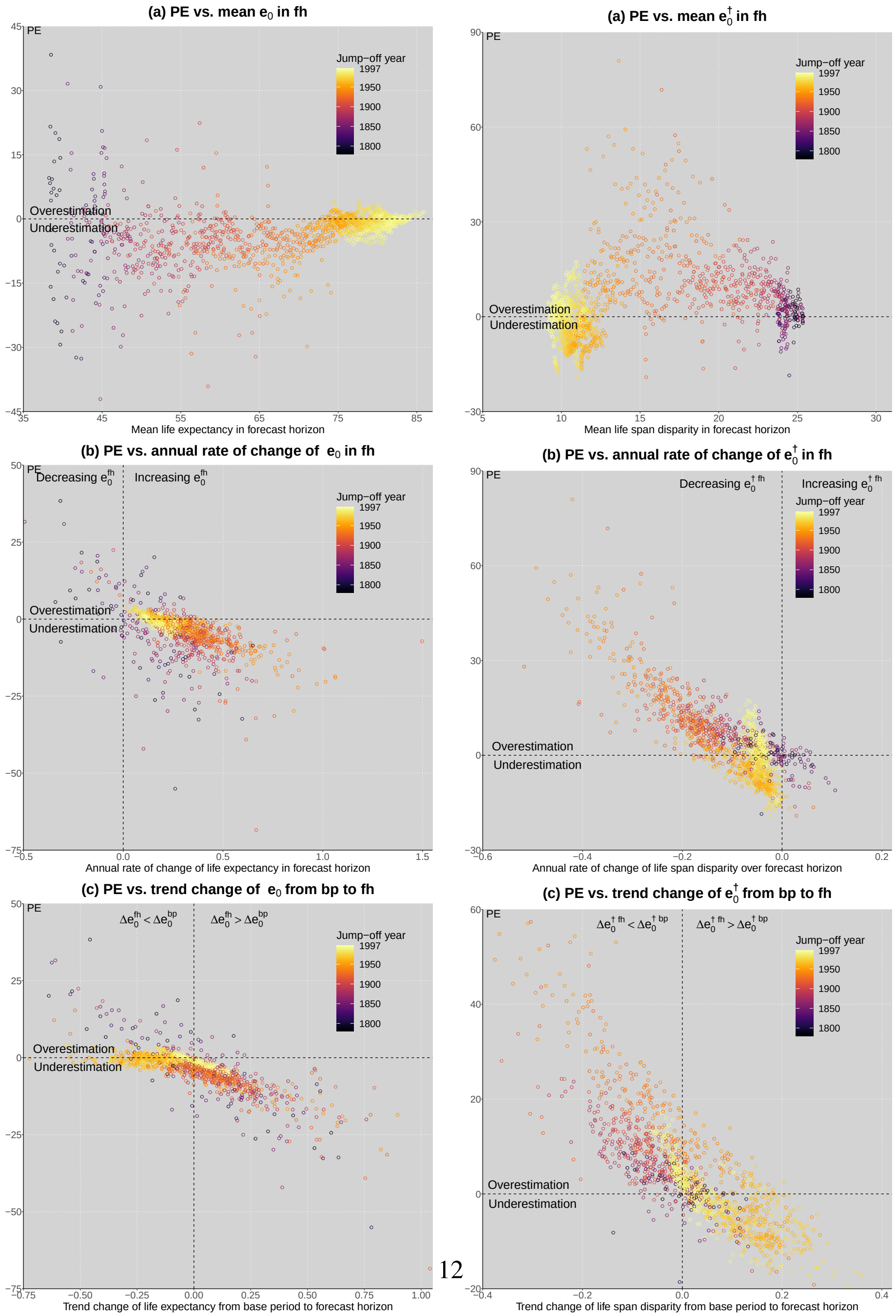
We do not observe an effect of the rate of change on the PE. The annual rate of change of e_0 in the forecast horizon (Figure 5) is positive for all world regions. For $e_0^\dagger$, the rate of change is negative, besides for the two most recent JOYs for Northern America.

Regarding the trend change from the base period to the forecast horizon, the effect is visible as a diagonal line: the higher the trend change, the more severe is the underestimation. The smaller the trend change, the more severe is the overestimation of the mortality measures. The PE is the smallest for forecasts with a trend change close to zero, which is especially visible for $e_0^\dagger$. The findings for the third analytical setting are in line with the results from the HMD data.

3.3 Final Recommendation

Taking all sensitivity checks into account, we can reassess our recommendation for Lee-Carter forecasts. The suitability of the Lee-Carter method varies with the country or world region of interest. Further, different error measures allow for different implications regarding the forecast performance compared between world regions. Based on this exemplary application for 20-years forecasts with a base period of 30 years, we would recommend using Lee-Carter for mortality forecasts only for high-income, low-mortality countries, e.g. in Europe, Northern America, Asia and Oceania. Further, the mortality measure of interest influences the assessment of the forecast performance. The forecast performance of Lee-Carter forecasts assessed on $e_0^\dagger$ are generally worse than for e_0.

