Quality & Quantity
https://doi.org/10.1007/s11135-022-01537-z

Backtesting stochastic mortality models by prediction interval-based metrics

Salvatore Scognamiglio[1] · Mario Marino[2]

Accepted: 17 September 2022

Abstract

Human lifespan increments represent one of the main current risks for governments and pension and health benefits providers. Longevity societies imply financial sustainability challenges to guarantee adequate socioeconomic conditions for all individuals for a longer period. Consequently, modelling population dynamics and projecting future longevity scenarios are vital tasks for policymakers. As an answer, the demographic and the actuarial literature have been introduced and compared to several stochastic mortality models, although few studies have thoroughly tested the uncertainty concerning mortality projections. Forecasting mortality uncertainty levels have a central role since they reveal the potential, unexpected longevity rise and the related economic impact. Therefore, the present study poses a methodological framework to backtest uncertainty in mortality projections by exploiting uncertainty metrics not yet adopted in mortality literature. Using the data from the Human Mortality Database of the male and female populations of five countries, we present some numerical applications to illustrate how the proposed criterion works. The results show that there is no mortality model overperforming the others in all cases, and the best model choice depends on the data considered.

Keywords Stochastic mortality models · Prediction interval forecasting · Models comparison · Backtesting metrics

1 Introduction

During the last century, populations around the world have experienced a continuous longevity growth, despite country-specific dynamics and disparities in age-at-death distribution (Aburto and van Raalte 2018; Nigri et al. 2022). Broadly speaking, mortality has

✉ Salvatore Scognamiglio
salvatore.scognamiglio@uniparthenope.it

Mario Marino
m.marino@uniroma1.it

[1] Department of Management and Quantitative Studies, University of Naples Parthenope, Naples, Italy

[2] Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome, Rome, Italy

Published online: 29 September 2022

declined at all ages, with different intensities, due to the action of heterogeneous factors affecting the human lifespan. For instance, the diffusion of preventative health measures and improved medical care have positively impacted on population's life expectancy (see, e.g. Vaupel et al. 2021 and Zarulli et al. 2021). Moreover, socioeconomic and environmental conditions have been determinants of how mortality changed over time and among populations (see, e.g. Cairns et al. 2019 and Khomenko et al. 2021). Although human lifespan increments are enjoyable evidence, as longevity increases, economic challenges emerge to ensure an adequate income and health treatments for all individuals throughout their old age. The social costs to cater for the needs of increasingly elderly populations involve a burdensome financial position for governments, annuity providers and pension schemes. These entities face the risk of paying out benefits for a much longer period, suffering financial instability conditions. In addition, such a circumstance may be exacerbated by future social advances. For instance, it is plausible that scientific and technological progresses will be able to boost medicine and lifestyles behaviours, making future improvements in life expectancy highly unpredictable (see, e.g. Keilman 2019). Indeed, population forecasts are heavily based on the current knowledge in vital processes, from which to extrapolate future trends. As a consequence, unexpected changes may occur due to random fluctuations or structural changes in the mortality trends. Therefore, uncertainty in predicting future longevity scenarios is a fact and achieving accuracy in mortality forecasts is crucial to support a long-lived society financially. To this end, both the demographic and the actuarial literature have been enriched over time by excellent works focused on stochastic mortality modelling. Models widely used to forecast mortality rates are extrapolative and draw inspiration from the pioneering Lee-Carter model (Lee and Carter 1992, hereinafter LC). The latter is a log-bilinear model embedding age-period effects and an additive Gaussian error structure. Various LC model extensions have been proposed in the literature to overcome model weaknesses and describe other relevant mortality patterns. For instance, Brouhns et al. (2002) formulated the canonical LC model in terms of Poisson regression, later extended in a Bayesian framework by Czado et al. (2005). A multi-factor age–period extension of LC was proposed by Booth et al. (2002), while Renshaw and Haberman (2003) exploited the LC model to provide mortality forecasting with age-specific enhancement. Prominent developments in stochastic mortality modelling were furnished by Renshaw and Haberman (2006), introducing cohort effects, and by Cairns et al. (2006). The latter proposed a two-factor stochastic representation for the logit of death probabilities, opening the way for further generalizations embodying multiple period and cohort effects (Cairns et al. 2009). Afterwards, Plat (2009) gathered the LC, the Renshaw and Haberman and the Cairns-Blake-Dowd models to establish a unified approach. Since a large number of mortality models have been proposed in the literature, Hunt and Blake (2014) designed a general procedure for constructing parametric mortality models able to catch different age-period-cohort effects present in the data.

Designating the best mortality model is not trivial, and the choice stems from the satisfaction of suitable criteria. However, a model may not necessarily dominate all others based on the selected criteria. For instance, a model could be better than others looking at its goodness of fit to historical data; at the same time, such a model could be less robust or provide less accurate projections than others. Furthermore, the model comparison depends on the mortality data investigated and the purposes of the analysis. Subsequently, many articles in the literature have compared different mortality models exploiting specific mortality experiences and befitting metrics. Cairns et al. (2009) offered an extensive analysis comparing empirical fits of eight stochastic mortality models on US and England & Wales males mortality data. The authors examined model performances by means of both qualitative and quantitative model selection criteria. The former refers to desirable mortality model's features, that is, parsimony, transparency, ability to generate sample paths, incorporation of cohort effects, aptitude to provide non-trivial correlation structures between ages. Quantitative criteria allow for assessing consistency and robustness of parameter estimates with respect to the period of data observed. Dowd et al. (2010b) tailored hypothesis tests to strengthen models evaluation in terms of goodness-of-fit and to consider England & Wales males mortality data. A different perspective to benchmark mortality models was developed in Dowd et al. (2010a). The latter designed a backtesting framework to gauge (ex-post) the forecasting performances of six stochastic mortality models fitted on England & Wales male mortality data. Cairns et al. (2011) focused on the plausibility of stochastic mortality model forecasts by means of innovative qualitative criteria, namely: biological reasonableness, plausibility of predicted levels of uncertainty in forecasts at different ages, projections robustness with respect to the sample period used to fit the model. A considerable model comparison is also promoted in Haberman and Renshaw (2011), whose numerical experiment include both US and England & Wales female's mortality experiences. Additional quantitative model comparisons have been performed in Lovász (2011), concerning Finnish and Swedish populations, and in Biffi and Clemente (2014) and Carfora et al. (2017), concerning Italian population.

