Comparative Population Studies
Vol. 50 (2025): 95-128 (Date of release: 23.06.2025)

A Practitioner-Oriented Evaluation of Mortality Forecasting Methods: The Case of Australia*

Irina Grossman, Tom Wilson

Abstract: Practitioners seeking a suitable mortality model for forecasting population by age and sex are presented with many possible choices from the large and growing academic literature on mortality forecasting. Despite this abundance, there is relatively little practical guidance on selecting the most appropriate models for their needs. This study evaluates the accuracy of mortality forecasting methods and provides guidance on model selection. The evaluation includes eight methods from the `StMoMo` and `demography` R packages, and a benchmark extrapolative method based on the *Ediev* (2008) model. We also consider the accuracy of simple combinations of individual methods. We evaluate models by preparing mortality 'forecasts' for Australia for past periods using data obtained from the Human Mortality Database. For each method, we created five sets of 30-year retrospective forecasts and evaluated the accuracy of the forecast mortality rates, life expectancies at birth, and life expectancy at age 65. We also evaluated the accuracy of mortality forecasts in terms of projected total deaths calculated using a pseudo-projection method. The Age-Period-Cohort model from the `StMoMo` R package, based on the *Cairns et al.* (2009) implementation, was the standout performer in our evaluation, followed by the benchmark extrapolative method. This study presents a comprehensive evaluation of mortality forecasting methods using a variety of metrics, including a new way to evaluate mortality forecasts using a pseudo-projection method. We hope that this evaluation proves useful for practitioners looking to select a mortality forecasting method.

Keywords: Forecasting · Mortality forecasting methods · Life expectancy · Australia · Pseudo-projection method

* This article has an Online Appendix with supplementary material URL: https://www.comparativepopulationstudies.de/index.php/CPoS/article/view/635/435.

URL: www.comparativepopulationstudies.de
DOI: https://doi.org/10.12765/CPoS-2025-07
URN: urn:nbn:de:bib-cpos-2025-07en

1 Introduction

Mortality forecasts are an essential input to any cohort-component forecast of a population. Mortality is nearly always the main demographic process affecting populations in the older age groups, and it is therefore particularly important that forecasts of older populations are as accurate as possible for applications including those related to pensions, health care, and aged care. An extensive and impressive body of academic literature on mortality forecasting methods has been developed over many decades (*Basellini et al.* 2023; *Bengtsson/Keilman* 2019). For practitioners, however, the preparation of high quality mortality forecasts remains challenging. Whilst an abundance of mortality forecasting models is described in the academic literature, there is limited guidance for practitioners on which models and tools are most appropriate for particular populations and which generate the most plausible and accurate forecasts. One of the most commonly used mortality forecasting models is the Lee-Carter method (*Lee/Carter* 1992), and there is a large body of work developing extensions to this model (*Basellini et al.* 2023). Several retrospective evaluations have shown that these Lee-Carter extensions perform better than the original model when forecasting death rates (*Bergeron-Boucher/Kjærgaard* 2022; *Booth et al.* 2006). Other evaluations have compared the accuracy of the Lee-Carter method with other methods, such as the Age-Period-Cohort (APC) model (*Osmond* 1985), the Cairns-Blake-Dowd (CBD) model (*Cairns et al.* 2006), the Renshaw-Haberman model (*Renshaw/Haberman* 2006), the coherent functional demographic model (CFDM: *Hyndman et al.* 2013), and the Plat model (*Plat* 2009). *Devi Fokeer* and *Narsoo* (2022) evaluated mortality forecast models for male mortality data from five countries and found that the Renshaw-Haberman, Lee-Carter and APC models performed best for 0-19-year-olds while the Plat model performed well for other age groups.

Shang et al. (2011) evaluated ten variants and extensions of the Lee-Carter method using one-step ahead forecast errors on data from 14 countries. They found that the weighted Hyndman-Ullah method (*Hyndman/Ullah* 2007; *Shang et al.* 2011) produced the most accurate point forecasts of male and female mortality rates and of female life expectancy at birth, whereas the Lee-Miller method *(Lee/Miller* 2001) performed best for male life expectancy at birth. The weighted Hyndman-Ullah method performed best on the majority of datasets for female age-specific death rates (ASDRs) and life expectancy, while the Lee-Miller method excelled for the majority of datasets for male life expectancy. *Shang et al.* (2011) suggested the strong performance of the weighted Hyndman-Ullah method was linked to its emphasis on recent mortality trends in its forecasts. *Shang* (2015) investigated model performance for forecasting male and female mortality rates and life expectancies for males and females in 20 countries and found that the CFDM model performed well for males. Similarly, *Li* (2022) evaluated the accuracy of traditional statistical models, machine learning models, and combinations thereof (ensemble models) for forecasting death rates and life expectancies of older US males. They found that a stacked neural network approach to combining a diverse group of base models produced the best forecast.

Dowd et al. (2010) evaluated the accuracy and robustness of mortality forecasting methods using data for males aged 60-89 years in England and Wales. They considered several metrics, including how well forecasts converged to actual mortality rates as forecast horizons decreased, forecast accuracy across a range of forecast horizons, and mortality probability density forecasts. Model performance varied across metrics, with five models performing similarly (Lee-Carter, APC, CBD, and two variants of the CBD model). However, the Renshaw–Haberman model showed limited robustness in their evaluation. *Terblanche* (2015) conducted an evaluation of mortality forecasting methods for Australians aged 50-100 years of age and found the top performer to be the Ediev method (*Ediev* 2008). In the *Booth et al.* (2006) evaluation of the Lee-Carter method and its extensions for mortality forecasting, the authors found that the original Lee-Carter method did not perform as well as its variants in terms of mean error in forecasted logged death rates. However, this difference did not translate to differences in terms of forecast life expectancy. In summary, the literature on mortality forecast evaluations demonstrates how results often vary by dataset, jump-off year, and evaluation metric.

As discussed by *Dowd et al.* (2010), mortality models can be used to produce a range of variables, including age-specific-mortality rates, life expectancy at birth, and survival rates. Practitioners typically value criteria such as forecast accuracy, ease of implementation, accessibility of tools, computational efficiency, and the ability to generate forecasts across different horizons and demographic measures. Ease of use is a significant criterion for practitioners, and a key component of this is the availability of established, validated, and accessible tools. By focusing on models implemented in popular R packages such as `StMoMo` and `demography`, this study ensures that practitioners can readily adopt the methods evaluated here. Such tools are not only user-friendly, but also widely supported within the demographic forecasting community, enhancing their practical utility. Most evaluations focus on the accuracy of forecasts in terms of mortality rates or life expectancies. A key use case for mortality forecasts for demographers is in the production of accurate forecasts of population and deaths. However, relatively little work has been conducted to evaluate the accuracy of mortality forecasting models in terms of numbers of deaths, particularly for longer-term forecasts. A challenge in forecasting the number of deaths is that future exposures are unknown, and a robust population projection framework is often required alongside the mortality forecasting method. Exposures are relatively accurate for past periods, and retrospective evaluations can be conducted to provide valuable insights into potential errors in forecasted deaths.

Recent years have seen substantial developments in mortality forecasting methods and in the tools available to practitioners to create forecasts. However, model selection guidance for practitioners producing and using mortality forecasts in real-world settings remains limited. The aim of this paper is to evaluate the accuracy of the methods available in two of the most popular R packages used for mortality forecasts, `StMoMo` and `demography`, and simple combinations thereof, for forecasts of Australian national mortality data. By evaluating methods that are readily accessible and widely used through these packages, we aim to provide practical guidance for practitioners seeking effective and implementable

mortality forecasts. We use an Extrapolative Model, based on *Ediev's* (2008) work, as a benchmark. A comparison of forecast accuracy for age-specific mortality rates, life expectancies at birth, and life expectancies at age 65 are presented. We also develop and evaluate a new method to evaluate the accuracy of mortality forecasts in terms of total deaths generated through a pseudo-projection method (described later). In summary, the key research aims of this paper are to:

1) evaluate mortality forecast models for Australia for up to 30 years out,

2) develop and evaluate a pseudo-projection method for evaluating forecasts in terms of projected numbers of deaths, and

3) provide guidance to practitioners.

We generated retrospective mortality forecasts for Australia over 30-year periods. The data, mortality models, and metrics are detailed in Section 2, with code and data available at https://github.com/irigrossman/Mortality_Projection_Methods. Section 3 presents the results of the evaluation, which are then discussed in Section 4. Concluding remarks are made in Section 5.

2 Methods

2.1 Data

Age-specific death rates (ASDRs) and age-specific populations-at-risk (Exposures) data for males and females in Australia from 1921–2020 for ages 0–110+ were downloaded from the Human Mortality Database (*HMD* 2023) using the `hmd.mx` function from the R `demography` package (*Hyndman* 2023). As the data at the highest ages was very sparse, we aggregated data for ages 105 years and above into 105+ groups for males and females separately. Even with this aggregation, rates at the oldest ages were atypical due to small populations. Atypical rates included 0s, rates greater than 1, infinite values, and undefined values (caused by division by 0). We employed a simple data cleaning procedure where values greater than 1 were set to 1. 0s, infinite, missing, and undefined values were replaced with the average of the surrounding values that were real numbers (up to 1 year before and after and up to 1 year older and younger – the smoothing window is smaller for edge cases). Whilst there are more sophisticated methods for anomalous data, such as fitting models within a Poisson framework, we selected this simple model to allow for consistency between models – as the model implementation functions handle these values differently (see Online Appendix A), and we wanted to ensure comparability across all models. ASDRs may also be greater than 1; however, by capping at 1, we aimed to decrease potential errors when encountering extreme values, and to support interpretability across age groups. The cleaned data was used to fit the models and evaluate the forecasts.

Input data used for the pseudo-projections (described below) consisted of forecast ASDRs from each of the evaluated mortality models, actual births, and actual net overseas migration. Births were obtained from ABS Historical Population Statistics (*ABS* 2019) for most years and the ABS online Data Explorer tool, https://explore.data.abs.gov.au/, for recent years. Age-sex-specific net overseas migration was calculated as residual cohort population change after accounting for deaths. Population estimates by sex and single years of age for 1971 onwards were sourced from the ABS publication 'National State and Territory Population' (*ABS* 2023), while earlier populations were obtained from the Australian Demographic Databank (*Smith* 2009).

2.2 Packages and forecasting models: An overview

We evaluated methods available in the `demography` (*Hyndman* 2023) and `StMoMo` (*Villegas et al.* 2018) software packages, two of the most popular R packages for mortality forecasts. `Demography` includes a comprehensive set of functions to support data handling, demographic forecasting (including for fertility, migration, and mortality), forecast evaluation, visualisations, and life table functions (*Hyndman* 2023). `StMoMo` is specifically focused on mortality forecasting and provides functions to support the implementation of a wide range of stochastic mortality forecast models. Both packages have significant flexibility in terms of how models are configured. Where possible, we use the default values when using the methods. Additionally, we created custom code for the Extrapolative Model (EM) which we used as a benchmark. EM is described in greater depth in the Forecast Methods section below.

We evaluated nine individual mortality forecast models: the Lee-Carter model implemented in the `demography` package (LC_D, with '_D' indicating the package), a modified version of it with automated base period selection (LC_Dc, where the 'c' indicates this modification), the Booth-Maindonald-Smith variant of the Lee-Carter model (BMS), the Coherent Functional Demographic Model (CFDM), the Lee-Carter model implemented in the `StMoMo` package (LC_S, with '_S' indicating the package), the Age-Period-Cohort model (APC), the Cairns-Blake-Dowd model (CBD), the Plat model, and EM. We also considered three ensemble models. Each of these is described further in the following. Several methods available through the `StMoMo` Package were excluded from our evaluation, including the *Renshaw* and *Haberman* (2006) method due to known issues with its robustness (*Cairns et al.* 2009). The M6, M7, and M8 forecast methods were also excluded (*Cairns et al.* 2009) because they are better suited for higher ages rather than for the full age range (*Plat* 2009) and we wanted to avoid an over-representation of models that are not suited for the full age range in our analysis.

For each of the models, five sets of 30-year forecasts of ASDRs were created with the following jump-off years: 1950, 1960, 1970, 1980, and 1990. Forecast accuracy was evaluated at 1, 10, 20, and 30 years. We provided the full dataset, from 1921 up to each respective jump-off year, to all models. Several models – including BMS, LC_Dc, and EM – automatically select the optimal fitting period based on data

characteristics, resulting in fitting periods that may vary across the five forecasts. Including these models allows us to assess whether methods with adaptive fitting periods provide more accurate forecasts compared to those using a fixed fitting period.