Figure 2: Percentage error (PE) of Lee-Carter forecasts of female mortality with a base period (bp) of 30 years and a forecast horizon (fh) of 20 years by jump-off year (dark purple to yellow) plotted against three measures of life expectancy at birth (e_0, left column) and life span disparity at birth ($e_0^\dagger$, right column): (a) against the mean value in the fh, (b) against the annual rate of change in the fh and (c) against the trend change from bp to fh. Data source: Human Mortality Database (**?**).



4. Discussion

4.1 Summary

How can we objectively decide if a method is suitable to forecast mortality in a country of interest? In recent decades, researchers have developed numerous mortality forecast approaches with increasing levels of methodological complexity. Considering that so many simple and complex methods are available, our study's main goal is to establish validation as a test prior to mortality forecasting to make sure a method is selected that is suitable for the particular forecast setting. In this paper, we have validated if the basic assumption of the Lee-Carter 1992 method holds in a country of interest. Namely, that mortality changes in the forecast horizon will develop in the same way they have had in the base period. Therefore, we have applied the Lee-Carter method to all available mortality data for 24 countries of the Human Mortality Database according to our own validation design. We have assessed forecast accuracy and bias of life expectancy and lifespan disparity at birth over all countries by jump-off year in three analytical settings. First, we have analysed the forecast percentage error in relation to the mean level of mortality in the forecast horizon. Second, we have examined how the percentage error reacts to mortality changes in the forecast horizon. Third, we have shown the relationship between the percentage error and mortality trend changes from base period to forecast horizon.

Based on these validation results, we have found that the Lee-Carter method is indeed suitable for forecasting mortality for many high-income countries in most recent years. However, there have been mortality regimes in the past, during which Lee-Carter's forecasts performance was unsatisfactory. This applies to jump-off years ranging from the late nineteenth century to approximately 1960, when discontinuous mortality developments have been prevalent in many of those countries. Caused by, for example, mortality developments such as the advancement of large survival improvements from younger to increasingly older ages with ongoing time.

Further, we have checked the sensitivity of our validation results by performing additional analyses. We have looked at bias, accuracy and uncertainty of the Lee-Carter forecasts using additional error measures and an additional data source. With data from the UN World Population Prospects (2019 edition) we have been able to explore the forecast performance in other mortality regimes. The results show that mortality forecasts are slightly more inaccurate for men. Further, comparing several error measures has shown that the assessment of the forecast performance is influenced by the chosen measure. Regarding the data from the UN WPP 2019, our analyses have shown that the forecast performance of the Lee-Carter model varies by world region.

Based on the initial results and the sensitivity checks, we have argued that the suitability of Lee-Carter to forecast mortality depends on the country of interest and the mortality measure of interest. Lee-Carter's forecast performance is adequate for high-income, low-mortality countries with a smooth mortality development and for life expectancy at birth as the mortality outcome. In contrast, the Lee-Carter model is not able to produce adequate results for mortality regimes with discontinuous development observed e.g., in countries of Africa and the forecast performance is worse for life span disparity at birth, compared to life expectancy at birth.

4.2 Limitations

Our design is suitable to validate mortality forecast methods that use an extrapolation approach. These methods have the same underlying assumption as the assumption that our recommendation after validation is based on: The past mortality trajectories are assumed to continue in the future. However, researchers developed other approaches to forecast mortality. Booth and Tickle (2008) classify the different approaches into three groups: extrapolation methods, explanation models, and methods based on expert opinions. Forecasts build on expert opinions follow a different approach that is based on subjective expectations, e.g. the method of targeting of life expectancy (Pollard, 1987). In principal, our validation design can be used for this type of models, as well. However, it would require an extensive history of forecasts using the same approach. Further, our validation design is not applicable for models that use an explanation approach. Such models use specific causes of death and their risk factors to forecast mortality. In contrast to the extrapolation approach, mortality is not forecasted directly, but the relationship between risk factors and mortality is modelled and the risk factors are forecast. Therefore, the assumption of continuing mortality development is omitted and assessing the forecast performance using our validation design is unreasonable.

The evaluation of the Lee-Carter model served as an exemplary choice to showcase the application of our validation design. The Lee-Carter model itself has proven problems that have been addressed and solved by multiple extensions (for a review of models see Basellini et al. (2022)). However, the comparison of different models was not in the scope of this research.