In the spirit of the aforementioned literature, the present study furnishes a comparative analysis of the forecasting ability of stochastic mortality models. As suggested in Dowd et al. (2010a), a suitable mortality model should offer both proper in-sample results and plausible ex-ante forecasts, as well as generate adequate ex-post out-of-sample performances. In particular, forecasted uncertainty levels have a central role since it reveals the potential, unexpected longevity rise and its biological plausibility. Therefore, we implement an ex-post assessment of the projected prediction intervals, comparing and measuring their plausibility through uncertainty metrics not yet adopted: the Prediction Interval Coverage Probability (hereinafter PICP) and the Mean Prediction Interval Width (hereinafter MPIW). The former generally represents the probability of observing mortality outcomes over the backtesting horizon falling within the prediction intervals. At the same time, the latter expresses the average amplitude of such intervals. Both the PICP and the MPIW are usually employed for rating neural networks prediction intervals (see, e.g. Khosravi et al. 2011, but they can be efficiently utilized for any predictive model. It is straightforward to note that prediction intervals whose PICP is the highest possible value are a matter of interest; we aim for all mortality realizations within the variability range. Such a scenario may be simply achieved through a wide prediction interval, but the latter could indicate poor predictive believability. For instance, too large prediction intervals are not informative about the likely uncertainty of future mortality outcomes; furthermore, too wide intervals may be associated with low or null PICP values. The latter is the worst-case scenario as it indicates the model's failure to capture the mortality trend. Hence, it is important to consider jointly the PICP and the MPIW, defining an associated criterion for the model comparison and selection. From a backtesting perspective, we advance a preference criterion that gives credibility to the mortality model with the highest PICP and the lower MPIW. Our proposal identifies an additional backtesting strategy whose joint application with commonly used quantitative metrics improves the model selection process. As further contribution, we illustrate how the suggested criterion performs ranking seven stochastic mortality models in an empirical analysis. To this end, we deal with the mortality experiences of five populations, namely Australia, England & Wales, Italy, Japan and the USA, and for both genders by exploiting the Human Mortality Database data (Human Mortality Database 2018). Our analysis shows that, as expected, no mortality model outperforms all the others. Despite both age- and country-specific characteristics, our proposal grants to identify models with mortality density forecasts more balanced across populations. The remainder of the paper is the following. Section 2 contains the methodological framework for investigating the stochastic mortality models. Section 3 explains the PICP and the MPIW metrics and the proposed preference criterion. Section 4 refers to the numerical experiments and related results obtained. Finally, Sect. 5 sets out concluding remarks.

2 Methodological framework

Let $\mathcal{X} = \{x_1, x_2, \dots, x_\omega\}$ be the age set, and $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$ the calendar year set. For each last birthday age $x \in \mathcal{X}$ during calendar year $t \in \mathcal{T}$, we introduce the following variables of interest:

- $D_{x,t}$, i.e. the random number of deaths;
- $d_{x,t}$, i.e. the observed number of deaths;
- $E^0_{x,t}$, i.e. the initial exposed to risk;
- $q_{x,t}$, i.e. the one-year probability of death.

In the present work we resort the Generalized Age-Period-Cohort (hereinafter GAPC) as unified mortality modeling framework, allowing for a fair comparison among stochastic mortality models (see e.g. Currie 2016; Villegas et al. 2018). In particular, we assume that:

$$D_{x,t} \sim \text{Binomial}(E^0_{x,t}, q_{x,t}), \quad \forall x \in \mathcal{X}, \forall t \in \mathcal{T}, \tag{1}$$

coherently representing the systematic mortality component by the following predictor:

$$\log \frac{q_{x,t}}{1 - q_{x,t}} = a_x + \sum_{l=1}^{L} b_x^{(l)} \kappa_t^{(l)} + c_x \gamma_{t-x}. \tag{2}$$

By the means of Eq. (2), one-year probabilities of death are characterized through a sigmoidal transformation of age-dependent parameters $a_x, b_x^{(l)}$ and c_x, time-dependent parameters $\kappa_t^{(l)}$, and the cohort parameter γ_p. Most of the stochastic mortality models in the literature are special cases extracted from the GAPC setup. For instance, consider whether the age-dependent parameter a_x or the cohort parameter γ_{t-x}, implies different mortality representations; again, varying the order of L, mortality models are shaped according to specific period effects portrayals.

2.1 Stochastic mortality models

We choose to deal with seven stochastic mortality models, summarized in Table 1, depicting the hinges of the two main stochastic mortality models families: the Lee-Carter family and the Cairns-Blake-Dowd family. We briefly illustrated their analytical structure in the following sections.

Table 1 Stochastic mortality models considered in the present work

Model	Notation	References
Lee-Carter model family		
Lee-Carter	LC	Lee and Carter (1992)
Renshaw-Haberman	RH	Renshaw and Haberman (2006)
Age-Period-Cohort	APC	Currie (2006)
Cairns-Blake-Dowd model family		
Cairns-Blake-Dowd	CBD	Cairns et al. (2006)
CBD with cohort effects	M6	Cairns et al. (2009)
Quadratic CBD with cohort effects	M7	Cairns et al. (2009)
Variant of CBD with cohort effects	M8	Cairns et al. (2009)

2.1.1 LC model

The LC model provides two age-specific effects, a period effect and it does not acknowledge for cohort influences. Thus, by the means of Eq. (2) we have $L = 1$ and $c_x, \gamma_{t-x} = 0$, obtaining:

$$\log \frac{q_{x,t}}{1 - q_{x,t}} = a_x + b_x^{(1)} \kappa_t^{(1)}. \tag{3}$$

We need to make the LC model invariant with respect to parameters' transformations, and the following constraints are then applied:

$$\sum_{x \in \mathcal{X}} b_x^{(1)} = 1, \quad \sum_{t \in \mathcal{T}} \kappa_t^{(1)} = 0.$$

2.1.2 RH model

The RH model develops the LC structure introducing an age-cohort bi-linear term, that is:

$$\log \frac{q_{x,t}}{1 - q_{x,t}} = a_x + b_x^{(1)} \kappa_t^{(1)} + c_x \gamma_{t-x}. \tag{4}$$

Similarly to the LC model, identifiability problems are avoided imposing the following constraints:

$$\sum_{x \in \mathcal{X}} b_x^{(1)} = 1, \quad \sum_{x \in \mathcal{X}} c_x = 1, \quad \sum_{t \in \mathcal{T}} \kappa_t^{(1)} = 0, \quad \sum_{p=t_0-x_\omega}^{t_n-x_0} \gamma_p = 0.$$

As suggested in Haberman and Renshaw (2011), we consider the specification $c_x = 1$ aiming to bypass robustness issues suffered by the predictor in Eq. (4).

2.1.3 APC model

The APC model was originally proposed in the fields of medicine and demography (see e.g. Clayton and Schifflers 1987) and later introduced in actuarial literature by Currie (2006). Referring to the latter, the APC model involves an age effect, a period effect and a cohort effect, that is:

$$\log \frac{q_{x,t}}{1-q_{x,t}} = a_x + \kappa_t^{(1)} + \gamma_{t-x}. \tag{5}$$

Such a model requires parameters constraints defined as below:

$$\sum_{t\in\mathcal{T}} \kappa_t^{(1)} = 0, \qquad \sum_{p=t_0-x_\omega}^{t_n-x_0} \gamma_p = 0.$$

2.1.4 CBD model

The CBD model assumes $L=2$ time-dependent parameters, while a_x, c_x, $\gamma_{t-x}=0$. In addition, Cairns et al. (2006) accounts for $b_x^{(1)}=1$ and $b_x^{(2)}=(x-\bar{x})$, where $\bar{x}$ denotes the average age for the age set considered. Hence, the predictor has the following expression:

$$\log \frac{q_{x,t}}{1-q_{x,t}} = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)}, \tag{6}$$

and it is not affected by identifiability issues.