2.3 The Lee-Carter Model and its variants

The Lee-Carter method and its variants are amongst the most popular models for producing forecasts of ASDRs. The original model by *Lee* and *Carter* (1992) involves modelling the log ASDR *m* as a function of age *x* and time *t* (equation 1):

$$\ln[m_{x,t}] = a_x + b_x k_t + \varepsilon_{x,t} \quad (1)$$

The right-hand side of the equation features an age-specific constant a_x, which models the age-specific general pattern of mortality, k_t, which models the main time trend of mortality, and an age-specific constant b_x which modifies k_t, thereby capturing whether the mortality rate at age x will be greater or less than the main trend (*Booth et al.* 2002). The error term relates to time-varying age-specific factors which are not otherwise captured by the model. A key benefit of the model is that it requires relatively few assumptions. However, it has several disadvantages. First, the Lee-Carter model does not consider cohort effects and may be a poor fit for datasets that display them (*Plat* 2009). Secondly, it considers ASDRs to be perfectly correlated as the time index variable k_t is shared by all the age groups; this only allows for uniform shifts in mortality over time to be modelled across all ages.

We implemented several versions of the Lee-Carter model and its variants using both the `Demography` and `StMoMo` packages. The `lca` function was used to fit the Lee-Carter model to the base data using the `demography` package, with `max.age` set to 105. We examined two variants of this model: LC_D, which used default settings, and LC_Dc, where `chooseperiod` was set to `TRUE` to allow for automated fitting period selection using Bai's method, which is the default for the `lca` function. By default, the `lca` function sets the `adjust` parameter to `dt`, which scales the time component k_t to match the total number of deaths. The `lca` function estimates model parameters using singular value decomposition (SVD) of the log mortality rates (*Booth et al.* 2003), consistent with the original Lee-Carter approach (*Lee/Carter* 1992).

We also implemented the Lee-Carter model using the `StMoMo` package (LC_S) using the `LC` function with preset defaults. We then used the `fit` function to fit the model to the base data. In contrast to the SVD approach, the default `StMoMo` implementation employs Maximum Likelihood Estimation assuming that deaths follow a Poisson distribution. Other distribution variants can also be selected by the user (*Villegas et al.* 2025). The `forecast` function was used to generate forecasts for both the `demography` and `StMoMo` packages (*Hyndman et al.* 2023; *Hyndman/Khandakar* 2008).

We also considered the Booth-Maindonald-Smith (BMS) methodology, an extension of the original Lee-Carter model proposed by *Booth et al.* (2002). The BMS

model improves the fit of the base model by identifying the best-fitting period and adjusting k_t to fit the age-specific distribution of deaths. We used the `bms` function in the `Demography` package with `max.age` set to 105, keeping other parameters at their default values. The `bms` function is a wrapper for the `lca` function, with the `adjust` parameter set to `dxt` to fit k_t to the age-specific distribution of deaths. The `bms` function will set the `chooseperiod` parameter to be `TRUE` and the `breakmethod` parameter – which determines the method used to determine the optimal fitting period – to `bms`, which is based on the method by *Booth et al.* (2002). While both LC_Dc and BMS involve automated fitting period selection, they differ in how the optimal periods are selected (Bai's method is the default of *lca* and the bms method is the default for *bms*). Additionally, `bms` adjusts k_t to match the age distribution of deaths, whilst, by default, the LC_D and LC_Dc methods will set the `adjust` parameter to `dt` and fit k_t to total deaths. Thus, whilst both LC_Dc and BMS allow for automated base period selection, there are differences in how they are implemented, which leads to different results.

2.4 Coherent Functional Demographic Model

The CFDM by *Hyndman et al.* (2013) creates consistent forecasts for subpopulations (such as males and females), reducing the risk that forecasts for different subgroups diverge from each other. The method involves the calculation of the product and ratio of the smoothed ASDRs of the subpopulations. For example, if the subpopulations of interest are males and females, as in this study, the product is the geometric mean of the smoothed male and female death rates (or the square root of the smoothed male death rates multiplied by the smoothed female death rates), whilst the ratio is the square root of the ratio of the smoothed death rates.

The product and ratio are modelled using functional time series models, which decompose the product and ratio into a set of weighted basis functions, with weights given by time varying coefficients. These coefficients are then forecast separately using ARIMA(p,q) models for the product and ARMA(p,q) or ARFIMA(p,d,q) models for the ratio. The forecast coefficients are then multiplied by the basis functions, thereby creating forecast curves for the product and ratios into the future. Each coefficient indicates the contribution of a basis function to the forecast death rate at each timepoint in the forecast horizon. The more basis functions included, the greater the potential complexity of the model. However, as more basis functions are included, there is a danger of the model overfitting the base period data. We implement the CFDM model using the `coherentfdm` function from the `demography` package.

2.5 The Age-Period-Cohort Model

The age-period-cohort (APC) mortality model is widely used for modelling age-specific mortality rates as a function of age, period, and birth cohort and can be represented as:

$$\ln[m_{x,t}] = \alpha_x + k_t + \gamma_{t-x} \tag{2}$$

where α, κ, and γ represent the age, period, and cohort effects on the logged mortality rate. In the `StMoMo` implementation of the method, several constraints are implemented such that the period effect across the time periods is 0 and the average cohort effect is 0 with no obvious linear trend (*Villegas et al.* 2018), based on the APC model implementation in *Cairns et al.* (2009).

2.6 The CBD model

The CBD mortality forecast model (*Cairns et al.* 2006) can be expressed as:

$$\text{logit}[q_{x,t}] = ln\left(\frac{q_{x,t}}{1 - q_{x,t}}\right) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) \tag{3}$$

In equation 3, the logit of the mortality rate is modelled such that:

- $k_t^{(1)}$ captures the period effect, which represents general changes in mortality occurring across ages, and
- $k_t^{(2)}(x - \bar{x})$ captures the age-specific effect, allowing time-dependent changes in mortality rates to be age-specific. Here, x represents the age being modelled at time t, and the mean age of the dataset is given by $\bar{x}$.

The CBD model was designed primarily for modelling the 60+ mortality curve, relying on the assumption that cohort effects are linear for the logit transformed mortality rate. This assumption holds better for older ages than for younger ages. Because of the substantial decline in mortality at older ages in Australia, and given the popularity of the method, we felt that it was appropriate to include it as part of a comprehensive evaluation of available mortality forecasting methods. However, the results need to be interpreted with caution given that it is being applied outside its original design. Additionally, due to this application outside its original design, the CBD method was excluded from the ensemble models described below.

The CBD model typically models ${}_1q_x$, the probability of someone exactly x years of age dying before reaching age $x + 1$. Conversely, the other models evaluated in this paper model m_x, the death rate, which is the ratio of the number of deaths occurring at age x during a year divided by the number of person-years lived during that year. However, q_x and m_x can be converted between each other as needed.

To implement the CBD model, we define the model using the `cbd` function from `StMoMo`. We transform the input data from m_x to q_x to using the `central2initial` function and use this transformed data to fit the CBD model. We then create the forecast using the `forecast` function. To allow comparisons with the other models, we convert the output to m_x. As is common practice (e.g., *Thatcher et al.* 1998), we approximate the death rates based on the probability of dying assuming a constant hazard within the age interval (i.e., exponential distribution):

$$m_x \approx -\ln(1 - q_x) \tag{4}$$

2.7 The Plat model

The Plat model was built on prior mortality forecast models, with the goals of building a relatively simple model, able to forecast mortality across the full age range, include a non-trivial correlation structure, and model the cohort effect (*Plat* 2009). The formula for the Plat model is:

$$\ln[m_{x,t}] = \alpha_x + k_t^{(1)} + k_t^{(2)}(x - \bar{x}) + k_t^{(3)}(\bar{x} - x)^+ + \gamma_{t-x} \tag{5}$$

The α_x term represents the basic shape of the age-specific mortality curve from historical data. $K_t^{(1)}$ models general changes in mortality that affect all ages. $K_t^{(2)}(x - \bar{x})$ allows time dependent changes in mortality rates to vary by age, where $\bar{x}$ is the mean age. $K_t^{(3)}(\bar{x} - x)^+$ models age-specific dynamics in mortality rates for younger ages, where the superscript '+' indicates that the term only applies when x is less than the mean age. The cohort effect is modelled with $\gamma(t - x)$.

To implement the Plat model, we use the `StMoMo` package to first define the Plat model using the `plat` function, with the default parameters. Then we fit it to the base data using the `fit` function and then use the `forecast` function to create the forecasts.

2.8 The Extrapolative Model (EM)

The *Ediev* (2008) mortality forecasting method extrapolates logged death rates and applies a set of adjustments and constraints to produce sensible forecasts. Optimal fitting periods are calculated for each age and sex group and long-term trajectories of mortality decline rates are estimated. Convergence parameters bridge the past and long-term mortality rates. The Ediev method was one of the top performers in *Terblanche's* (2015) evaluation of mortality forecasting methods for Australia's very elderly population. The original Ediev method used absolute deviation to determine the optimal fitting period. *Terblanche* (2015) proposed an Ediev model variant that identifies the optimal fitting period by selecting the base period associated with the lowest R^2 value from linear regression, arguing that deviations are more volatile, complicating the selection of a unique fitting period. In the *Terblanche* (2015) evaluation, the two variants performed similarly, however, the R^2 method performed marginally better in the rankings. The EM method used here is strongly influenced by the R^2 variant of the Ediev method. It also incorporates important constraints that reflect demographic realities, ensuring that male ASDRs are greater than or equal to female ASDRs and that, after the age of 15 years, ASDRs do not decrease with increasing age – the ASDR of each successive age group is greater than or equal to that of the younger age group. This approach aligns with *Ediev's* (2008) method, which applies adjustments to ensure that forecasts are plausible and enhance the model's robustness, particularly for extended forecast horizons. A brief description

of the EM method is provided here, with the code provided in the supplementary materials.

1) *Data Preparation*. Input data was smoothed and ASDRs were extended to age 110. The ASDRs at age 110 were set to 0.8 for females and 0.9 for males, based on a twenty-year average of Australian death rates at the highest ages.

2) *Linear Models Using Optimal Fitting Periods*. Smoothed ASDRs were logged, and linear models fitted to determine the optimal fitting periods for each age and sex group, defined by the lowest R^2 value, with a minimum fitting period of 20 years. The intercepts and slopes of the models fitted to the calculated optimal fitting periods were used to create a set logged ASDR projections without any adjustments. Slope values were heavily smoothed across ages and adjustments were applied so that ASDRs from age 15 years onwards did not decrease with increasing age. These heavily smoothed slope values are referred to as the 'long-run slope values.'

3) S*moothing and Adjustments to Slope Values*. The jump-off year ASDRs were smoothed to avoid a bumpy age profile in the final year of the base period. The long-run slope values were trended into the slope projections over 20 years and then smoothed.

4) *Projections with Adjusted Slope*. First, the logged ASDRs were projected again, this time using the logged and smoothed ASDRs from the jump-off year and the adjusted and smoothed forecast slope values. These adjusted projections were more closely aligned with the long-run trends in mortality in the base data. Taking the exponential of these projections generated a set of forecasted ASDRs which incorporated the slope adjustment. To ensure a smooth transition from the historical (base) data to our forecasts, we applied a jump-off year adjustment to better align the forecasted ASDRs with the historical data. The jump-off year adjusted forecasts were gradually merged with the slope-adjusted forecasted ASDRs over a period of 50 years. This ensured a smooth transition from the adjusted initial ASDRs to the long-term mortality rates, aligning the forecasts with the jump-off data and the expected future trends.

5) *Adjusting forecasted Male ASDRs*. Male and Female ASDRs were forecast separately. Male ASDRs were then adjusted to ensure they were always greater or equal to the corresponding female ASDRs.

6) *Modifying forecast maximum age*. While the forecasts extended to age 110 to incorporate calculated mortality rate determined for this age, forecasts were adjusted to age 105, which is the maximum age for this study due to data inconsistencies beyond this age. This was achieved by using life table calculations to aggregate death rates from ages 105 to 110+ into a 105+ category.

2.9 Ensemble models

The models selected for this study are known to perform variably across different datasets. Ensemble models, which are combinations of individual models, may be more reliable. Whilst ensemble models do not necessarily produce the best forecasts, they tend to produce fewer bad forecasts than individual models, a desirable property for demographic forecasting (*Grossman et al.* 2022). Whilst there

are many ways to combine forecasts, research has shown that the mean is a reliable and simple method (*Grossman et al.* 2022) and was thus chosen for this study. We test three ensemble models:

- D ensemble – an ensemble model created by taking the simple mean of three individual methods from the `demography` package: Lee-Carter, BMS, and CFDM.
- S ensemble – an ensemble model created by taking the simple mean of three individual methods from the `StMoMo` package: Lee-Carter, APC, and Plat.
- U ensemble – an ensemble model created by taking the simple mean of five unique methods evaluated in this study: Lee-Carter (LC_D), CFDM, APC, Plat, and EM.