4.3 Conclusion

In conclusion, we have shown that validation serves as a meaningful first test to decide whether a method is likely to be appropriate to forecast mortality in a country of interest. Our results for the exemplary application of our validation design have shown that the validation process needs to take the mortality development of the population of interest into account. Further, conclusions drawn from different forecast performance measures and analytical settings emphasize the importance of a comprehensive validation design. However, applying the showcased steps, method validation is an extra effort in the research process. Therefore, researchers might want to balance the effort of method validation against the impact or consequences of the forecast results. E.g., mortality forecasts that will be used for political decisions and forecast many years would certainly benefit from prior method validation.

References

Bardoutsos, A., J. de Beer, and F. Janssen (2018, December). Projecting delay and compression of mortality. *Genus 74*(1), 17.

Basellini, U. and C. G. Camarda (2019, January). Modelling and forecasting adult age-at-death distributions. *Population Studies 73*(1), 119–138.

Basellini, U., C. G. Camarda, and H. Booth (2022, December). Thirty years on: A review of the Lee–Carter method for forecasting mortality. *International Journal of Forecasting*, S0169207022001455.

Bohk-Ewald, C., M. Ebeling, and R. Rau (2017, August). Lifespan Disparity as an Additional Indicator for Evaluating Mortality Forecasts. *Demography 54*(4), 1559–1577.

Bohk-Ewald, C., P. Li, and M. Myrskylä (2018, September). Forecast accuracy hardly improves with method complexity when completing cohort fertility. *Proceedings of the National Academy of Sciences 115*(37), 9187–9192.

Bohk-Ewald, C. and R. Rau (2017, December). Probabilistic mortality forecasting with varying age-specific survival improvements. *Genus 73*(1), 1.

Booth, H., R. J. Hyndman, L. Tickle, and P. de Jong (2006, October). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research 15*, 289–310.

Booth, H. and L. Tickle (2008, September). Mortality Modelling and Forecasting: a Review of Methods. *Annals of Actuarial Science 3*(1-2), 3–43.

Camarda, C. G. (2019, October). Smooth constrained mortality forecasting. *Demographic Research 41*, 1091–1130.

D. Pascariu, M., M. J. Dańko, J. Schöley, and S. Rizzi (2018, September). ungroup: An R package for efficient estimation of smooth distributions from coarsely binned data. *Journal of Open Source Software 3*(29), 937.

de Beer, J., A. Bardoutsos, and F. Janssen (2017, June). Maximum human lifespan may increase to 125 years. *Nature 546*(7660), E16–E17.

Gampe, J. (2010). Human mortality beyond age 110. In H. Maier, J. Gampe, B. Jeune, J.-M. Robine, and J. W. Vaupel (Eds.), *Supercentenarians*, pp. 219–230. Berlin, Heidelberg: Springer Berlin Heidelberg. Series Title: Demographic Research Monographs.

Haberman, S. and A. Renshaw (2012, May). Parametric mortality improvement rate modelling and projecting. *Insurance: Mathematics and Economics 50*(3), 309–333.

Hyndman, R. J., H. Booth, and F. Yasmeen (2013, February). Coherent Mortality Forecasting: The Product-Ratio Method With Functional Time Series Models. *Demography 50*(1), 261–283.

Hyndman, R. J., A. B. Koehler, R. D. Snyder, and S. Grose (2002, July). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting 18*(3), 439–454.

Janssen, F., A. Bardoutsos, S. El Gewily, and J. De Beer (2021, July). Future life expectancy in Europe taking into account the impact of smoking, obesity, and alcohol. *eLife 10*, e66590.

Janssen, F., L. J. G. van Wissen, and A. E. Kunst (2013, August). Including the Smoking Epidemic in Internationally Coherent Mortality Projections. *Demography 50*(4), 1341–1362.

Lee, R. D. and L. R. Carter (1992, September). Modeling and Forecasting U. S. Mortality. *Journal of the American Statistical Association 87*(419), 659.

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

Li, N., R. Lee, and P. Gerland (2013, December). Extending the Lee-Carter Method to Model the Rotation of Age Patterns of Mortality Decline for Long-Term Projections. *Demography 50*(6), 2037–2051.

Poler, R. and J. Mula (2011, November). Forecasting model selection through out-of-sample rolling horizon weighted errors. *Expert Systems with Applications 38*(12), 14778–14785.

Pollard, J. H. (1987). Projection of age-specific mortality rates. *Population Bulletin of the United Nations 21*(22), 55–69.

Preston, S. H., P. Heuveline, and M. Guillot (2001). *Demography: measuring and modeling population processes*. Oxford: Blackwell Publisher Ltd.

Raftery, A. E., J. L. Chunn, P. Gerland, and H. Ševčíková (2013, June). Bayesian Probabilistic Projections of Life Expectancy for All Countries. *Demography 50*(3), 777–801.