2.1.5 M6 model

The M6 model is a CBD's extension due to the cohort effect inclusion. The latter is weighted by $c_x=1$, and the predictor becomes:

$$\log \frac{q_{x,t}}{1-q_{x,t}} = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + \gamma_{t-x}. \tag{7}$$

Contrary to the CBD model, constraints have to be introduced:

$$\sum_{p=t_0-x_\omega}^{t_n-x_0} \gamma_p = 0, \qquad \sum_{p=t_0-x_\omega}^{t_n-x_0} p\gamma_p = 0.$$

2.1.6 M7 model

The M7 model boosts the M6 one, favoring the introduction of a third period effect weighted by a quadratic coefficient. In particular, the M7's predictor is defined setting $L=3$, $a_x=0$, $b_x^{(1)}=1$, $b_x^{(2)}=(x-\bar{x})$, $b_x^{(2)}=((x-\bar{x})^2-\hat{\sigma}_x^2)$, $c_x^{(1)}=1$, where $\hat{\sigma}_x$ denotes the standard deviation of ages. In analytical terms, it holds that:

$$\log \frac{q_{x,t}}{1-q_{x,t}} = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + ((x-\bar{x})^2-\hat{\sigma}_x^2)\kappa_t^{(3)} + \gamma_{t-x}. \tag{8}$$

Regarding identifiability problems, the following constraints are imposed:

$$\sum_{p=t_0-x_\omega}^{t_n-x_0} \gamma_p = 0, \qquad \sum_{p=t_0-x_\omega}^{t_n-x_0} p\,\gamma_p = 0, \qquad \sum_{p=t_0-x_\omega}^{t_n-x_0} p^2\,\gamma_p = 0.$$

2.1.7 M8 model

The M8 model is M6 model's variant contemplating a non-constant coefficient related to the cohort effect, that is $c_x^{(1)} = (x_c - x)$ for some constant parameter x_c to be estimated. Thus, the predictor takes the form:

$$\log \frac{q_{x,t}}{1-q_{x,t}} = \kappa_t^{(1)} + (x-\bar{x})\kappa_t^{(2)} + (x_c - x)\gamma_{t-x}, \tag{9}$$

and the associated parameter constraint is:

$$\sum_{p=t_0-x_\omega}^{t_n-x_0} \gamma_p = 0.$$

2.2 Model fitting

Stochastic mortality models falling within the GAPC framework are examples of generalized, or non-generalized, linear models and they can be fitted by maximum likelihood estimation (see e.g. Currie 2016). Recalling assumption (1) and assuming i.i.d. death counts for each age and calendar year, we consider the log-likelihood below:

$$\log \mathcal{L}(d_{x,t}, \hat{d}_{x,t}) = \sum_{x,t} w_{x,t} \left[d_{x,t} \log\left(\frac{\hat{d}_{x,t}}{E^0_{x,t}}\right) + (E^0_{x,t} - d_{x,t}) \log\left(\frac{E^0_{x,t} - \hat{d}_{x,t}}{E^0_{x,t}}\right) + \binom{E^0_{x,t}}{d_{x,t}} \right], \tag{10}$$

where $w_{x,t}$ is a 0/1 weight taking value 0 if a data cell, (x, t), is ignored or 1 if the cell is incorporated, and $\hat{d}_{x,t}$ is the estimated number of deaths, that is:

$$\hat{d}_{x,t} = E^0_{x,t}\hat{q}_{x,t} = E^0_{x,t} \frac{\exp\left(\hat{a}_x + \sum_{l=1}^{L} \hat{b}_x^{(l)} \hat{\kappa}_t^{(l)} + \hat{c}_x \hat{\gamma}_{t-x}\right)}{1 + \exp\left(\hat{a}_x + \sum_{l=1}^{L} \hat{b}_x^{(l)} \hat{\kappa}_t^{(l)} + \hat{c}_x \hat{\gamma}_{t-x}\right)}. \tag{11}$$

It is straightforward to note that the parameters are estimated by solving the problem:

$$\hat{\theta} = \arg\max_{\theta} \log \mathcal{L}(d_{x,t}, \hat{d}_{x,t})$$

where $\hat{\theta} = \left(\hat{a}_x, \hat{b}_x^{(l)}, \hat{\kappa}_t^{(l)}, \hat{c}_x, \hat{\gamma}_{t-x}\right)$.

2.3 Model forecasting

Stochastic mortality models forecasts arise from time-series techniques describing period and cohort effects, while age-specific parameters are time invariant. Let $t_n \in \mathcal{T}$

be the forecasting year and let $\mathcal{T}' = \{t_n + h,\ h = 1, \dots, s\}$ be the forecasting horizon. Then, the h-step ahead forecast for the predictor is defined as:

$$\log \frac{q_{x,t_n+h}}{1 - q_{x,t_n+h}} = \hat{a}_x + \sum_{l=1}^{L} \hat{b}_x^{(l)} \kappa_{t_h+h}^{(l)} + \hat{c}_x \gamma_{t_n+h-x}, \quad \forall x \in \mathcal{X} \tag{12}$$

Periods effects at time $t \in \mathcal{T}'$, namely $\boldsymbol{\kappa}_t = (\kappa_t^{(1)}, \kappa_t^{(2)}, \dots \kappa_t^{(L)}) \in \mathbb{R}^L$, are shaped through a multivariate random walk:

$$\boldsymbol{\kappa}_t = \boldsymbol{\theta} + \boldsymbol{\kappa}_{t-1} + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \Sigma), \tag{13}$$

where $\boldsymbol{\theta} \in \mathbb{R}^L$ is the drift term and $\Sigma \in \mathbb{R}^{L \times L}$ is the variance-covariance matrix of $\boldsymbol{\epsilon}_t$. Obviously, for stochastic mortality models with $L = 1$, the random walk collapses to the univariate case. Concerning the cohort's parameter, we follow previous studies in literature (see e.g. Haberman and Renshaw 2011; Lovász (2011)) in order to generate projections embedding cohort effects. Therefore, the latter are modeled by an univariate ARIMA process, assuming stochastic independence with respect to $\boldsymbol{\kappa}_t$. In Table 2 we summarize the specific ARIMA(p,d,q) considered in the present work.