2.10 Forecast evaluation measures

At each forecast horizon, and for each model, the following metrics were used to evaluate forecast accuracy for ASDRs:

- *The mean absolute error in forecast ASDRs over all ages (MAE^d).* The MAE^d for year *y* is given as the average of the absolute difference between the actual ASDR (*m*) and the forecast ASDR for year *y* across all age groups (*x*), expressed as:

$$MAE_y^d = \frac{1}{n}\sum_{x=1}^{n} |Forecast\ m_{x,y} - Actual\ m_{x,y}| \quad (6)$$

- *The absolute error in forecasting life expectancies at birth and at age 65, e_0 and e_{65} respectively.* These metrics were calculated using the `lifetable` function in the `demography` package to create forecasted life tables and life expectancy forecasts, using the forecast ASDRs. The absolute error in forecasting life expectancy at birth ($AE^{e(0)}$) is given in equation 7, which takes the absolute value of the difference between actual life expectancy at birth for year *y* and forecast life expectancy at birth for year *y*. Similarly, equation 8 represents the calculation of the absolute error in the forecast life expectancy at age 65 ($AE^{e(65)}$). e_0 and e_{65} are both measured in years, thereby giving us an easy-to-understand metric:

$$AE_Y^{e(0)} = |\ Actual\ e_{0,y} - Forecast\ e_{0,y}\ | \quad (7)$$

$$AE_Y^{e(65)} = |\ Actual\ e_{65,y} - Forecast\ e_{65,y}\ | \quad (8)$$

- *Absolute Percentage Error in Projected Total Deaths (APE_Y^{TD}).* To evaluate performance in terms of total deaths, as calculated using the pseudo projection method (described below), we calculate the absolute percentage error in projected total deaths (equation 9). A percentage-based metric was selected because the population can change significantly over time, and it normalises errors relative to population size:

$$APE_y^{TD} = \left| \frac{Actual\ D_y - Forecast\ D_y}{Actual\ D_y} \right| \times 100 \tag{9}$$

For each metric, we calculate TOT^m, which represents the average summed error across the five sets of projections with the different jump-off years, as defined by equation 10. We also present a central measure of forecast error, Av^m (equation 11), which is the average error across the forecast horizon, averaged over the forecasts with the different jump-off years.

$$TOT^m = \frac{1}{k} \sum_{j=1}^{k} \sum_{y=1}^{h} E_{y,j} \tag{10}$$

$$Av^m = \frac{1}{k} \sum_{j=1}^{k} \frac{\sum_{y=1}^{h} E_{y,j}}{h} \tag{11}$$

Here, m denotes the specific error metric being aggregated. k is the number of forecasts considered for each model, set at five for this study to correspond to the five jump-off years, and h is the maximum forecast horizon. As an example, let us consider the process of calculating $\mathrm{TOT}^{e(0)}$ and , the total and central errors for life expectancy at birth (e_0), for two three-year forecasts by the same model, with jump-off years y_1 and y_2. Let us assume errors for years 1 to 3 are 0.5 years, 1.2 years and 1.3 years for the first forecast, and 0.2 years, 1.3 years, and 1.2 years for the second. Thus,

$$TOT^{e(0)} = \frac{1}{2}\big((0.5 + 1.2 + 1.3) + (0.2 + 1.3 + 1.2)\big) = \frac{1}{2}(3 + 2.7) = 2.85\ years$$

whilst

$$Av^{e(0)} = \tfrac{1}{2}\left(\tfrac{0.5+1.2+1.3}{3} + \tfrac{0.2+1.3+1.2}{3}\right) = \tfrac{1}{2}(1 + 0.9) = 0.95\ years.$$

By compiling and averaging errors from forecasts across multiple forecast periods and across the entire forecasting window, TOT^m and Av^m serve as indicators of the model's capacity to generate accurate forecasts over time and under different scenarios.

2.11 Pseudo-Projection Method

A key contribution of this study is the introduction of the pseudo-projection method, which evaluates mortality 'projections' over past periods in terms of their impact on total deaths. The pseudo-projection method uses a standard cohort-component model (*Wilson/Rees* 2021) with forecast deaths, but observed historical data for fertility and migration. This substitution isolates the impact of mortality forecast errors on total deaths, allowing us to evaluate model performance for a key demographic outcome without interference from errors in fertility or migration. The term 'pseudo' reflects this deliberate replacement of actual deaths with forecast deaths, distinguishing the method from a full population projection.

To calculate total deaths for the pseudo-projections, forecast ASDRs are combined with observed births and observed net overseas migration to produce a population projection. The resulting total deaths represent what would have occurred if mortality had been forecast imperfectly, but fertility and migration had remained as observed. This method allows us to assess the implications of mortality forecast errors in terms of total deaths, providing a straightforward and intuitive measure which offers a practical evaluation tool for practitioners, particularly those using mortality forecasts to inform planning and decision-making.

By linking forecast mortality rates to total deaths, the pseudo-projection method highlights that strong performance in traditional metrics does not necessarily translate to accuracy in key demographic outcomes. This evaluation underscores the importance of considering total deaths as an independent metric to assess the practical implications of mortality forecast errors.

3 Results

3.1 Mean absolute error for forecast ASDRs over all ages

Table 1 presents the MAE^d for males across the ages of 0-105, while Table 2 presents equivalent results for females. For males, there were considerable differences between jump-off years. If we focus on the mean measures, the APC and Plat models are the standout performers, particularly for longer forecast horizons. The EM and the StMoMo ensemble also performed well across horizons, while the Lee-Carter models had higher errors overall. For females, little separates the forecast errors for shorter horizons. However, over longer horizons, the performance of the CFDM model was poor, whilst EM, APC and the StMoMo ensemble performed well. The latter three models were also the top performers in terms of the TOT metric. The D ensemble model did not perform particularly well, nor did the `StMoMo` implementation of the Lee-Carter model (LC_S). Figure 1 presents the average MAE^d for different forecast horizons by sex, depicting how mean absolute errors in forecast ASDRs change with time and between males and females.

Tab. 1: Mean absolute forecast error of ASDRs, across ages 0-105 for males

Jump-off year	EM	*demography* LC_D	LC_Dc	C FDM	BMS	*StMoMo* LC_S	APC	CBD	Plat	*ensemble models* D	S	U
Point forecast errors after 1 year												
1950	0.014	0.015	0.024	0.015	0.017	0.017	0.022	0.050	0.026	0.015	0.016	**0.013**
1960	0.014	0.027	0.026	0.010	0.019	0.019	0.015	0.036	0.021	0.019	0.010	**0.008**
1970	0.021	0.017	0.019	**0.016**	0.019	0.018	0.024	0.029	0.021	0.017	0.021	0.019
1980	0.013	0.060	0.012	**0.010**	0.025	0.017	0.015	0.019	0.014	0.028	0.014	0.015
1990	0.013	0.022	0.015	0.015	0.014	0.012	0.014	0.022	0.013	**0.009**	0.013	0.011
mean	0.015	0.028	0.019	**0.013**	0.019	0.016	0.018	0.031	0.019	0.018	0.015	**0.013**
Point forecast errors after 10 years												
1950	0.030	0.054	0.064	0.029	0.058	0.051	0.017	0.022	0.020	0.047	**0.013**	0.020
1960	0.016	0.061	0.053	0.031	0.047	0.046	0.018	0.019	0.024	0.046	0.017	0.020
1970	0.026	0.028	0.028	**0.022**	0.037	0.029	0.030	0.022	0.022	0.029	0.026	0.025
1980	0.013	0.093	**0.011**	0.012	0.044	0.028	0.014	0.014	0.015	0.049	0.017	0.026
1990	0.018	0.031	0.025	0.029	0.024	0.019	0.023	0.020	**0.017**	0.019	0.018	**0.017**
mean	0.021	0.053	0.036	0.025	0.042	0.035	0.020	0.020	0.020	0.038	**0.018**	0.022
Point forecast errors after 20 years												
1950	0.044	0.127	0.132	0.064	0.134	0.113	0.021	0.023	**0.021**	0.108	0.039	0.047
1960	0.022	0.123	0.099	0.054	0.097	0.092	0.021	0.032	**0.016**	0.091	0.036	0.042
1970	0.033	0.045	0.035	0.032	0.059	0.046	0.031	0.049	**0.023**	0.045	0.033	0.033
1980	0.022	0.113	0.035	0.025	0.043	0.023	0.020	0.022	0.020	0.052	**0.017**	0.021
1990	0.011	0.042	0.022	0.025	0.025	0.013	0.020	0.015	0.018	0.017	**0.010**	0.011
mean	0.027	0.090	0.064	0.040	0.072	0.057	0.023	0.028	**0.020**	0.063	0.027	0.031
Point forecast errors after 30 years												
1950	0.056	0.257	0.249	0.117	0.271	0.216	**0.021**	0.078	0.023	0.215	0.080	0.090
1960	0.027	0.199	0.149	0.075	0.156	0.144	0.017	0.069	**0.016**	0.143	0.053	0.062
1970	0.025	0.044	0.023	0.023	0.063	0.044	**0.018**	0.063	0.025	0.043	0.019	0.020
1980	0.016	0.165	0.029	0.017	0.065	0.036	0.016	0.046	0.017	0.080	**0.014**	0.035
1990	0.012	0.057	0.022	0.025	0.028	0.020	0.019	0.036	0.035	0.022	0.014	0.016
mean	0.027	0.144	0.094	0.051	0.117	0.092	**0.018**	0.058	0.023	0.101	0.036	0.045
Cumulative forecast errors after 30 years, averaged over the five forecasts												
TOT[d]	0.674	2.328	1.575	0.968	1.831	1.483	0.565	0.869	**0.555**	1.621	0.700	0.823
Average forecast errors after 30 years, averaged over the five forecasts												
Av[m]	0.022	0.078	0.053	0.032	0.061	0.049	0.019	0.029	**0.019**	0.054	0.023	0.027

Notes: Point forecast errors refer to the difference between the actual and forecast values for specific years. Point errors are reported for 1-, 10-, 20-, and 30-year forecast horizons, for each jump-off year. Mean errors across jump-off years are presented for each of the evaluated forecast horizons. The TOT metric is the average of the cumulative errors over the 30-year forecast horizon, for the five evaluated forecasts for the five jump-off years. The Av metric is equal to the TOT metric divided by the forecast horizon, providing a central measure of error across the forecast horizon.

Abbreviations: EM – extrapolative smoothing model, LC_D – Lee-Carter model from the `demography` package, LC_Dc Lee-Carter model from the `demography` package with automated base period selection, CFDM – coherent functional demographic model, LC_S

– the `StMoMo` implementation of the Lee-Carter model, APC – Age-period-cohort model, CBD – Cairns-Blake-Dowd model, Plat – Plat model, D – demography ensemble, S – StMoMo ensemble, U – unique ensemble.

Source: Authors' calculations based on *HMD* (2023).