Rau, R., E. Soroko, D. Jasilionis, and J. W. Vaupel (2008, December). Continued Reductions in Mortality at Advanced Ages. *Population and Development Review 34*(4), 747–768.

Rizzi, S., J. Gampe, and P. H. C. Eilers (2015, July). Efficient Estimation of Smooth Distributions From Coarsely Grouped Data. *American Journal of Epidemiology 182*(2), 138–147.

Rizzi, S., S. Kjærgaard, M.-P. Bergeron Boucher, C. G. Camarda, R. Lindahl-Jacobsen, and J. W. Vaupel (2021, January). Killing off cohorts: Forecasting mortality of non-extinct cohorts with the penalized composite link model. *International Journal of Forecasting 37*(1), 95–104.

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

Shang, H. L. (2015, September). Statistically tested comparisons of the accuracy of forecasting methods for age-specific and sex-specific mortality and life expectancy.

Population Studies 69(3), 317–335.

Shang, H. L., H. Booth, and R. Hyndman (2011, July). Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. *Demographic Research 25*, 173–214.

Shcherbakov, M. V., A. Brebels, N. L. Shcherbakova, A. P. Tyukov, T. A. Janovsky, and V. A. Kamaev (2013). A Survey of Forecast Error Measures.

Sohl, J. E. and A. Venkatachalam (1995, December). A neural network approach to forecasting model selection. *Information & Management 29*(6), 297–303.

van Raalte, A. A., I. Sasson, and P. Martikainen (2018, November). The case for monitoring life-span inequality. *Science 362*(6418), 1002–1004.

Vaupel, J. W. and V. C. Romo (2003). Decomposing change in life expectancy: A bouquet of formulas in honor of Nathan Keyfitz' 90th Birthday. *40*(2).

Venkatachalam, A. R. and J. E. Sohl (1999, May). An intelligent model selection and forecasting system. *Journal of Forecasting 18*(3), 167–180.

Vogt, T., A. van Raalte, P. Grigoriev, and M. Myrskylä (2017, June). The German East-West Mortality Difference: Two Crossovers Driven by Smoking. *Demography 54*(3), 1051–1071.

Wang, H. and S. H. Preston (2009, January). Forecasting United States mortality using cohort smoking histories. *Proceedings of the National Academy of Sciences 106*(2), 393–398.

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

Ševčíková, H., N. Li, V. Kantorová, P. Gerland, and A. E. Raftery (2016). Age-Specific Mortality and Fertility Rates for Probabilistic Population Projections. In R. Schoen (Ed.), *Dynamic Demographic Analysis*, Volume 39, pp. 285–310. Cham: Springer International Publishing. Series Title: The Springer Series on Demographic Methods and Population Analysis.

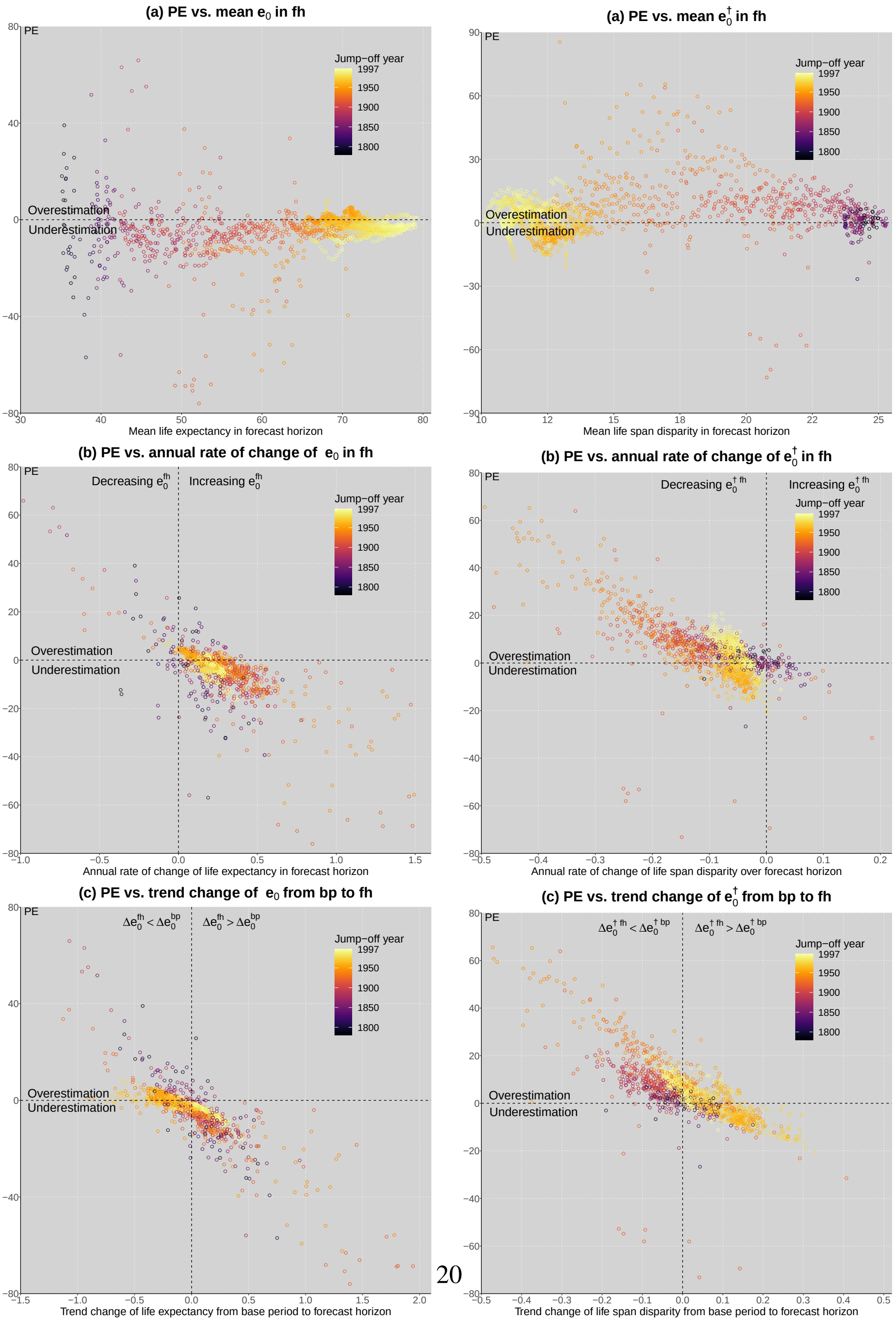
A. Data from the Human Mortality Database (HMD)