Point mortality forecasts are obtained determining the expectation of both time-dependent and cohort-dependent parameters, that is:

$$\log \frac{\hat{q}_{x,t_n+h}}{1 - \hat{q}_{x,t_n+h}} = \hat{a}_x + \sum_{l=1}^{L} \hat{b}_x^{(l)} \mathbb{E}\left(\kappa_{t_h+h}^{(l)}\right) + \hat{c}_x \mathbb{E}\left(\gamma_{t_n+h-x}\right), \quad \forall x \in \mathcal{X}, \tag{14}$$

so that future mortality trend is outlined without considering prediction errors. However, the latter generates uncertainty in mortality forecasts, as well as the estimation error related to $\hat{\boldsymbol{\theta}}$. Therefore, prediction intervals are substantial in supporting mortality projection's reliability and to perform consistent mortality/longevity risk assessments. Any stochastic mortality model have to provide prediction intervals with a certain coverage probability, namely $(1 - \alpha)$, aiming that future mortality outcomes belong to the interval defined by the lower and upper bounds. More in detail, indicating with $\hat{q}^{LB}$ the projected prediction interval lower bound and with $\hat{q}^{UB}$ the projected upper bound, a desirable model satisfies the following:

$$\mathbb{P}\left(\hat{q}_{x,t_n+h}^{LB} \leq q_{x,t_n+h} \leq \hat{q}_{x,t_n+h}^{UB}\right) \geq (1 - \alpha), \quad \alpha \in (0, 1). \tag{15}$$

In addition, prediction interval's quality depends on the distance between $\hat{q}_{x,t_n+h}^{LB}$ and $\hat{q}_{x,t_n+h}^{UB}$ for each time point in $\mathcal{T}'$. For our purposes, we appraise models' out-of-sample judging the quality of the projected prediction intervals in terms of their probability coverage and their width. In doing so, we take into account both model and parameter uncertainty. The method considered to generate such uncertainty is explained in Appendix A.

Table 2 ARIMA models for the cohort effect in the different mortality models

Mortality model	ARIMA model for γ_{t-x}
APC	ARIMA(1,1,0) with drift
RH	ARIMA(1,1,0) with drift
M7 and M8	ARIMA(2,0,0) with non-zero intercept

3 Prediction interval based metrics and forecasts comparison

To appreciate the prediction interval's goodness, we propose to jointly examine specific prediction interval-based metrics, namely PICP and MPIW. The former inspects the prediction interval coverage by counting how many mortality realisations are wrapped in the probabilistic range, given a confidence level $(1-\alpha)$. It is defined as

$$PICP = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \mathbb{1}_{\{q_{x,t}\,\in\,[\hat{q}^{LB}_{x,t},\,\hat{q}^{UB}_{x,t}]\}}, \tag{16}$$

where $\mathbb{1}_{\{\cdot\}} = 1$ if $q_{x,t} \in [\hat{q}^{LB}_{x,t},\ \hat{q}^{UB}_{x,t}]$, and $\mathbf{1}_{\{\cdot\}} = 0$ otherwise. Broadly speaking, the PICP furnishes the estimate for the probability in Eq. (15) so that, theoretically, PICP should be greater or equal to the nominal value $(1-\alpha)$. The higher the PICP value, the more likely the coverage the prediction interval provides for future mortality realizations. However, PICP values lower than the confidence level could occur due to different reasons, such as in the presence of noisy mortality data or when the model under-fitting or over-fitting crops up. Larger PICP values could be achieved simply by considering wider prediction intervals, but the latter suggests poor predictive believability, and they are not of practical usefulness. Therefore, rating prediction interval quality by PICP without considering the prediction interval width is a little choice. We must evaluate prediction interval accuracy, referring simultaneously to both PICP and MPIW. The latter represents the prediction interval mean width over the forecasting horizon, that is:

$$MPIW = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \hat{q}^{UB}_{x,t} - \hat{q}^{LB}_{x,t}. \tag{17}$$

The joint use of PICP and MPIW requires formulating a criterion to compare mortality models given their projected prediction intervals. Our approach traces a backtesting exercise. We provide a plain ex-post evaluation of prediction intervals inspecting the associated PICP and MPIW for each mortality model considered. In particular, we advance a preference criterion relying on the mortality model with the highest PICP and the lower MPIW. For instance, given two stochastic mortality models, namely M_k and M_j, we propose to promote the model satisfying the following criterion:

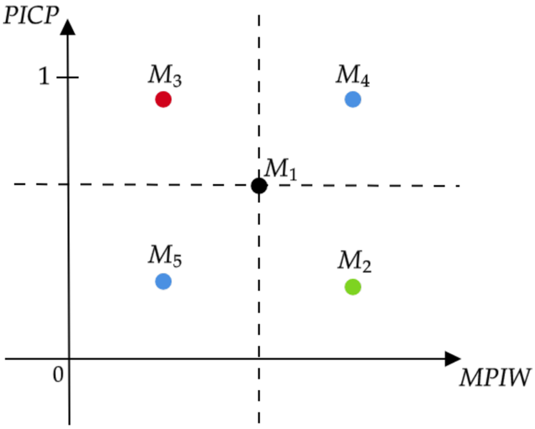
$$M_k \succ M_j \quad \text{if} \quad PICP_k > PICP_j \ \wedge \ MPIW_k \leq MPIW_j. \tag{18}$$

The criterion in (18) may be drafted also in terms of weak preference, that is:

$$M_k \succeq M_j \quad \text{if} \quad PICP_k \geq PICP_j \ \wedge \ MPIW_k \leq MPIW_j, \tag{19}$$

allowing for the indifference situation $M_k \sim M_j$ iff $PICP_k = PICP_j$ and $MPIW_k = MPIW_j$. The means of criterion (18), or (19), is streamlined: we aim to select a model whose prediction intervals enclose future mortality outcomes and, at the same time, the model (and associated parameters) must not yield an excessive uncertainty. For instance, Fig. 1 shows a graphical visualization of how the proposed preference criterion works. Let us consider five stochastic mortality models and let M_k, $k = 1, \dots, 5$, the coordinates representing such models in the plane $(MPIW, PICP)$. It is straightforward to note that $M_3 \succ M_1 \succ M_2$, but criterion (18) does not allow preference judgment in comparing M_1 and M_4 or M_1 and M_5. However, for financial and actuarial purposes, the latter case should be unambiguous: we are inclined to prefer the model M_1 as it provides a greater probability coverage and is more

Fig. 1 An illustrative representation of mortality models in the plane (*MPIW*, *PICP*)



informative regarding uncertainty. Thus, we interpret the preference criterion (18) as a gold standard, facilitating model selection if there is a prominent mortality model in terms of density forecasts; otherwise, intermediate situations occur, and we need to introduce an additional rule in defining the preference criterion. In particular, we say that $M_k \succ M_j$ if the following criterion is satisfied:

$$\begin{aligned} & PICP_k \geq PICP_j \wedge \ MPIW_k \leq MPIW_j \vee \\ & PICP_k \geq PICP_j \wedge \ MPIW_k \leq MPIW_j \end{aligned} \tag{20}$$

Recalling the example in Fig. 1, the previous criterion implies the preference chain $M_3 \succ M_4 \succ M_1 \succ M_5 \succ M_2$. The model M_2 represents the worst case, since it has simultaneously the lowest PICP and the higest MPIW.