Tab. 2: Mean absolute forecast error of ASDRs, across ages 0-105 for females

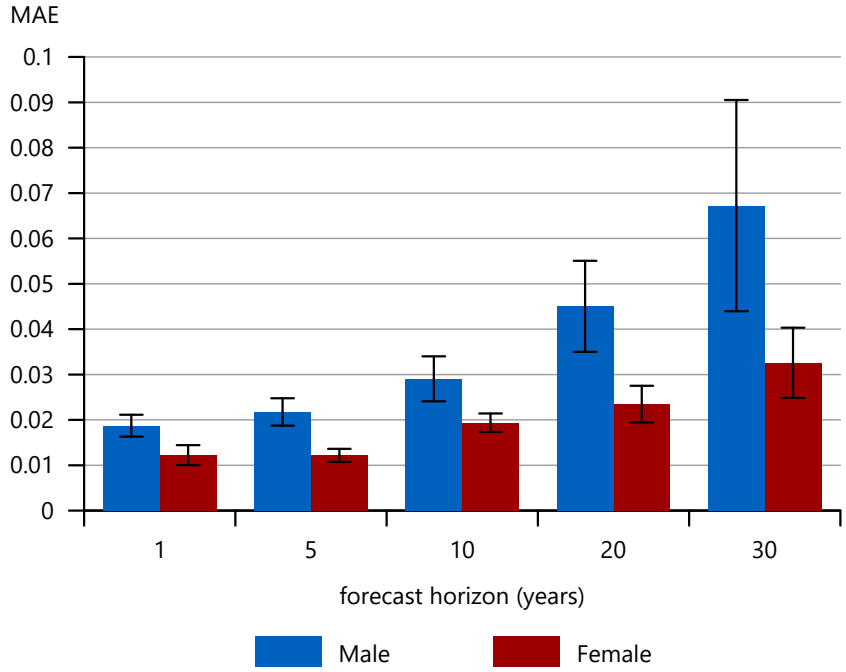
Jump-off		*demography*				*StMoMo*				*ensemble models*		
year	EM	LC_D	LC_Dc	CFDM	BMS	LC_S	APC	CBD	Plat	D	S	U
1-year forecast horizon												
1950	0.020	0.013	**0.012**	0.013	0.013	0.016	0.018	0.053	0.015	0.013	0.015	0.013
1960	0.013	0.012	0.008	0.009	0.011	0.008	0.007	0.033	0.007	0.010	**0.006**	0.007
1970	0.013	0.014	0.014	**0.013**	0.014	0.014	0.016	0.027	**0.013**	0.014	0.014	**0.013**
1980	0.008	0.014	0.008	0.009	**0.007**	0.010	0.009	0.014	0.009	0.009	0.009	0.009
1990	0.008	0.008	0.009	0.009	0.009	0.007	0.006	0.013	**0.005**	0.008	0.006	0.007
mean	0.012	0.012	**0.010**	0.011	0.011	0.011	0.011	0.028	**0.010**	0.011	**0.010**	**0.010**
10-year forecast horizon												
1950	0.031	0.054	0.053	0.072	0.053	0.050	0.023	**0.015**	0.027	0.060	0.033	0.041
1960	0.017	0.016	0.013	0.019	0.014	0.013	0.011	0.021	0.014	0.016	0.011	**0.008**
1970	0.012	0.023	0.011	0.023	0.023	0.018	0.019	0.010	**0.006**	0.023	0.013	0.016
1980	0.008	0.012	0.006	0.010	0.009	0.005	0.006	0.006	0.011	0.006	**0.005**	**0.005**
1990	**0.011**	0.012	0.017	0.020	0.018	0.013	0.017	0.013	0.021	0.016	0.016	0.015
mean	0.016	0.023	0.020	0.029	0.023	0.020	0.015	**0.013**	0.016	0.024	0.016	0.017
20-year forecast horizon												
1950	0.019	0.056	0.054	0.092	0.055	0.050	0.011	0.020	**0.010**	0.068	0.018	0.033
1960	0.008	0.038	0.026	0.054	0.036	0.021	0.009	0.020	**0.008**	0.042	**0.008**	0.019
1970	0.008	0.025	0.014	0.030	0.025	0.018	0.012	0.028	0.012	0.027	**0.007**	0.013
1980	0.016	0.016	0.018	0.018	0.021	0.013	0.019	**0.011**	0.032	0.014	0.020	0.015
1990	0.009	0.012	0.017	0.021	0.019	0.011	0.021	**0.008**	0.029	0.016	0.019	0.016
mean	**0.012**	0.029	0.026	0.043	0.031	0.023	0.014	0.017	0.018	0.033	0.015	0.019
30-year forecast horizon												
1950	0.035	0.100	0.096	0.163	0.098	0.091	**0.012**	0.073	0.012	0.120	0.037	0.064
1960	0.010	0.044	0.028	0.072	0.041	0.020	0.007	0.043	0.014	0.053	**0.003**	0.020
1970	0.013	0.022	0.026	0.031	0.022	0.017	**0.008**	0.044	0.031	0.025	0.013	0.010
1980	0.016	0.017	0.017	0.016	0.021	**0.012**	0.023	0.027	0.039	**0.012**	0.022	0.015
1990	**0.007**	0.013	0.015	0.020	0.017	0.010	0.021	0.026	0.031	0.014	0.018	0.013
mean	0.016	0.039	0.036	0.061	0.040	0.030	**0.014**	0.043	0.026	0.045	0.019	0.024
Cumulative forecast errors after 30 years, averaged over the five forecasts												
TOT[d]	**0.380**	0.758	0.658	1.039	0.768	0.590	0.391	0.644	0.500	0.819	0.410	0.488
Average forecast errors after 30 years, averaged over the five forecasts												
Av[m]	**0.013**	0.025	0.022	0.035	0.026	0.020	0.013	0.021	0.017	0.027	0.014	0.016

Notes: Point forecast errors refer to the difference between the actual and forecast values for specific years. Point errors are reported for 1-, 10-, 20-, and 30-year forecast horizons, for each jump-off year. Mean errors across jump-off years are presented for each of the evaluated forecast horizons. The TOT metric is the average of the cumulative errors over the 30-year

forecast horizon, for the five evaluated forecasts for the five jump-off years. The Av metric is equal to the TOT metric divided by the forecast horizon, providing a central measure of error across the forecast horizon.
Abbreviations: EM – extrapolative smoothing model, LC_D – Lee-Carter model from the `demography` package, LC_Dc Lee-Carter model from the `demography` package with automated base period selection, CFDM – coherent functional demographic model, LC_S – the `StMoMo` implementation of the Lee-Carter model, APC – Age-period-cohort model, CBD – Cairns-Blake-Dowd model, Plat – Plat model, D – demography ensemble, S – StMoMo ensemble, U – unique ensemble.

Source: Authors' calculations based on *HMD* (2023).

Fig. 1: Mean absolute error in forecast ASDRs over all ages, averaged across models, by forecast horizon



Notes: The figure presents mean absolute error in forecast ASDRs over all ages (MAE^{d}), averaged across jump-off years and across models. Error bars represent 95% confidence intervals of the model errors.

Source: Authors' calculations based on *HMD* (2023).

3.2 Error for forecast life expectancy at birth

Tables 3 and 4 present the absolute error in forecasting life expectancy at birth, $AE^{e(0)}$, for males and females, respectively. Actual life expectancies are presented in the lefthand columns of the tables. The $AE^{e(0)}$ is presented for each model, each evaluated jump-off year, and forecast horizon. For males, the APC model is the

Tab. 3: Absolute error for life expectancy at birth forecasts for males

Jump-off year	Actual	*demography* EM	LC_D	LC_Dc	C FDM	BMS	*StMoMo* LC_S	APC	CBD	Plat	*ensemble models* D	S	U
1-year forecast horizon													
1950	66.08	0.60	0.48	0.57	0.85	0.58	0.62	0.74	**0.41**	0.70	0.63	0.69	0.67
1960	68.04	0.25	0.35	0.39	0.28	0.11	0.09	0.30	0.13	0.14	**0.02**	0.12	0.05
1970	68.25	0.58	1.25	1.43	**0.24**	0.35	0.37	0.32	0.87	0.71	0.62	0.47	0.63
1980	71.30	0.26	1.12	**0.02**	0.45	1.16	0.81	0.24	0.02	0.11	0.19	0.24	0.10
1990	74.48	0.77	0.45	0.25	0.67	0.20	0.71	**0.08**	0.23	0.39	0.15	0.40	0.30
	mean	0.49	0.73	0.53	0.50	0.48	0.52	0.34	0.33	0.41	**0.32**	0.38	0.35
10-year forecast horizon													
1950	67.95	0.06	0.06	0.43	0.73	**0.03**	0.10	0.88	0.73	0.55	0.23	0.50	0.40
1960	67.43	1.68	2.00	1.81	1.81	1.55	1.59	2.93	**1.12**	2.18	1.79	2.21	2.11
1970	71.09	2.42	3.26	4.23	1.78	2.38	2.40	**1.53**	3.27	2.66	2.48	2.20	2.34
1980	73.98	1.84	0.91	1.65	2.19	3.10	2.66	**0.55**	2.10	1.79	2.09	1.70	1.48
1990	76.95	1.15	1.14	0.82	1.73	**0.26**	2.16	0.44	2.11	1.89	1.07	1.53	1.29
	mean	1.43	1.48	1.79	1.65	1.46	1.78	**1.27**	1.87	1.81	1.53	1.63	1.52
20-year forecast horizon													
1950	67.43	1.68	1.61	0.87	2.59	1.70	1.79	3.93	**0.28**	2.56	1.96	2.72	2.44
1960	71.09	0.71	0.76	1.14	0.41	1.19	1.13	1.70	2.05	0.60	0.79	**0.10**	0.22
1970	73.98	4.31	5.38	7.09	3.44	4.53	4.55	**2.46**	5.81	5.19	4.47	4.12	4.20
1980	76.95	3.68	3.29	3.54	3.95	5.38	4.85	**1.29**	4.60	4.00	4.23	3.48	3.30
1990	79.80	1.97	3.12	1.76	2.89	**0.77**	4.00	0.86	4.46	3.89	2.34	3.04	2.63
	mean	2.47	2.83	2.88	2.66	2.71	3.26	**2.05**	3.44	3.25	2.76	1.63	1.52
30-year forecast horizon													
1950	71.09	0.92	1.15	2.17	0.02	1.08	0.97	3.00	3.10	0.26	0.76	0.43	**0.01**
1960	73.98	2.51	2.96	3.47	2.13	3.35	3.28	1.31	4.63	3.55	2.83	2.08	2.13
1970	76.95	6.43	7.70	10.03	5.38	6.90	6.91	**3.37**	8.56	8.40	6.68	6.40	6.37
1980	79.80	5.47	5.64	5.38	5.72	7.64	7.02	**1.72**	7.11	6.39	6.38	5.28	5.13
1990	81.62	1.96	4.13	1.83	3.07	0.54	4.87	**0.08**	5.88	5.03	2.73	3.66	3.09
	mean	3.46	4.32	4.57	3.26	3.90	4.61	**1.90**	5.86	4.73	3.87	3.57	3.35
Cumulative forecast errors after 30 years, averaged over the five forecasts													
	$TOT^{e(0)}$	55.0	64.1	68.2	59.4	60.2	72.7	**43.0**	81.4	73.9	60.7	61.4	57.4
Average forecast errors after 30 years, averaged over the five forecasts													
	Av^{m}	1.83	2.14	2.27	1.98	2.01	2.42	**1.43**	2.71	2.46	2.02	2.05	1.91

Notes: Actual life expectancies are presented in the second column from the left. If the jump-off year is 1970 and the forecast horizon is 10 years, the relevant cell will show the estimated life expectancy at birth for 1980. The 12 rightmost columns present the absolute error in years of the life expectancy forecasts for the 12 evaluated models for each of the evaluated jump-off years and forecast horizons. The mean of the absolute errors across jump-off years for each evaluated forecast horizon are also presented. The TOT metric is the average of the cumulative errors over the 30-year forecast horizon, for the five evaluated forecasts and the five jump-off years. The Av metric is equal to the TOT metric divided by the forecast horizon, providing a central measure of error across the forecast horizon.

Abbreviations: EM – extrapolative smoothing model, LC_D – Lee-Carter model from the `demography` package, LC_Dc Lee-Carter model from the `demography` package with automated base period selection, CFDM – coherent functional demographic model, LC_S – the `StMoMo` implementation of the Lee-Carter model, APC – Age-period-cohort model, CBD – Cairns-Blake-Dowd model, Plat – Plat model, D – demography ensemble, S – StMoMo ensemble, U – unique ensemble.

Source: Authors' calculations based on *HMD* (2023).