Table 2: Chosen countries and data availability in HMD with classification into world regions.

World Region (n=7)	Country (n=24)	Code	Available Years
Northern America	Canada	CAN	1921 – 2016
	USA	USA	1933 – 2017
Australia & New Zealand	Australia	AUS	1921 – 2016
	New Zealand	NZL_NP	1948 – 2013
Japan	Japan	JPN	1947 – 2017
Eastern Europe	Bulgaria	BGR	1947 – 2010
	Czechia	CZE	1950 – 2017
	Hungary	HUN	1950 – 2017
	Slovakia	SVK	1950 – 2017
Northern Europe	Denmark	DNK	1835 – 2016
	Finland	FIN	1878 – 2015
	Ireland	IRL	1950 – 2014
	Norway	NOR	1846 – 2014
	Sweden	SWE	1751 – 2017
	United Kingdom	GBR_NP	1922 – 2016
Southern Europe	Italy	ITA	1872 – 2014
	Portugal	PRT	1940 – 2015
	Spain	ESP	1908 – 2016
Western Europe	Austria	AUT	1947 – 2017
	France	FRATNP	1816 – 2017
	Germany West	DEUTE	1956 – 2017
	Germany East	DEUTW	1956 – 2017
	Netherlands	NLD	1850 – 2016
	Switzerland	CHE	1876 – 2016

Note: Classification follows United Nations Geoscheme (United Nations 1999)

Figure 3: Percentage error (PE) of Lee-Carter forecasts of male mortality with a base period (bp) of 30 years and a forecast horizon (fh) of 20 years by jump-off year (dark purple to yellow) plotted against three measures of life expectancy at birth (e_0, left column) and life span disparity at birth ($e_0^{\dagger}$, right column): (a) against the mean value in the fh, (b) against the annual rate of change in the fh and (c) against the trend change from bp to fh. Data source: Human Mortality Database (**?**).



B. Data from the United Nations World Population Prospects (UN WPP 2019)

Table 3: UN WPP 2019 Data Availability and Classification into World Regions and Subregions.