4 Empirical experiments

To test our proposal, we implement an empirical experiment involving mortality experiences worldwide. In particular, we exploit mortality data for both genders in five countries, namely Australia (AUS), England & Wales (GBRTENW), Japan (JPN), Italy (ITA) and the USA. We indicate the population set by $\mathcal{I} = \{\texttt{Male}, \texttt{Female}\} \times \{\texttt{AUS}, \texttt{GBRTENW}, \texttt{JPN}, \texttt{ITA}, \texttt{USA}\}$. We consider such countries representative in terms of mortality dynamics. For instance, Australia, England & Wales and Italy populations experienced a deep, non-linear shaped, reduction of mortality rates after WWII, despite they differ in terms of life expectancy acceleration. On the other hand, Japan population exhibits a linear decline in mortality, while the USA population is characterized by the well-known life expectancy stagnation (Mehta et al. 2020).

Data were downloaded from the Human Mortality Database, which concerns calendar years from 1960 to 2019 and the age set $\mathcal{X} = \{60, \ldots, 89\}$. The latter is specially chosen to investigate ages for which the longevity risk may arise. Such a data are split in two sub-dataset with respect to the forecasting year $t = 2000$: a training dataset composed by mortality data indexed by the calendar years set $\mathcal{T} = \{1960, \ldots, 2000\}$ and a testing dataset concerning mortality data for the calendar years set $\mathcal{T}' = \{2001, \ldots, 2019\}$. The former is used to fit stochastic mortality models, while the latter act as a test for the backtesting analysis. We refer to the Bayesian Information Criterion (hereinafter BIC) to assess the mortality model's in-sample performance. At the same time, the out-of-sample accuracy of point forecasts is evaluated by looking at the Mean Squared Error (hereinafter MSE). We inspect the latter metric as an overall performance measure for each mortality model, and it is defined as:

$$MSE_i = \frac{1}{|\mathcal{T}| \cdot |\mathcal{X}|} \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}} (q_{x,t,i} - \hat{q}_{x,t,i})^2, \quad \forall i \in \mathcal{I}. \tag{21}$$

We also investigated the MSE distribution by age, that is:

$$MSE_{x,i} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} (q_{x,t,i} - \hat{q}_{x,t,i})^2, \quad \forall x \in \mathcal{X}, i \in \mathcal{I}. \tag{22}$$

In addition, we assess the mortality density forecasts using uncertainty metrics explained in Sect. 3. In particular, we firstly scrutinize PICP and MPIW distribution by age and separately. As the final step, we apply the proposed preference criterion displaying the results for ages $x = 65, 75, 85$. The experiments are performed using the R software (R Core Team 2021, version 3.6.3) and the package StMoMo (Villegas et al. 2018, version 0.4.1).

4.1 Results

The present section exposes and discusses the findings of our numerical experiments. Foremost, in Table 3 we list in-sample BIC values for each mortality model and different populations. The best performance is reported in bold for each population. We outline that RH and M7 fitting overperform the other models: the former presents the lowest BIC in 5/10 cases, while the latter has the best BIC in 4/10 cases. The M8 provides the lowest BIC only for the US female population. Thus, from a proper perspective, both periods and cohort effects seem to be prevalent in explaining the mortality surfaces considered. For completeness, we report the heatmaps of the binomial residuals for all the considered models and populations in Appendix B.

Concerning the out-of-sample analysis, Table 4 shows MSE values applying Eq. (21). We observe that models selected in terms of BIC are not optimal from a forecasting

Table 3 In-sample BIC for the different populations and different stochastic models considered

Country	Sex	LC	RH	CBD	M6	M7	M8	APC
AUS	Female	14222.9	**14007.0**	15852.4	14608.2	14755.8	14556.47	14549.3
	Male	14960.0	**14665.4**	15492.7	15238.3	15424.6	15178.7	15061.5
GBRTENW	Female	19915.6	**15955.9**	22196.4	16545.2	16318.8	16360.9	18356.9
	Male	19899.3	**16017.9**	21122.8	16209.1	16106.3	16139.5	16887.6
ITA	Female	17906.7	16271.2	23285.7	17518.1	**15781.2**	16807.03	18043.1
	Male	19805.2	16676.7	20790.5	16141.4	**15949.5**	16103.5	17030.1
JPN	Female	25038.5	16392.7	37239.9	18715.7	**16257.9**	17591.7	17347.1
	Male	25299.3	16429.2	28188.5	17019.0	**15945.7**	16387.1	16918.1
USA	Female	27459.8	21646.5	80536.6	27864.3	26259.9	**25798.9**	29303.3
	Male	30110.8	**21615.3**	42856.1	25973.5	25426.3	25573.3	25675.8

The lowest BIC for each population is reported in bold

Table 4 Out-of-sample MSE_i, $i \in \mathcal{I}$ for the different stochastic models considered

Country	Sex	LC	RH	CBD	M6	M7	M8	APC
AUS	Female	**0.0806**	0.3806	0.1445	1.0595	0.4252	0.1089	0.4720
	Male	0.3128	1.0581	0.3111	**0.2624**	0.2627	0.4614	0.8046
GBRTENW	Female	**0.1720**	0.3545	0.1890	1.1626	0.7468	0.2105	0.2245
	Male	0.6339	1.2497	0.5935	0.4448	**0.3703**	3.5868	0.4801
ITA	Female	0.1259	0.8940	0.2623	0.4258	2.0138	**0.1104**	0.4543
	Male	0.3412	0.4227	0.3747	1.1891	0.4678	**0.3049**	0.5614
JPN	Female	0.0621	0.9355	0.2627	0.4914	1.9322	**0.0239**	0.3407
	Male	0.2171	1.1591	0.8170	0.3437	3.0701	**0.1105**	0.8848
USA	Female	**0.1083**	0.1901	0.1685	5.8476	0.2573	0.2690	0.1235
	Male	0.4543	0.4606	0.5169	3.1440	0.5978	0.6117	**0.3472**

Values are in 10^{-4}

perspective. For instance, the RH forecasts underperform for all populations investigated, and the M7 results are profitable only for the GBRTENW male mortality data. From a forecasting accuracy's point of view, the models appearing fruitful are the LC and the M8. Such evidence remarks on the backtesting process's usefulness. Indeed, a mortality model should offer both an in-sample representativeness and an accurate out-of-sample performance. Our analysis points out potential over-fitting generated by RH and M7 models so that their forecasting accuracy is poor overall. However, we also examine the MSE distribution by age for an in-depth vision.

Referring to Fig. 2, MSE increases with age, and some models show peculiar trends for the elderly. For all female populations, excluding the US, the M7's MSE tends to be greater for initial ages, then decreases until about age 85 and increases for the final ages. Such behaviour also occurs in both Italian and Japanese male populations. For the latter mortality data, the M6's MSE shrinks after age 85, indicating the M6 goodness in predicting mortality for older ages. However, the M6 shows lower accuracy for other mortality experiences. Within the CBD family, the MSE related to the M8 model exhibits greater regularity with increasing age. The RH model seems to be befitting, respecting the age range 60–73, but then loses predictive accuracy at older ages. The LC forecasts produce quite increasing errors, except for US and Australian female populations. On the other hand, the APC's MSE boasts significantly after age 80 for all populations.