Tab. 4: Absolute error for life expectancy at birth forecasts for females

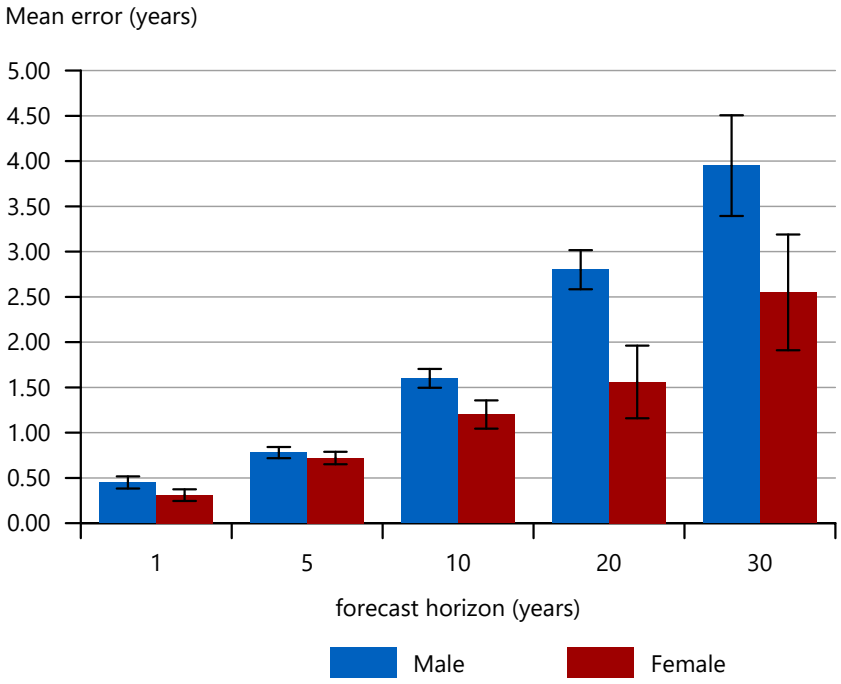
Jump-off		*demography*					*StMoMo*				*ensemble models*		
year	Actual	EM	LC_D	LC_Dc	C FDM	BMS	LC_S	APC	CBD	Plat	D	S	U
						1-year forecast horizon							
1950	71.59	0.23	0.37	0.40	0.17	0.28	0.31	0.18	0.19	0.38	**0.16**	0.29	0.20
1960	74.51	0.41	0.16	**0.10**	0.75	0.56	0.49	0.13	0.53	0.23	0.38	0.29	0.28
1970	74.89	0.42	0.40	0.55	0.71	0.61	0.57	**0.18**	0.82	0.53	0.57	0.43	0.45
1980	78.43	0.25	0.55	0.08	0.86	0.06	0.22	0.17	0.03	**0.01**	0.13	0.02	0.08
1990	80.50	0.60	0.21	0.16	0.51	0.13	0.29	**0.01**	0.14	0.15	0.15	0.15	0.21
	mean	0.38	0.34	0.26	0.60	0.33	0.38	**0.13**	0.34	0.26	0.28	0.24	0.24
						10-year forecast horizon							
1950	74.05	0.92	0.57	0.72	1.78	0.65	0.60	0.13	1.72	**0.08**	1.01	0.27	0.71
1960	74.17	1.94	1.85	1.77	0.41	1.14	1.24	2.48	**0.39**	2.19	1.13	1.95	1.74
1970	78.22	2.28	2.58	2.71	3.29	2.77	2.70	**1.45**	3.74	2.08	2.88	2.10	2.36
1980	80.15	0.70	**0.01**	0.30	2.11	0.13	0.66	0.72	1.21	0.37	0.78	0.09	0.41
1990	82.27	0.49	0.39	0.12	1.26	0.13	0.79	0.51	1.47	**0.03**	0.53	0.14	0.38
	mean	1.27	1.08	1.12	1.77	0.96	1.20	1.06	1.70	0.95	1.27	**0.91**	1.12
						20-year forecast horizon							
1950	74.17	0.18	0.65	0.36	0.79	0.59	0.67	2.34	1.53	1.46	**0.14**	1.45	0.71
1960	78.22	0.13	0.95	1.04	2.79	1.65	1.50	1.21	3.24	**0.06**	1.81	0.18	0.64
1970	80.15	2.72	3.39	3.45	4.48	3.57	3.45	**0.80**	5.36	2.55	3.82	2.38	2.90
1980	82.27	1.53	0.96	0.86	3.49	0.52	1.50	1.32	2.95	**0.03**	1.75	0.26	1.17
1990	84.20	0.50	1.14	**0.20**	1.90	0.27	1.42	1.30	3.02	0.52	0.97	0.37	0.70
	mean	1.01	1.42	1.18	2.69	1.32	1.71	1.39	3.22	**0.93**	1.70	0.93	1.22
						30-year forecast horizon							
1950	78.22	2.85	2.31	2.72	3.92	2.37	2.26	**1.13**	5.47	1.67	2.88	**1.07**	2.09
1960	80.15	**0.29**	1.83	1.91	4.00	2.51	2.32	2.37	4.93	0.61	2.80	0.45	1.17
1970	82.27	3.47	4.53	4.49	5.94	4.70	4.54	**0.06**	7.30	4.50	5.07	3.33	4.00
1980	84.20	2.25	1.84	1.30	4.74	**0.80**	2.26	2.34	4.61	1.48	2.62	0.90	2.09
1990	85.69	0.26	1.57	**0.02**	2.14	0.67	1.72	2.70	4.25	1.67	1.11	0.60	0.92
	mean	1.82	2.42	2.09	4.15	2.21	2.62	1.72	5.31	1.99	2.90	**1.27**	2.05
						Cumulative forecast errors after 30 years, averaged over the five forecasts							
	TOT	31.4	35.2	32.0	67.5	32.0	38.8	33.9	76.8	26.6	42.5	**22.2**	30.3
	Av^m	1.05	1.17	1.07	2.25	1.07	1.29	1.13	2.56	0.89	1.42	**0.74**	1.01

Notes: Actual life expectancies are presented in the second column from the left. If the jump-off year is 1970 and the forecast horizon is 10 years, the relevant cell will show the estimated life expectancy at birth for 1980. The 12 rightmost columns present the absolute error in years

of the life expectancy forecasts for the 12 evaluated models for each of the evaluated jump-off years and forecast horizons. The mean of the absolute errors across jump-off years for each evaluated forecast horizon are also presented. The TOT metric is the average of the cumulative errors over the 30-year forecast horizon, for the five evaluated forecasts and the five jump-off years. The Av metric is equal to the TOT metric divided by the forecast horizon, providing a central measure of error across the forecast horizon.
Abbreviations: EM – extrapolative smoothing model, LC_D – Lee-Carter model from the `demography` package, LC_Dc Lee-Carter model from the `demography` package with automated base period selection, CFDM – coherent functional demographic model, LC_S – the `StMoMo` implementation of the Lee-Carter model, APC – Age-period-cohort model, CBD – Cairns-Blake-Dowd model, Plat – Plat model, D – demography ensemble, S – StMoMo ensemble, U – unique ensemble.

Source: Authors' calculations based on *HMD* (2023).

Fig. 2: Mean absolute error in forecasted life expectancy at birth, averaged across models, by forecast horizon



Notes: This figure visualises the mean absolute error in forecasted life expectancy at birth, averaged across jump-off years and then across models. Error bars represent 95% confidence intervals of the model errors.

Source: Authors' calculations based on *HMD* (2023).

standout performer, with the difference particularly evident for the 30-year forecast horizon. When we consider the TOT metric, EM is the second-best performer and CFDM also performs reasonably well. Conversely, CBD performs poorly. For females,

when we consider the TOT metric, the Plat and EM models were the top performers, whilst the CFDM and CBD models were the worst performers.

The $AE^{e(0)}$ is greater for males than for females, with similar variance in model performance (Fig. 2). Across the models and jump-off years, the average $AE^{e(0)}$ for a 30-year forecast horizon increases from 1.60 years for a 10-year forecast horizon to 3.95 years for males and from 1.20 years for a 10-year forecast horizon to 2.55 years for a 30-year horizon.

3.3 Error for forecasted life expectancy at age 65

Tables 5 and 6 present the absolute error in life expectancy at age 65, $AE^{e(65)}$, for males and females. Actual life expectancies are presented in the lefthand columns, and the $AE^{e(65)}$ is presented for each model, each evaluated jump-off year, and each forecast horizon. When we consider the TOT metric, APC, EM, LC_Dc, and CFDM were the best performing models for males, whilst the APC, EM, Plat, and LC_Dc models were the best performing methods for females. The $AE^{e(65)}$ is greater for males than for females (Fig. 3). The average errors across the models and jump-off years increases from 0.98 years for a 10-year forecast horizon to 3.16 years for a 30-year forecast horizon for males, and from 0.85 years for a 10-year forecast horizon to 2.21 years for a 30-year horizon for females.

3.4 Error for total projected deaths using the pseudo-projection method

The results of our evaluation using the pseudo-projections method are presented in Table 7. The TOT^{m} metric is presented in the bottom row, whilst the absolute percentage errors between the actual and forecast number of total deaths is presented in the relevant cells for each of the models. The APC model was the top performer, particularly for the 20- and 30-year forecast horizons. The EM model was the second top performer according to the TOT^{m} metric. If we consider the mean absolute percentage error in forecast total deaths using the pseudo-projection method across models and jump-off years, we find that it increases from 1.7 percent for a forecast horizon of 1 year, to 8.4 percent for a 10-year horizon, to 12.8 percent for a 20-year horizon, and to 17.7 percent for a 30-year horizon.

3.5 Age-specific absolute errors in forecast ASDRs

The average absolute errors in the forecast ASDRs for males and females after 10, 20 and 30 years are reported in Appendices B1-B6. Here, the average represents the mean value across the five forecasts produced for jump-off years 1950, 1960, 1970, 1980, and 1990. In each table, the top model is bolded for each age group, whilst the worst performing model is shown in red. The APC model does not perform particularly well for the younger age groups and has few top performances for any age group for the 10- and 20-year forecast horizons. However, it performs very well relative to other models for age 50 and above for the 30-year forecast horizon. The EM model performs well across the age groups for the 10-, 20-, and 30-year forecast

Tab. 5: Absolute error for life expectancy at age 65 forecasts for males

Jump-off		*demography*						*StMoMo*			*ensemble models*		
year	Actual	EM	LC_D	LC_Dc	C FDM	BMS	LC_S	APC	CBD	Plat	D	S	U
						1-year forecast horizon							
1950	12.10	0.22	0.20	0.12	0.31	0.20	0.21	**0.04**	1.24	0.10	0.24	0.12	0.17
1960	12.60	0.19	0.27	0.32	0.27	0.26	0.24	0.26	0.84	**0.07**	0.27	0.19	0.21
1970	12.47	0.34	0.17	0.12	**0.06**	0.18	0.19	0.61	0.28	0.38	0.14	0.40	0.32
1980	13.84	0.04	0.78	0.04	0.17	1.10	0.85	0.16	0.43	**0.03**	0.70	0.33	0.23
1990	15.45	0.47	0.73	0.14	0.27	**0.13**	0.83	0.25	0.20	0.21	0.38	0.43	0.39
	mean	0.25	0.43	**0.15**	0.21	0.38	0.46	0.26	0.60	0.16	0.34	0.30	0.27
						10-year forecast horizon							
1950	12.51	0.21	0.23	0.38	0.12	0.23	0.21	**0.02**	0.24	0.19	0.12	0.14	0.11
1960	11.98	0.57	0.34	**0.26**	0.56	0.34	0.38	1.02	0.67	1.15	0.41	0.84	0.72
1970	13.76	1.56	1.46	1.41	**1.10**	1.48	1.49	1.21	1.70	1.14	1.35	1.29	1.30
1980	15.22	0.93	2.03	0.91	1.41	2.39	2.06	**0.64**	1.37	0.97	1.96	1.25	1.22
1990	17.05	0.97	1.96	0.71	1.15	**0.42**	2.00	0.69	1.65	1.75	1.21	1.51	1.33
	mean	0.85	1.21	0.73	0.87	0.97	1.23	**0.72**	1.13	1.04	1.01	1.00	0.94
						20-year forecast horizon							
1950	11.98	0.41	0.26	**0.05**	0.77	0.26	0.30	1.58	0.73	0.85	0.43	0.90	0.76
1960	13.76	1.00	1.45	1.56	0.93	1.44	1.39	0.66	1.87	0.32	1.28	**0.19**	0.52
1970	15.22	2.90	2.94	2.87	2.29	2.97	2.97	**1.63**	3.84	2.22	2.74	2.29	2.41
1980	17.05	2.26	3.72	2.18	2.85	4.13	3.70	**1.33**	3.64	2.67	3.59	2.63	2.61
1990	19.01	1.80	3.52	1.56	2.16	**0.99**	3.49	1.03	3.87	3.59	2.29	2.82	2.51
	mean	1.67	2.38	1.64	1.80	1.96	2.37	**1.25**	2.79	1.93	2.07	1.76	1.76
						30-year forecast horizon							
1950	13.76	1.23	1.55	1.83	0.90	1.55	1.49	1.59	3.37	0.24	1.34	**0.14**	0.54
1960	15.22	2.22	2.93	3.07	2.13	2.92	2.84	0.89	4.00	**0.70**	2.67	1.05	1.55
1970	17.05	4.58	4.79	4.70	3.87	4.82	4.82	**2.35**	6.29	4.36	4.51	3.91	4.04
1980	19.01	3.74	5.54	3.57	4.40	6.00	5.48	**1.77**	6.00	4.66	5.35	4.13	4.13
1990	20.57	2.27	4.67	2.04	2.72	1.23	4.57	**0.65**	5.67	5.21	3.01	3.79	3.33
	mean	2.81	3.90	3.04	2.81	3.30	3.84	**1.45**	5.07	3.03	3.37	2.60	2.72
						Cumulative forecast errors after 30 years, averaged over the five forecasts							
	TOT	37.7	55.6	38.5	39.0	45.7	55.4	**26.6**	65.4	43.0	47.0	39.5	39.0
						Average forecast errors after 30 years, averaged over the five forecasts							
	Av^m	1.3	1.9	1.3	1.3	1.5	1.8	**0.9**	2.2	1.4	1.6	1.3	1.3

Notes: Actual life expectancies are presented in the second column from the left. If the jump-off year is 1970 and the forecast horizon is 10 years, then the relevant cell will show the estimated life expectancy at birth for 1980. The 12 rightmost columns present the absolute error in years of the life expectancy forecasts for the 12 evaluated models for each of the evaluated jump-off years and forecast horizons. The mean of the absolute errors across jump-off years for each evaluated forecast horizon are also presented. TOT is the average cumulative error over the five 30-year forecasts and is presented at the bottom of the table. The Av metric is equal to the TOT metric divided by the forecast horizon, providing a central measure of error across the forecast horizon.