World Region (n=7)	Subregion (n=22)	Countries (n=209)
Africa	Eastern Africa	Burundi, Comoros, Djibouti, Eritrea, Ethiopia, Kenya, Madagascar, Malawi, Mauritius, Mayotte, Mozambique, Réunion, Rwanda, Seychelles, Somalia, South Sudan, Uganda, United Republic of Tanzania, Zambia, Zimbabwe
	Middle Africa	Angola, Cameroon, Central African Republic, Chad, Congo, Democratic Republic of the Congo, Equatorial Guinea, Gabon, Sao Tome and Principe
	Southern Africa	Botswana, Eswatini, Lesotho, Namibia, South Africa
	Western Africa	Benin, Burkina Faso, Cabo Verde, Côte d'Ivoire, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Mali, Mauritania, Niger, Nigeria, Senegal, Sierra Leone, Togo
	Northern Africa	Algeria, Egypt, Libya, Morocco, Sudan, Tunisia, Western Sahara
Asia	Western Asia	Armenia, Azerbaijan, Bahrain, Cyprus, Georgia, Iraq, Israel, Jordan, Kuwait, Lebanon, Oman, Qatar, Saudi Arabia, State of Palestine, Syrian Arab Republic, Turkey, United Arab Emirates, Yemen
	Central Asia	Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan
	Southern Asia	Afghanistan, Bangladesh, Bhutan, India, Iran (Islamic Republic of), Maldives, Nepal, Pakistan, Sri Lanka
	Eastern Asia	China, China, Hong Kong SAR, China, Macao SAR, China, Taiwan Province of China, Dem. People's Republic of Korea, Japan, Mongolia, Republic of Korea
	South-Eastern Asia	Brunei Darussalam, Cambodia, Indonesia, Lao People's Democratic Republic, Malaysia, Myanmar, Philippines, Singapore, Thailand, Timor-Leste, Viet Nam
Latin America and the Caribbean	Caribbean	Antigua and Barbuda, Aruba, Bahamas, Barbados, Cuba, Curaçao, Dominican Republic, Grenada, Guadeloupe, Haiti, Jamaica, Martinique, Puerto Rico, Saint Lucia, Saint Vincent and the Grenadines, Trinidad and Tobago, United States Virgin Islands
	Central America	Belize, Costa Rica, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama
	South America	Argentina, Bolivia (Plurinational State of), Brazil, Chile, Colombia, Ecuador, French Guiana, Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela (Bolivarian Republic of)
Oceania	Australia/New Zealand	Australia, New Zealand
	Melanesia	Fiji, New Caledonia, Papua New Guinea, Solomon Islands, Vanuatu
	Micronesia	Guam, Kiribati, Micronesia (Fed. States of)
	Polynesia	French Polynesia, Samoa, Tonga
Europe	Eastern Europe	Belarus, Bulgaria, Czechia, Hungary, Poland, Republic of Moldova, Romania, Russian Federation, Slovakia, Ukraine
	Northern Europe	Channel Islands, Denmark, Estonia, Finland, Iceland, Ireland, Latvia, Lithuania, Norway, Sweden, United Kingdom
	Southern Europe	Albania, Bosnia and Herzegovina, Croatia, Greece, Italy, Malta, Montenegro, North Macedonia, Portugal, Serbia, Slovenia, Spain
	Western Europe	Austria, Belgium, France, Germany, Luxembourg, Netherlands, Switzerland
Northern America	Northern America	Canada, United States of America
World		

Preparing and Un-Grouping Data from the UN World Population Prospects It is necessary to prepare and un-group the data from the United Nations World Population Prospects 2019 (UN WPP 2019) to use them for mortality forecast validation (**?**). The original data for population counts is available annually from 1950 to 2020 by five-year age groups and sex. However, death counts by sex are published in period groups of five years from 1950-1954 to 2015-2019 by five-year age groups and sex. Further, there are differences in the highest age groups: the last open age interval starts at age 95 for death counts and for population counts at age 100. Therefore, we need to alter the data sets to, first, achieve the same data format for death counts and population counts and second, to un-group period groups and age groups. The preparation steps 1, 2, 3, 4 and 9 are necessary for all available countries and world regions. However, the preparation steps 5, 6, 7, 8 and 10 are applicable to exceptional cases. For an overview of these exceptions, see Table 4.

To obtain annual age-specific mortality rates by single ages and sex, the following steps were applied using the statistical software R (version 4.0.2), separately for females and males:

1. Contrary to the deaths counts, the data on population exposures has information for the year 2020. Therefore, we excluded the population exposure data for the year 2020.

2. To achieve the same data format of the annual population exposures as for the death counts, first, we grouped the population exposures into five-year periods starting with 1950-1954 and ending with 2015-2019. Second, we summarized the age groups 95-99 and 100+ into one open age group 95+.

3. Using the number of deaths $d(x, n)$ and number of person-years lived $L(x, n)$ from life table data from the UN WPP 2019, we calculated the proportion of deaths and person-years lived in the age groups 0 and 1-4. Then, we multiplicated these proportions with the first age group 0-4 of death counts $D(x, n)$ and population exposures $Pop(x, n)$ to obtain values for the age groups 0 and 1-4. This was done to account for the larger variability in this age group due to infant mortality. The following formula is exemplary for this procedure and shows the calculation of the death counts in age 0 $D(0, 1)$:

$$D(0,1) = D(0,5) * \frac{d(0,1)}{d(0,1) + d(1,4)} \tag{6}$$

4. For death counts and population exposures, we un-grouped the five-year period groups into single years using a cubic smoothing spline with the *smooth.spline* function from the R Stats package (R Core Team 2020).