Selecting a mortality model considering only the point predictions accuracy could be difficult and, sometimes, misleading. Therefore, we evaluate projected prediction intervals exploiting both PICP and MPIW. We begin inspecting the former as depicted in Fig. 3, separately from the latter represented in Fig. 4. As expected, mortality models within the CBD family demonstrate larger prediction intervals at old ages. These models imply more uncertainty due to the presence of multiple periods effects with respect to LC-based models. Prediction interval width influences the prediction interval coverage, making CBD-based models more appealing ex-ante to anticipate unexpected mortality outcomes. Nonetheless, the ex-post perspective about PICP by age, as in Fig. 3, delineates particular evidence. For instance, the best M6 and M7 density forecasts occur for both English & Welsh and Australian males. For all other populations, such models regain coverage probability only at higher ages, thanks to their MPIW's exponential

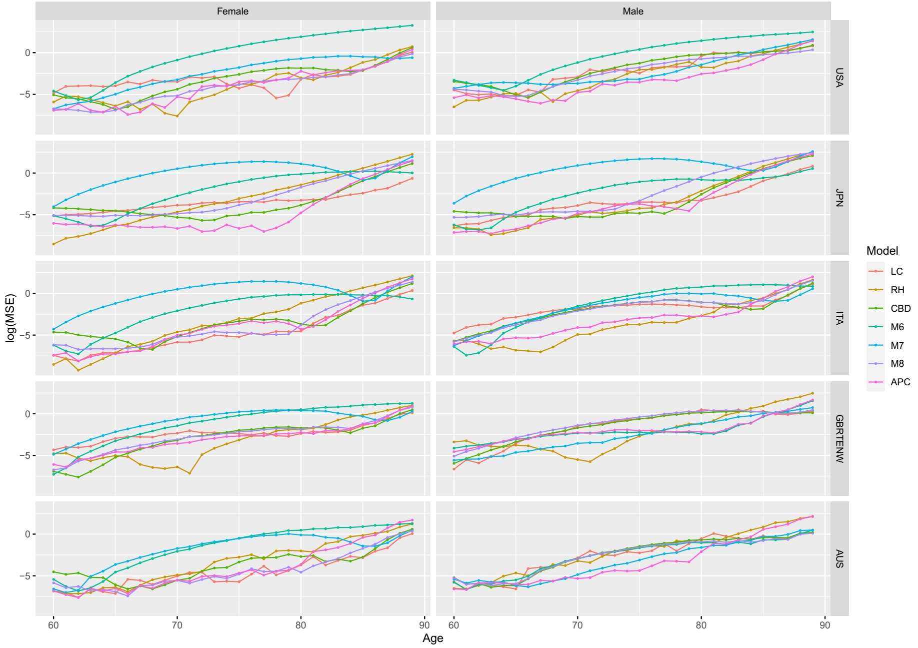


Fig. 2 Out-of-sample $MSE_{x,i}$ distribution by age, for $i \in \mathcal{I}$. $MSE_{x,i}$ values are expressed on log-scale for simplicity of graphical display

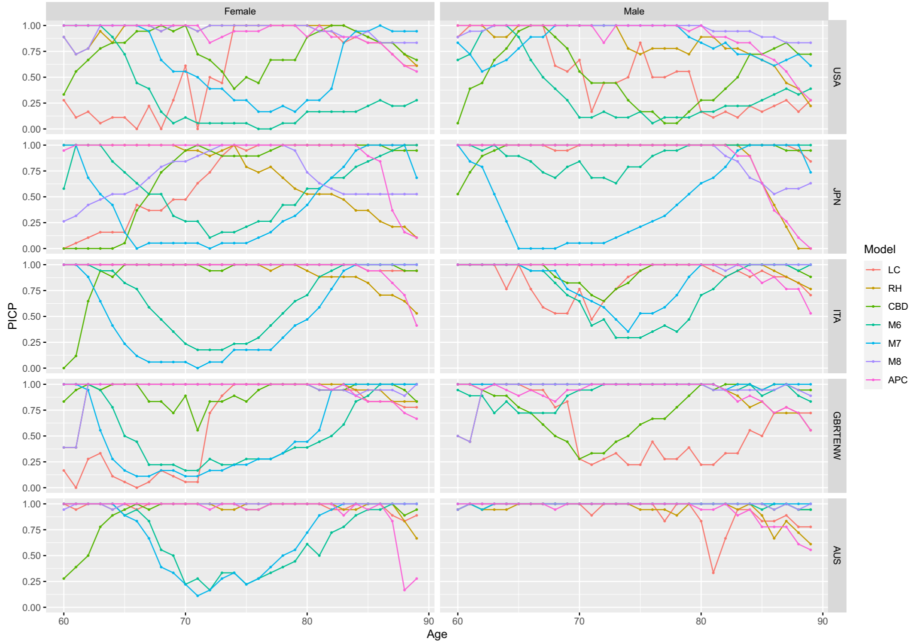


Fig. 3 $PICP_i$ distribution by age, for $i \in \mathcal{I}$. Prediction intervals are calculated at level 95%, accounting for parameter uncertainty

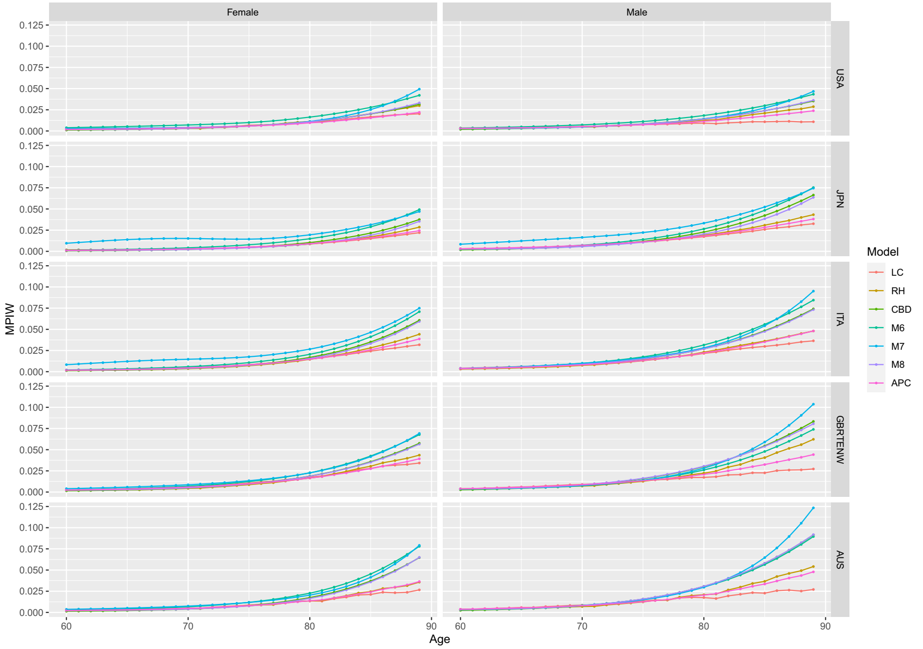


Fig. 4 $MPIW_i$ distribution by age, for $i \in \mathcal{I}$. Prediction intervals are calculated at level 95%, accounting for parameter uncertainty

growth. In addition, we observe that the APC's PICP departs from the complete coverage from age 80 onwards, despite APC's MPIW increasing with age. The LC model presents the lowest MPIW in all populations and allows a reasonable coverage probability for Australian females, Italian females and Japanese males. Regarding PICP, the LC model with cohort effect is profitable for both Australian and English & Welsh female mortality experiences.