Abbreviations: EM – extrapolative smoothing model, LC_D – Lee-Carter model from the `demography` package, LC_Dc Lee-Carter model from the `demography` package with automated base period selection, CFDM – coherent functional demographic model, LC_S – the `StMoMo` implementation of the Lee-Carter model, APC – Age-period-cohort model, CBD – Cairns-Blake-Dowd model, Plat – Plat model, D – demography ensemble, S – StMoMo ensemble, U – unique ensemble.

Source: Authors' calculations based on *HMD* (2023).

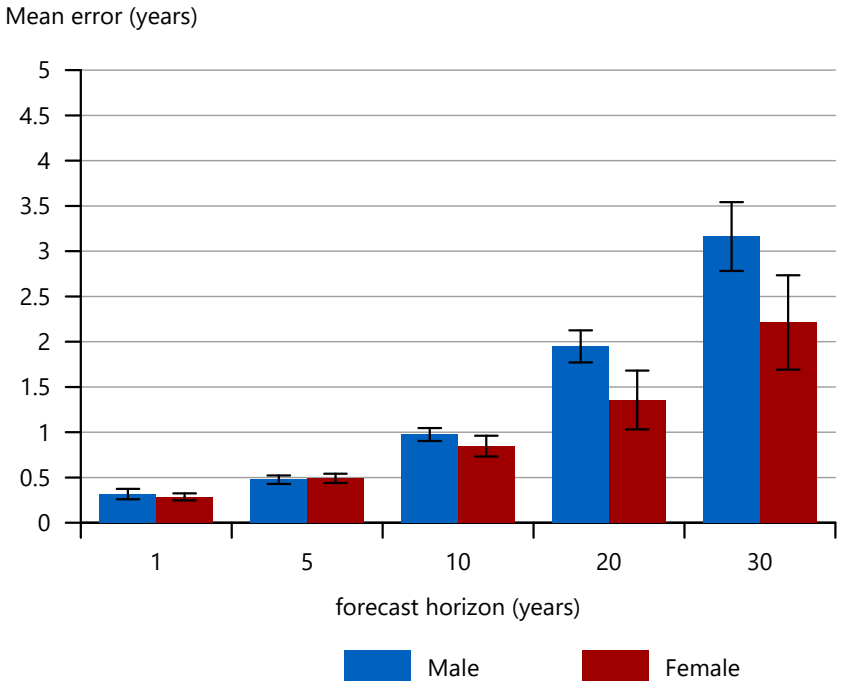
Tab. 6: Absolute error for life expectancy at age 65 forecasts for females

Jump-off		*demography*					*StMoMo*				*ensemble models*		
year	Actual	EM	LC_D	LC_Dc	C FDM	BMS	LC_S	APC	CBD	Plat	D	S	U
						1-year forecast horizon							
1950	14.66	0.11	0.07	0.05	0.06	0.06	0.08	0.19	0.99	0.12	**0.02**	0.13	0.08
1960	15.92	0.20	0.37	0.34	0.54	0.55	0.48	0.22	0.57	**0.06**	0.49	0.26	0.28
1970	16.17	0.32	0.47	0.40	0.51	0.53	0.46	0.57	**0.16**	0.28	0.50	0.44	0.43
1980	18.12	0.13	0.33	0.06	0.61	0.06	0.51	0.18	**0.05**	0.12	0.30	0.19	0.23
1990	19.32	0.46	0.32	0.07	0.39	0.07	0.43	0.20	0.27	**0.03**	0.26	0.20	0.27
	mean	0.24	0.31	0.18	0.42	0.25	0.39	0.27	0.41	**0.12**	0.32	0.25	0.26
						10-year forecast horizon							
1950	15.76	0.82	0.78	0.84	1.23	0.80	0.76	**0.26**	1.21	0.28	0.94	0.44	0.69
1960	15.71	0.76	0.22	0.27	0.31	**0.01**	0.12	0.93	0.07	1.29	0.03	0.75	0.55
1970	18.02	1.54	1.97	1.78	2.28	2.03	1.93	1.50	2.63	**0.75**	2.09	1.42	1.63
1980	19.08	0.41	0.69	**0.06**	1.58	0.17	0.81	0.26	1.32	1.14	0.74	0.14	0.33
1990	20.57	0.56	0.90	0.16	1.07	**0.01**	0.94	0.11	1.71	0.73	0.68	0.07	0.38
	mean	0.82	0.91	0.62	1.29	0.60	0.91	0.61	1.38	0.84	0.90	**0.56**	0.72
						20-year forecast horizon							
1950	15.71	0.58	0.48	0.58	1.10	0.49	0.43	0.93	2.20	0.65	0.69	0.35	**0.16**
1960	18.02	0.65	1.68	1.62	2.52	1.92	1.76	**0.20**	2.93	0.43	2.04	0.47	0.95
1970	19.08	1.89	2.65	2.32	3.22	2.72	2.56	0.88	4.29	**0.08**	2.87	1.29	1.86
1980	20.57	1.17	1.53	0.39	2.83	**0.15**	1.58	0.76	3.12	1.95	1.59	0.17	0.81
1990	21.92	0.63	1.51	0.24	1.61	**0.07**	1.46	0.88	3.23	1.57	1.06	0.16	0.42
	mean	0.98	1.57	1.03	2.25	1.07	1.56	0.73	3.15	0.93	1.65	**0.49**	0.84
						30-year forecast horizon							
1950	18.02	2.70	2.53	2.68	3.30	2.55	2.46	**0.40**	5.38	0.75	2.80	1.05	1.92
1960	19.08	0.78	2.34	2.27	3.46	2.61	2.41	1.50	4.60	0.96	2.82	**0.24**	1.12
1970	20.57	2.66	3.77	3.30	4.58	3.84	3.64	0.12	6.32	**0.11**	4.07	1.52	2.51
1980	21.92	1.78	2.25	0.69	3.91	0.33	2.21	1.84	4.73	2.80	2.32	**0.32**	1.22
1990	23.21	0.71	2.09	0.30	2.05	**0.17**	1.95	2.11	4.69	1.98	1.42	0.31	0.52
	mean	1.73	2.60	1.85	3.46	1.90	2.53	1.19	5.15	1.32	2.69	**0.69**	1.46
						Cumulative forecast errors after 30 years, averaged over the five forecasts							
	TOT	25.9	38.3	26.2	55.1	28.7	38.4	19.6	72.8	25.2	40.2	**13.4**	22.1
						Average forecast errors after 30 years, averaged over the five forecasts							
	Av^m	0.9	1.3	0.9	1.8	1.0	1.3	0.7	2.4	0.8	1.3	**0.4**	0.7

Notes: Actual life expectancies are presented in the second column from the left. If the jump-off year is 1970 and the forecast horizon is 10 years, then the relevant cell will show the estimated life expectancy at birth for 1980. The 12 rightmost columns present the absolute error in years of the life expectancy forecasts for the 12 evaluated models for each of the evaluated jump-off years and forecast horizons. The mean of the absolute errors across jump-off years for each evaluated forecast horizon are also presented. TOT is the average cumulative error over the five 30-year forecasts and is presented at the bottom of the table. The Av metric is equal to the TOT metric divided by the forecast horizon, providing a central measure of error across the forecast horizon.
Abbreviations: EM – extrapolative smoothing model, LC_D – Lee-Carter model from the `demography` package, LC_Dc Lee-Carter model from the `demography` package with automated base period selection, CFDM – coherent functional demographic model, LC_S – the `StMoMo` implementation of the Lee-Carter model, APC – Age-period-cohort model, CBD – Cairns-Blake-Dowd model, Plat – Plat model, D – demography ensemble, S – StMoMo ensemble, U – unique ensemble.

Source: Authors' calculations based on *HMD* (2023).

Fig. 3: Mean absolute error in forecasted life expectancy at age 65, averaged across models, by forecast horizon



Notes: This figure visualises the mean absolute error in forecasted life expectancy at age 65, which has been averaged across jump-off years and then across models. Error bars represent 95% confidence intervals.

Source: Authors' calculations based on *HMD* (2023).

Tab. 7: Absolute percentage error in projected total deaths using the pseudo-projection method

Jump-off year	*demography*					*StMoMo*				*ensemble models*		
	EM	LC_D	LC_Dc	CFDM	BMS	LC_S	APC	CBD	Plat	D	S	U
1-year forecast horizon												
1950	2.71	2.83	2.55	2.50	2.83	3.07	3.56	**1.40**	3.25	2.72	3.29	2.97
1960	0.32	1.15	1.02	2.54	1.71	1.23	2.21	**0.06**	1.86	1.03	0.94	0.47
1970	0.09	1.41	1.69	0.48	0.08	0.16	**0.04**	2.45	0.51	0.34	0.11	0.27
1980	0.30	0.71	0.97	4.19	5.84	5.92	2.11	0.38	1.61	3.12	0.74	**0.02**
1990	4.18	0.84	0.29	3.23	0.71	5.29	0.71	1.87	**0.14**	1.13	1.58	1.54
mean	1.52	1.39	1.30	2.59	2.23	3.13	1.73	1.23	1.47	1.67	1.33	**1.05**
10-year forecast horizon												
1950	1.61	0.76	2.26	1.43	0.72	**0.40**	5.37	11.39	2.83	0.99	2.56	0.83
1960	8.49	7.85	7.20	4.85	5.56	5.94	15.02	**1.51**	11.08	6.08	10.57	9.35
1970	16.07	18.95	20.90	16.88	17.51	17.34	**7.38**	29.65	14.38	17.79	13.16	14.85
1980	6.09	4.20	4.35	12.36	9.17	9.90	3.05	14.40	**2.34**	8.72	3.29	4.62
1990	5.45	9.13	3.91	11.40	1.02	12.65	**0.68**	19.64	8.19	7.32	7.37	7.16
mean	7.54	8.18	7.72	9.38	6.80	9.25	**6.30**	15.32	7.76	8.18	7.39	7.36
20-year forecast horizon												
1950	4.28	4.84	**2.61**	4.28	4.89	5.23	17.61	6.04	9.73	4.64	10.64	7.92
1960	7.19	10.52	11.64	12.79	12.63	12.24	7.10	23.80	6.68	12.02	**4.52**	6.56
1970	16.60	21.21	24.79	18.97	19.74	19.55	**1.84**	32.20	19.01	20.02	13.97	15.98
1980	17.73	15.79	15.37	23.14	18.89	20.33	**0.57**	29.15	16.14	19.59	13.15	15.50
1990	5.99	13.57	5.79	13.47	**1.56**	16.09	2.60	26.29	15.18	9.89	10.39	9.84
mean	10.36	13.19	12.04	14.53	11.54	14.69	**5.94**	23.50	13.35	13.23	10.53	11.16
30-year forecast horizon												
1950	13.66	13.67	16.60	14.17	13.62	13.26	10.14	24.98	10.92	13.87	**5.61**	9.32
1960	8.41	13.62	14.98	15.28	15.43	14.94	11.03	25.43	15.37	14.86	**7.92**	9.68
1970	25.14	31.28	36.01	29.05	29.91	29.61	**4.00**	41.12	38.87	30.17	25.66	26.97
1980	20.71	20.00	18.55	25.53	22.05	23.36	**3.80**	31.12	27.04	22.99	17.42	19.72
1990	3.19	14.19	3.48	11.09	**1.31**	15.60	9.40	26.12	22.16	8.59	11.38	9.80
mean	14.22	18.55	17.92	19.02	16.46	19.35	**7.67**	29.75	22.87	18.10	13.60	15.10
Cumulative forecast errors after 30 years, averaged over the five forecasts												
TOT	244.5	299.9	284.2	333.3	268.0	336.4	**171.7**	555.4	311.8	299.1	239.5	251.5
Average forecast errors after 30 years, averaged over the five forecasts												
Av^m	8.2	10.0	9.5	11.1	8.9	11.2	**5.7**	18.5	10.4	10.0	8.0	8.4

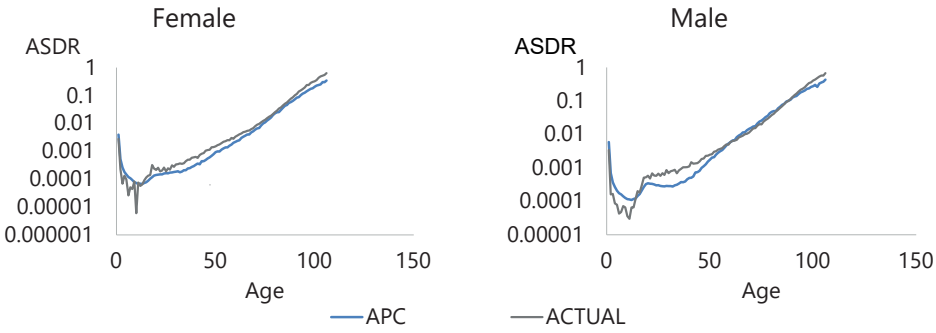
Notes: The 12 rightmost columns present the absolute percentage error in total deaths for the 12 evaluated models for each of the evaluated jump-off years and forecast horizons. The mean of the errors across jump-off years for each evaluated forecast horizon are also presented. TOT is the average cumulative error over the five 30-year forecasts and is presented at the bottom of the table. The Av metric is equal to the TOT metric divided by the forecast horizon, providing a central measure of error across the forecast horizon.
Abbreviations: EM – extrapolative smoothing model, LC_D – Lee-Carter model from the `demography` package, LC_Dc Lee-Carter model from the `demography` package with automated base period selection, CFDM – coherent functional demographic model, LC_S

– the `StMoMo` implementation of the Lee-Carter model, APC – Age-period-cohort model, CBD – Cairns-Blake-Dowd model, Plat – Plat model, D – demography ensemble, S – StMoMo ensemble, U – unique ensemble.