5. For countries with small populations, we inflated the data which allowed us to obtain results for some of the smallest countries, e.g. Kiribati. We multiplicated the absolute case numbers of death counts and population exposures with 100,000 if, at any point in time, the population in age-group 75-79 was smaller than 100 (see Table 4).

Table 4: UN Data Preparation Steps and their Application

Step	No. of Countries / Subregions	Ages (n=5)	Years
5	184 / 21	-	-
6, D_x	28 / 12	5-30, 85-95	1950-2019
6, Pop$_x$	32 / 13	80-95	1950-2019, 2005, 2008-2012
7	40 / 14	90, 95	1950-1987, 1992-1999, 2003-2007, 2013, 2019
8	143 / 19	40-95	1950-2019
10	5 / 5	90, 95	1950-2019

Note: Full table with information on each individual country available on request.

6. In few cases (see Table 4), the *smooth.spline* procedure produced negative and 0 values for death counts and population exposures. We replaced these implausible values with the last positive non-zero value from age-groups before.

7. Further, *smooth.spline* produced values for death counts and population exposures that resulted in implausible age-specific mortality rates ($m_x = \frac{D_x}{Pop_x}$) >= 1 in the highest age groups. If m_x in age groups 90-94 and 95+ were >= 1, we replaced the corresponding death counts with a value that corresponds to a value of probability of death $a_x = 0.5$ (Gampe, 2010).

8. We assume that mortality after reaching the adult ages increases. However, *smooth.spline* produced cases, for which the age-specific mortality rates m_x in the age groups 40-44 to 95+ are decreasing. We replaced the corresponding death counts with a value so that m_x in this age group equals m_x in the age group before.

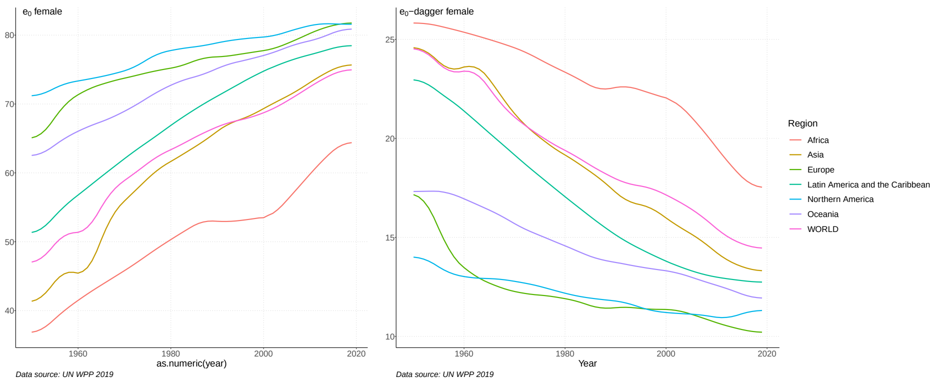
9. We un-grouped the five-year age groups of death counts and population exposures and simultaneously obtained age-specific mortality rates using a Penalized Composite Link Model (Rizzi et al., 2015). We used the *pclm* function from the R package *ungroup* (D. Pascariu et al., 2018) with an optimization interval for the smoothing parameter lambda of [0, 0.0001] and with the population exposures as offset. For a detailed description of the process of choosing the most suitable value for the highest single age, see the following section "Selecting suitable parameters for pclm".

10. In some cases, *pclm* produced decreasing age-specific mortality rates m_x in the higher ages. Under the assumption that mortality is not decreasing after reaching the adult age, we replaced the decreasing m_x values (starting from age 40) with the last positive value from the age before.

Selecting suitable parameters for *pclm* When applying the the *pclm* function from the R package *ungroup* (D. Pascariu et al., 2018), one must define the parameter *nlast*, the length of the last age interval of the un-grouped data. Let Ω be the highest age that determines the length of this last age interval. Depending on Ω, the results of the un-grouping differ substantially. Therefore, we selected four different ages (100, 110, 125, and 130) and applied the *pclm* function to data for the world regions (see Table 3 for list of world regions).

Our decision for the most suitable Ω to un-group the mortality data of the world regions and each of their countries is based on three different criteria. First, we face-validated the age-specific mortality rates resulting from *pclm* using the different value from Ω. Second, we compared the step-wise age-specific mortality rates for five-year age groups calculated from the input data for *pclm* with the age-specific mortality rates resulting from *pclm* to check whether the same mortality patterns can be found. Third, we calculated the difference in life expectancy at birth resulting from the age-specific mortality rates for five-year age groups and for single ages resulting from *pclm*. The selection of Ω for data on individual countries was based on the results for the respective world regions.

Figure 4: Female mortality development by world region from 1751 to 2019. Data source: UN WPP 2019.