Moreover, RH's MPIW increases exponentially at older ages for English & Welsh males, failing to provide a corresponding probability coverage. In our opinion, the M8 model seems to be the best compromise to achieve biological reasonableness and plausibility of predicted levels of uncertainty in forecasts at different ages. To verify our belief, we compare mortality model predictions using a statistical test about the global PICP-based performances.

To this end, we compute the global PICP for each mortality model and each population, that is:

$$PICP_i = \frac{1}{|\mathcal{T}'| \cdot |\mathcal{X}|} \sum_{x \in \mathcal{X}} \sum_{t \in \mathcal{T}'} \mathbb{1}_{\left\{q_{x,t,i} \in [\hat{q}^{LB}_{x,t,i}, \hat{q}^{UB}_{x,t,i}]\right\}}, \tag{23}$$

and we submit the values to the Wilcoxon signed-rank test (Wilcoxon 1992): the null hypothesis assumes that differences between two distributions come from a zero-median distribution, while the alternative hypothesis states that they come from a distribution with a median greater than zero. We perform the test by analysing the PICP values $(PICP_i)_{i \in \mathcal{I}}$ obtained by two models in the (ten) different populations across ages and calendar years.

Table 5 P-value of the left-side Wilcoxon rank test. The test is applied to the $(PICP_i)_{i\in\mathcal{I}}$ values of two models each time

	LC	RH	CBD	M6	M7	M8
RH	0.0000					
CBD	0.0544	0.9998				
M6	0.9999	1.0000	1.0000			
M7	0.9774	1.0000	0.9999	0.0647		
M8	0.0000	0.0037	0.0000	0.0000	0.0000	
APC	0.0000	0.0210	0.0000	0.0000	0.0000	0.7492

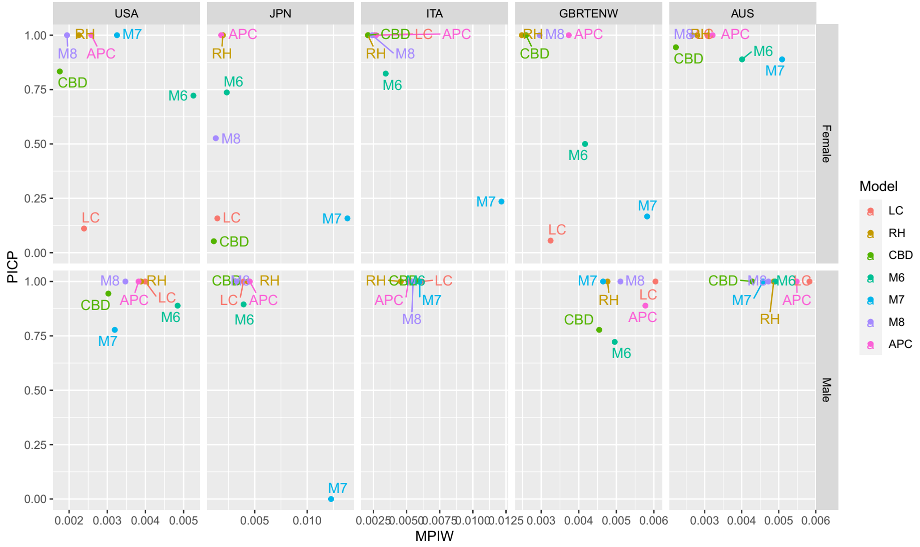


Fig. 5 Stochastic mortality models in the $(MPIW - PICP)$ plane for age $x = 65$

Table 5 contains the Wilcoxon's test results. Each table cell reports the p-value obtained by testing the PICP performances of the model in the row against those of the model in the column. We refer to the value 0.05 as the significance threshold: if the p-value is lower, we reject the null hypothesis concluding that the model in the row is superior to the model in the column. Green cells indicate, by row, which models are better, while red cells identify opposite cases. Interestingly, we observe that for both M8 and APC models alternative hypothesis is rejected in almost all cases. Therefore, these models are superior to others in terms of global PICP. In addition, the empirical evidence does not allow for rejection of the null hypothesis when the APC model is tested against M8 and vice versa, highlighting that these two models have comparable PICP performances.

Finally, we compare mortality forecasts employing jointly PICP and MPIW. In particular, we apply the preference criterion exposed in Eq. (20). Figures 5, 6 and 7 depict the $(MPIW, PICP)$ plane for each population under investigation and for ages 65, 75 and 85 respectively.

For $x = 65$, the M7 model produces modest performances, in particular for both Japanese genders and for the female gender of Italy, England & Wales and Australia. Similar