Source: Authors' calculations based on *HMD* (2023), *ABS data* (2019, 2023) and the Australian Demographic Data Bank (*Smith* 2009).

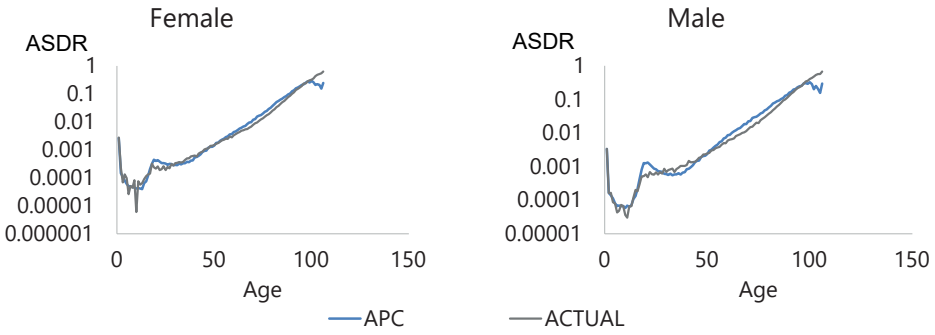
horizons for both males and females. CFDM performs poorly for older females (>50 years). ASDRs for the 30-year forecast horizon, for the 1990 jump-off year are shown in Figures 4, 5 and 6 for the APC, CFDM, and EM models, respectively. CFDM exhibits atypical forecasts for the highest ages (Fig. 5). When considering the forecasts at the highest ages (100+), as detailed in Tables B1–B6, it is evident that some methods produced more sensible forecasts, such as the APC and EM models. Conversely, the CFDM and Lee-Carter models (LC_D and LC_S) often produced poor forecasts for the highest ages.

Fig. 4: Forecast ASDRs with the APC model for a 30-year forecast horizon, 1990 jump-off



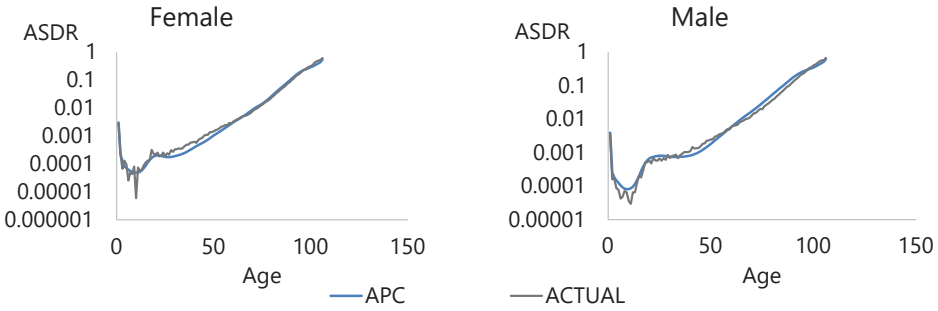
Source: Authors' calculations based on *HMD* (2023).

Fig. 5: Forecast ASDRs with the CFDM model for a 30-year forecast horizon, 1990 jump-off



Source: Authors' calculations based on *HMD* (2023).

Fig. 6: Forecast ASDRs with the EM model with a 30-year forecast horizon, 1990 jump-off



Source: Authors' calculations based on *HMD* (2023).

4 Discussion

4.1 Model Performance

Despite the variability of the results, there were three consistent findings: (1) the strong performance of the APC model, (2) the EM method's robustness, and (3) the poor performance of the CBD method when implemented across a full age range. The APC model's ability to capture cohort effects, in addition to age and period effects, may be a key factor in its strong performance, particularly for longer horizons and older age groups. Cohort-specific trends, such as health behaviours, medical advancements, and socio-economic factors, are important for understanding mortality trajectories. These cohort effects become more pronounced over longer forecast periods, aligning with the APC model's theoretical strengths. Conversely, *Dowd et al.* (2010) found that the APC model performed similarly to other models for forecasting the mortality of older males in England and Wales in an evaluation which considered the stability of the forecasts with shifting forecast horizons (*Dowd et al.* 2010). *Devi Fokeer* and *Narsoo* (2022) found that the APC model was most effective for ages 0-19, but performance was mixed for older ages for one year ahead mortality forecasts. In contrast, the Plat model, which also includes cohort effects, but uses a more detailed age-specific framework than the APC model, did not perform as well, particularly in forecasting total deaths using the pseudo-projection method. Specifically, the Plat model struggled with forecasting mortality rates for females aged 22-58, which likely contributed to its higher mean error of 22.9 percent in total death projections across the five 30-year forecasts from different jump-off years.

Our research aligns with previous work that indicated that the EM model was one of the best performers in an evaluation of the accuracy of forecasted ASDRs for Australian populations aged 50-100 (*Terblanche* 2016). The EM model's robustness may be linked to its incorporation of demographic constraints, such as ensuring that ASDRs do not decrease with increasing age after the age of 15 years, and that male ASDRs remain higher than or equal to female ASDRs. These assumptions

enhance the model's reliability, particularly for extended horizons, and align well with demographic realities. The EM method has not been used widely and is rarely included in evaluations of mortality forecasting methods. The results of this study suggest that it would be worthwhile to consider it as a potential useful mortality forecasting method. Next, we consider that one of the consistent findings of our study was the poor performance of the CBD method. This is not unexpected, given that the method was designed to be applied for older ages, both in terms of the data used to fit the model and the ages for which the mortality forecasts were created. Practitioners should be aware that this method is not suitable for the full age range.

Compared to other models, the CFDM method performed better for males than for females, where it was among the poorest performers. The CFDM method forecasts male and female data together – as products and ratios. Forecasts of male mortality generally have greater error than mortality forecasts for females. For example, in this study, the average error in forecast life expectancy at age 65, for a 30-year forecast horizon, was 3.27 years for males compared to 2.39 years for females. The greater error in the male forecasts may increase the relative uncertainty of the female forecasts when using the CFDM method. These findings align with *Booth* (2020), who observed that coherent mortality forecasting methods like CFDM may produce less accurate forecasts for lower-mortality populations when combined with higher-mortality populations. Specifically, Booth noted that when male and female mortality data are modelled together, forecasts for the lower-mortality group, typically females, can be negatively influenced by trends in the higher-mortality group. *Bergeron-Boucher* and *Kjærgaard* (2022) also found that symmetric mean absolute percentage errors were generally greater for males than for females in an evaluation of the Lee-Carter method and its variants using mortality data from Canada, Denmark, Italy, and Sweden. *Shang* (2015) found that forecast errors were greater for males than for females; however, the study found that CFDM produced relatively good forecasts of male mortality rates and suggested using CFDM as a benchmark method. It is important to note that the advantage of the CFDM method is not greater accuracy – rather, the method allows for the creation of coherent male and female mortality forecasts that do not diverge over longer forecast horizons. This is an important property given that mortality forecasts often need to be produced for longer horizons – even to 100 years (*Woods/Dunstan* 2014). We thus suggest that practitioners using the CFDM method be mindful that it may produce relatively poorer forecasts for females. Similarly, practitioners should be aware that not only are male mortality rates typically higher than female mortality rates, but forecasts of male mortality are generally more inaccurate.

Ensemble models, whilst not always more accurate than individual models, tend to reduce the frequency of bad forecasts (*Grossman et al.* 2022). However, *Li* (2022) found that simple averaging did not outperform individual base models. In this study, the ensemble models performed better than most of their individual constituent models. Whilst the S and U ensembles generally yielded robust results, the D ensemble did not perform particularly well, likely due to the inclusion of models with high errors (such as the CFDM model for females). These forecasts were notably different from those of the other individual models and could have

been identified and removed from the ensembles if a trimmed ensemble approach had been implemented (see *Rayer et al.* 2009). Future research should consider this accessible combination method, which has the potential to improve ensemble model reliability for practitioners.

4.2 The Pseudo-Projection Method for Mortality Forecast Evaluation

Unlike the sex-specific metrics presented here, the pseudo-projection method evaluates total deaths, providing a single, aggregate measure that connects mortality rate errors to population-level outcomes. While metrics such as life expectancy and ASDR errors highlight specific aspects of forecast performance, they do not fully capture the cumulative impact of mortality forecast errors on total deaths and thus older population forecasts, which are vital for planning in sectors such as healthcare and aged care. By offering this broader perspective, the pseudo-projection method complements traditional metrics and provides practitioners with actionable insights into the real-world implications of forecast errors.

The pseudo-projection results revealed notable differences in forecast performance across methods. For example, over a 30-year horizon, the APC model consistently outperformed others, with a mean error in total deaths of 7.7 percent across five forecasts. This was substantially lower than the next best methods: the S ensemble (13.6 percent) and EM (14.2 percent). The Plat model performed poorly, with a mean error of 22.9 percent, a result not immediately evident from other metrics. For instance, Plat's ASDR errors over the same horizon were comparable to better-performing methods, averaging 0.036 for females and 0.023 for males, while its life expectancy errors were similarly unremarkable. However, Online Appendix B6 reveals Plat's challenges with specific age groups – particularly females aged 22-58 – likely contributing to its weaker total death projections. This illustrates how errors in specific age groups can compound to influence aggregate outcomes, demonstrating the pseudo-projection method's ability to identify issues that other metrics might overlook.

These results also highlight the influence of male and female mortality on total deaths. Although deaths for males and females are calculated separately, their contributions to total errors reflect underlying differences in mortality rates and population sizes, with male mortality rates typically exhibiting greater variability. The APC model's strong performance across all metrics suggests it handles these dynamics particularly well, while Plat's high pseudo-projection errors underscore how even moderate ASDR inaccuracies for key age groups can cascade into larger aggregate errors. Future research could explore applying the pseudo-projection method to males and females separately, which may provide deeper insights into how sex-specific errors propagate to total deaths. Such an approach could complement aggregate evaluations and further enhance the method's value for practitioners.

5 Conclusion

This study evaluated nine individual and three ensemble mortality forecasting models by producing several retrospective forecasts for Australia. Results were evaluated in terms of errors in mortality rates, life expectancy at birth, life expectancy at age 65, and total numbers of deaths. A pseudo-projection method was used to evaluate mortality forecast accuracy in terms of total deaths. The methods evaluated in this study are implemented in widely used R packages (`StMoMo` and `demography`), which offer robust, well-documented, and user-friendly implementations. These packages allow even complex models such as APC and CFDM to be readily applied in real-world settings, bridging the gap between academic research and practical applications. By focusing on methods available through these established R packages, we hope that our evaluation supports practitioners in selecting methods that are not only accurate but also accessible and practical. The code used to create our forecasts is also made available, including for the EM model, which is not currently included in any of the popular R packages. In doing so, we hope to allow the findings of this paper to be relevant to practitioners. The main findings of the study are summarised here:

- The APC model performed strongly in our evaluation, particularly for older ages and longer forecast horizons, across several metrics for both males and females. These results suggest that the APC model is well-suited for long-term mortality forecasts. Practitioners may find the APC model a useful option for producing forecasts over extended horizons.

- The CBD method performed poorly, emphasising that it should not be used outside of its intended use for older populations.

- The inclusion of poor forecasts appeared to substantively decrease the accuracy of the ensemble models. Further investigation of the use of trimmed methods with a wide range of base models could prove useful.

- The EM method produced reliable forecasts across all metrics. The robustness of the EM model may be linked to its incorporation of demographic constraints, such as ensuring that ASDRs do not decrease with age after the age of 15 years, and that male ASDRs remain higher than or equal to female ASDRs. These assumptions enhance its reliability, particularly for extended forecast horizons, and align with demographic realities.

- Errors in mortality forecasts tend to be greater for males than for females.

- The use of methods which forecast male and female mortality rates together, such as CFDM, can contribute to relatively poorer performance

for forecasts of female mortality due to the relatively larger errors in the male mortality forecasts.