(a) Life Expectancy at Birth **(b)** Life Span Disparity at Birth

Figure 5: Percentage error (PE) of Lee-Carter forecasts of female mortality with a base period (bp) of 30 years and a forecast horizon (fh) of 20 years by jump-off year (dark purple to yellow) plotted against three measures of life expectancy at birth (e_0, left column) and life span disparity at birth ($e_0^{\dagger}$, right column): (a) against the mean value in the fh, (b) against the annual rate of change in the fh and (c) against the trend change from bp to fh. Data source: UN WPP 2019.

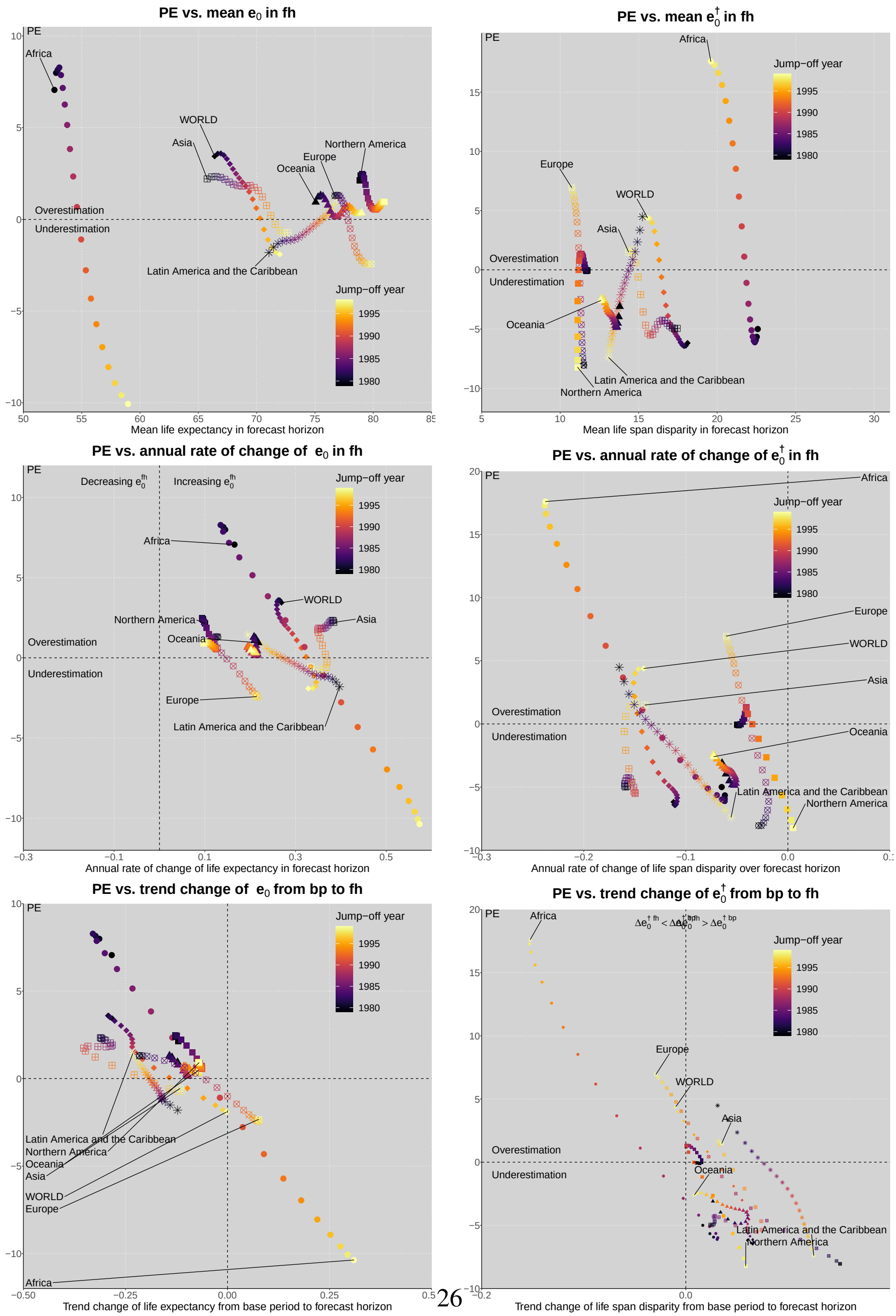


Figure 6: Percentage error (PE) of Lee-Carter forecasts of male mortality with a base period (bp) of 30 years and a forecast horizon (fh) of 20 years by jump-off year (dark purple to yellow) plotted against three measures of life expectancy at birth (e_0, left column) and life span disparity at birth ($e_0^\dagger$, right column): (a) against the mean value in the fh, (b) against the annual rate of change in the fh and (c) against the trend change from bp to fh. Data source: UN WPP 2019.