- The pseudo-projection method revealed significant differences in the error in total deaths generated by the best and worst performing methods over a 30-year forecast horizon (7.7 percent vs 29.8 percent), with the Plat model amongst the weaker performers (22.9 percent). These issues were not immediately apparent from the other metrics, such as life expectancy or ASDR errors. By linking forecast errors to their cumulative impact on total deaths, the pseudo-projection method provided a clearer understanding of how forecasting methods perform in practical applications, offering guidance for population projections and policy planning.

- Several methods resulted in 30-year errors of up to 20 percent in aggregate forecasts of total deaths when evaluated using the pseudo-projection method. This indicates that substantial errors can occur in long-term mortality forecasts, even at the national level, and such significant discrepancies in total death projections could have significant impacts on overall planning and policy decisions that rely on these figures.

- The top performing methods for shorter forecast horizons were not necessarily the top performing methods for longer horizons. Furthermore, there is often little separating the top methods for the shorter horizons. Differences between methods become more evident with longer horizons. Given that practitioners are often required to produce long-term mortality forecasts, where practical, longer forecast horizons are helpful when evaluating and developing mortality forecast methods.

There are also several limitations that readers should consider in their interpretation of the results. First, our study focused on specific implementations of forecast models. However, these methods can be configured in many ways, and the specific configuration can impact performance. This is evident in the differences in the performance of the `StMoMo` and `demography` package implementations of the Lee-Carter method, as well as between the LC_D and LC_Dc methods. We endeavoured to evaluate methods 'off the shelf' and to establish how they performed on raw data with minimal pre-processing. However, practitioners often employ additional techniques, such as smoothing and constraining, alongside forecasting methods. These steps can influence the results, and future evaluations should consider their impact on accuracy.

Our simple data preparation approach, which involved smoothing over zeros, NAs, and infinite values, and limiting ASDRs to a maximum of 1, may have influenced results. As the various model implementations have different approaches to handling such values (e.g., models implemented with `StMoMo` functions assign zero weights to non-positive exposures and missing values), we selected a data preparation approach that prioritised interpretability and consistency across models. Future

evaluations could explore alternative approaches to data preparation to further refine performance comparisons.

We only conducted our evaluation for the Australian context. For more generalisable recommendations, the methods need to be rigorously tested for long forecast horizons for multiple datasets. The outcomes of the evaluations will depend on the metrics chosen, and different metrics will support different use cases for the mortality forecasts. Some users also require indicators of forecast uncertainty, such as prediction intervals. This paper focuses on point forecast accuracy, which is vital for practitioners, but further research is required to address uncertainty measures comprehensively, including through the application and evaluation of probabilistic forecasting methods.

Furthermore, practitioners will often consider features other than accuracy when selecting a forecasting method. These can include input data requirements, ease of application, and the coherence of male and female forecasts, amongst others. These practical considerations underscore the importance of ensuring forecast methods are not only accurate but also user-friendly and fit for purpose. However, there is often limited information available to researchers and package developers about how forecasts are used in practice and which aspects are most critical to users. Bridging this gap will require greater collaboration between researchers, developers, and practitioners to ensure forecast methods align with real-world needs. By focusing this evaluation on methods that are readily available in robust, widely used R packages, we hope to have provided practitioners with a practical resource for selecting models that are not only accurate but also accessible and adaptable to real-world applications.

Acknowledgments

This work was supported by the Australian Government through the Australian Research Council's Linkage Project funding scheme (project LP210200733). We would like to thank Andres Villegas Ramirez for his help in the implementation of forecast models. Kim Dunstan kindly provided helpful comments on a draft of the paper. All errors and omissions remain the authors' responsibility.

References

ABS 2019: Historical population [Data set]. In: https://www.abs.gov.au/, 18.04.2019. [https://www.abs.gov.au/statistics/people/population/historical-population/2016, 31.10.2023].

ABS 2023: National, state and territory population [Data set]. In: https://www.abs.gov.au/ 14.12.2023. [https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/jun-2023, 31.10.2023.]

Basellini, Ugofilippo; Camarda, Carlo Giovanni; Booth, Heather 2023: Thirty years on: A review of the Lee-Carter method for forecasting mortality. In: International Journal of Forecasting 39,3: 1033-1049. https://doi.org/10.1016/j.ijforecast.2022.11.002

Bengtsson, Tommy; Keilman, Nico 2019: Old and new perspectives on mortality forecasting. Cham: Springer Nature. https://doi.org/10.1007/978-3-030-05075-7

Bergeron-Boucher, Marie Pier; Kjærgaard, Søren 2021: Mortality forecasting at age 65 and above: an age-specific evaluation of the Lee-Carter model. In: Scandinavian Actuarial Journal 2022, 1: 64-79. https://doi.org/10.1080/03461238.2021.1928542

Booth, Heather 2020: Coherent Mortality Forecasting with Standards: Low Mortality Serves as a Guide. In: *Mazzuco, Stefano; Keilman, Nico* (Eds.): Developments in Demographic Forecasting. The Springer Series on Demographic Methods and Population Analysis. Vol 49. Cham: Springer: 153-178. https://doi.org/10.1007/978-3-030-42472-5_8

Booth, Heather; Maindonald, John; Smith, Len 2002: Applying Lee-Carter under conditions of variable mortality decline. In: Population Studies 56,3: 325-336. https://doi.org/10.1080/00324720215935

Booth, Heather et al. 2003: demography/R/lca.R. GitHub repository: robjhyndman/demography [https://github.com/robjhyndman/demography/blob/master/R/lca.R, 14.04.2025].

Booth, Heather et al. 2006: Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. In: Demographic Research 15: 289-310. https://doi.org/10.4054/DemRes.2006.15.9

Cairns, Andrew J. G.; Blake, David; Dowd, Kevin 2006: A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. In: Journal of Risk and Insurance 73,4: 687-718. https://doi.org/10.1111/j.1539-6975.2006.00195.x

Cairns, Andrew J. G et al. 2009: A Quantitative Comparison of Stochastic Mortality Models Using Data From England and Wales and the United States. In: North American Actuarial Journal 13,1: 1-35. https://doi.org/10.1080/10920277.2009.10597538

Devi Fokeer, Oopashna; Narsoo, Jason 2022: Evaluation of the forecasting accuracy of stochastic mortality models: An analysis of developed and developing countries. In: Communications in Statistics: Case Studies, Data Analysis and Applications 8,3: 434-462. https://doi.org/10.1080/23737484.2022.2093294

Dowd, Kevin et al. 2010: Backtesting stochastic mortality models: an ex post evaluation of multiperiod-ahead density forecasts. In: North American Actuarial Journal 14,3: 281-298. https://doi.org/10.1080/10920277.2010.10597592

Ediev, Dalkhat M. 2008: Extrapolative projections of mortality: Towards a more consistent method part I: The central scenario. Vienna Institute of Demography Working Papers, 3/2008. Vienna Institute of Demography, Austrian Academy of Sciences. Vienna.

Grossman, Irina et al. 2022: Can machine learning improve small area population forecasts? A forecast combination approach. In: Computers, Environment and Urban Systems 95: 101806. https://doi.org/10.1016/j.compenvurbsys.2022.101806

HMD (Human Mortality Database) 2023: Human Mortality Database [Data set]. Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France) [www.mortality.org, 05.12.2023].

Hyndman, Rob J. 2023: demography: Forecasting Mortality, Fertility, Migration and Population Data. R package version 2.0. [https://CRAN.R-project.org/package=demography, 14.04.2025].

Hyndman, Rob J.; Ullah, Md Shahid 2007: Robust forecasting of mortality and fertility rates: A functional data approach. In: Computational Statistics & Data Analysis 51,10: 4942-4956. https://doi.org/10.1016/j.csda.2006.07.028

Hyndman, Rob J.; Khandakar, Yeasmin 2008: Automatic time series forecasting: the forecast package for R. In: Journal of Statistical Software 27: 1-22. https://doi.org/10.18637/jss.v027.i03

Hyndman, Rob J.; Booth, Heather; Yasmeen, Farah 2013: Coherent mortality forecasting: the product-ratio method with functional time series models. In: Demography 50,1: 261-283. https://doi.org/10.1007/s13524-012-0145-5

Hyndman, Rob J. et al. 2023: Forecast: Forecasting functions for time series and linear models. R package version 8.21. In: https://cran.r-project.org/. [https://CRAN.R-project.org/package=demography, 05.12.2023].

Lee, Ronald D.; Carter, Lawrence R. 1992: Modeling and forecasting U.S. mortality. In: Journal of the American Statistical Association 87,419: 659-671. https://doi.org/10.2307/2290201

Lee, Ronald; Miller, Timothy 2001: Evaluating the performance of the Lee-Carter method for forecasting mortality. In: Demography 38,4: 537-549. https://doi.org/10.1353/dem.2001.0036

Li, Jackie 2023: A model stacking approach for forecasting mortality. In: North American Actuarial Journal 27,3: 530-545. https://doi.org/10.1080/10920277.2022.2108453

Osmond, Clive 1985: Using age, period and cohort models to estimate future mortality rates. In: International Journal of Epidemiology 14,1: 124-129. https://doi.org/10.1093/ije/14.1.124

Plat, Richard 2009: On stochastic mortality modeling. In: Insurance: Mathematics and Economics 45,3: 393-404. https://doi.org/10.1016/j.insmatheco.2009.08.006

Rayer, Stefan; Smith, Stanley K.; Tayman, Jeff 2009: Empirical prediction intervals for county population forecasts. In: Population Research and Policy Review 28: 773-793. https://doi.org/10.1007/s11113-009-9128-7

Renshaw, Arthur E.; Haberman, Steven 2006: A cohort-based extension to the Lee-Carter model for mortality reduction factors. In: Insurance: Mathematics and Economics 38,3: 556-570. https://doi.org/10.1016/j.insmatheco.2005.12.001

Shang, Han Lin 2015: Statistically tested comparisons of the accuracy of forecasting methods for age-specific and sex-specific mortality and life expectancy. In: Population Studies 69,3: 317-335. https://doi.org/10.1080/00324728.2015.1074268

Shang, Han Lin; Booth, Heather; Hyndman, Rob J. 2011: Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. In: Demographic Research 25: 173-214. https://doi.org/10.4054/DemRes.2011.25.5

Smith, Leonard et al. 2009: The Australian Demographic Data Bank, 1901-2003: Populations [Data set]. Canberra, ACT, Australia: Australian Social Science Data Archive, The Australian National University.

Terblanche, Wilma 2015: Population estimates and projections for Australia's very elderly population at state and national level [PhD thesis]. In: The University of Queensland https://core.ac.uk/ [https://core.ac.uk/download/pdf/43380972.pdf, 03.01.2024].

Terblanche, Wilma 2016: Retrospective testing of mortality forecasting methods for the projection of very elderly populations in Australia. In: Journal of Forecasting 35,8: 703-717. https://doi.org/10.1002/for.2404

Thatcher, A. Roger; Kannisto, Väinö; Vaupel, James W. 1998: Appendix A: Definitions and approximations. In: The force of mortality at ages 80 to 120. Odense University Press [https://www.demogr.mpg.de/papers/books/monograph5/appenda.htm, 11.01.2024].

Villegas, Andrés; Kaishev, Vladimir; Millossovich, Pietro 2018: StMoMo: An R Package for Stochastic Mortality Modeling. In: Journal of Statistical Software 84,3: 1-38. https://doi.org/10.18637/jss.v084.i03

Villegas, Andrés; Millossovich, Pietro; Kaishev, Vladimir 2025: StMoMo: Stochastic Mortality Modelling [R package reference manual]. Comprehensive R Archive Network (CRAN) [https://cran.r-project.org/web/packages/StMoMo/StMoMo.pdf, 14.04.2025].

Wilson, Tom; Rees, Philip 2021: A brief guide to producing a national population projection. In: Australian Population Studies 5,1: 77-100. https://doi.org/10.37970/aps.v5i1.84

Woods, Carmel; Dunstan, Kim 2014: Forecasting mortality in New Zealand. In: https://www.stats.govt.nz/ [https://www.stats.govt.nz/assets/Uploads/Research/Forecasting-mortality-in-New-Zealand-2014/forecasting-mortality-14-01-17feb14.pdf, 28.01.2024].

Date of submission: 22.04.2024 *Date of acceptance: 17.04.2025*

Dr. Irina Grossman (✉). RMIT University, The University of Melbourne. Australia.
E-mail: irina.grossman@rmit.edu.au; https://orcid.org/0000-0002-5761-6194
URL: https://www.rmit.edu.au/profiles/g/irina-grossman

Dr. Tom Wilson. Advanced Demographic Modelling. Melbourne, Australia.
E-mail: tom.demographer@protonmail.com; http://orcid.org/0000-0001-8812-7556
URL: https://drtomwilson.com/

Comparative Population Studies

www.comparativepopulationstudies.de

ISSN: 1869-8980 (Print) – 1869-8999 (Internet)

Published by
Federal Institute for Population Research (BiB)
65180 Wiesbaden / Germany

E-mail: cpos@bib.bund.de