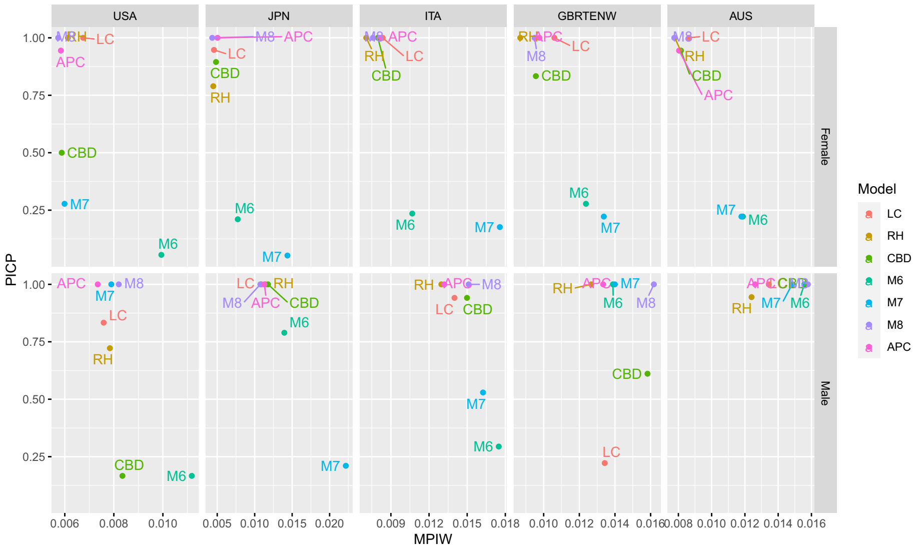


Fig. 6 Stochastic mortality models in the $(MPIW - PICP)$ plane for age $x = 75$

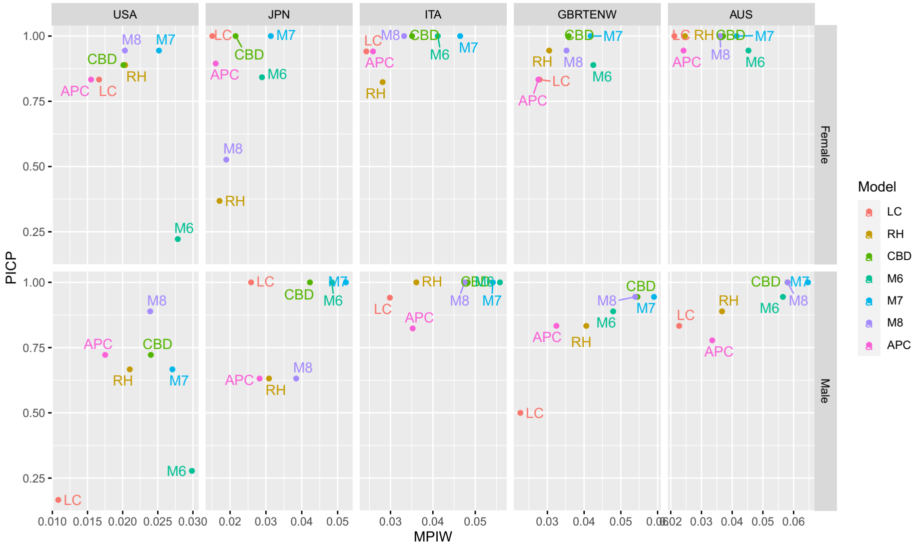


Fig. 7 Stochastic mortality models in the $(MPIW - PICP)$ plane for age $x = 85$

observations hold for the M6 model. In general terms, the M7 is the best model only for English & Wales aged 65. The M8 model is more advisable, but both the RH and the APC models can also guarantee a complete prediction interval coverage without large MPIW values. Moreover, we observe dense clusters of models for many populations, such as US males, Japanese males, Italian males and females, and Australian Males and

Table 6 Best (B) and worst (W) mortality forecasts for ages $x = 65, 75, 85$

		$x = 65$		$x = 75$		$x = 85$	
Country	Gender	(B)	(W)	(B)	(W)	(B)	(W)
AUS	Female	M8	M7	M8	M6	LC	M6
	Male	CBD	LC	APC	RH	CBD	APC
GBRTENW	Female	RH	LC	RH	M7	CBD	LC
	Male	M7	M6	RH	LC	M8	LC
ITA	Female	RH	M6	RH	M7	M8	RH
	Male	RH	LC	RH	M6	RH	APC
JPN	Female	APC	CBD	M8	M7	LC	RH
	Male	M8	M7	M8	M7	LC	M8
USA	Female	M8	LC	M8	M6	M8	M6
	Male	M8	M7	APC	M6	M8	LC

females. In these cases, mortality models provide similar density forecasts for populations aged 65, and the preference criterion's strength emerges when clusters dissolve with increasing age. Indeed, looking at populations aged 75, models M6 and M7 are far from the cluster's centre in many cases. On the other hand, prediction intervals stemming from models M8, RH and APC ensure coherent mortality boundaries with respect to mortality outcomes. Finally, from Fig. 7, some clusters of models disappear, e.g. for US males, and others change their composition. We can appreciate how the LC model becomes more profitable for Japanese populations, as well as for Australian female mortality. However, the M8 model plays the role of "best practice" in most cases.

In Table 6 we summarize models' ranking by the means of criterion (20). Generally speaking, our experiment highlights the M8 model performances are the more balanced. Secondly, also the RH model manifests appealing prediction interval-based results for many populations. From an age-based perspective, both M8 and RH models seem more proper for populations aged $x = 65$ and $x = 75$, while for higher ages, only the M8 maintains good performances. Adding country-specific considerations, both the LC and the CBD models' performances for age $x = 85$ may be explanatory. For instance, we notice that mortality profiles for Japanese populations aged 85 are strongly linear, and the LCs' density forecasts goodness confirms how such a model is more parsimonious than others. Thus, the LC model generates less uncertainty providing narrower prediction intervals and, at the same time, embedding mortality realizations. Similar suggestions hold about Australian populations. However, for populations aged 65, or 75, mortality experiences may be characterized by more complex influences so that mortality models also involving cohort effects are more accurate. Overall, the M6 model looks more unsatisfactory, while the M7 model owns peculiarities. The M7 is tailored for the English & Welsh male population, but it is also competitive for several elderly populations (see, e.g. Fig. 7).

5 Conclusions

The improvements in mortality observed after WWII have raised the need to create stochastic mortality models to measure longevity risk. The literature offers many models' specifications to provide accurate mortality projections. However, what is the best model remains an open question. Indeed, comparative studies available in the literature show that mortality model selection depends on both the mortality experiences and the criteria considered. In addition, few studies have thoroughly tested the uncertainty concerning mortality projections. Forecasting mortality uncertainty levels have a central role since they reveal the potential, unexpected longevity rise and the related economic impact. The present work proposes a methodological framework to backtest uncertainty in mortality projections exploiting uncertainty metrics not yet adopted in mortality literature. To this end, we employ the Prediction Interval Coverage Probability Coverage and the Mean Prediction Interval Width. In particular, such prediction interval-based measures allow quantifying both the plausibility and effectiveness concerning the predicted levels of uncertainty in future mortality outcomes. In addition, we define a new model selection criterion that combines the two metrics, allowing for a plain ex-post assessment of density forecasts at different ages. Numerical experiments are performed in five countries worldwide and both genders. As expected, there does not exist a mortality model overperforming all the others. Despite both age- and country-specific characteristics, our proposal grants to identify models with mortality density forecasts more balanced across populations. The empirical application of our proposal highlights that the RH model seems the best candidate within The LC family, while the M8 model overperforms the other CBD family members. Furthermore, the M7 model may be suited for elderly populations. The latter feature is particularly fruitful for governments, as well as for pension and health benefits providers. These entities bear the cost of increasingly elderly populations, facing the risk of paying out benefits for much longer. Therefore, selecting an appropriate mortality model to forecast mortality is necessary, and our proposal aims to support such a goal. Indeed, our analysis show the possibility to select a country-tailored mortality model in terms of uncertainty.

Future research will proceed in several directions. First, we intend to exploit our criterion to test and compare multi-population models that consider the dependence structure between the mortality dynamics of different countries, see Li and Lee (2005), Kleinow (2015). Second, we plan to develop a data-driven procedure based on machine-learning-based tools to select the most suited mortality model automatically following the approach suggested in Hunt and Blake (2014), Cairns et al. (2019).

Appendix A: Model and parameter uncertainty

In the present work, we construct prediction intervals considering both model and parameter uncertainty. Among others, we refer the reader to Dowd et al. (2010a) for more details concerning the role of both uncertainty causes in backtesting stochastic mortality models. For our purposes, we briefly recall that model uncertainty stems from forecasting errors about the period and cohort indexes, while parameter uncertainty arises from the estimation of the parameters of the GAPC model. Consequently, accounting for both uncertainty sources allows for portraying the randomness generated by a stochastic mortality model in projecting mortality. In particular, model uncertainty is easily assessed as it derives directly from the innovation component of the time series model considered for the time and the cohort indexes, see for instance, Eq. (13) and Table 2. On the other hand, due to the analytical intractability of many stochastic mortality models, parameter uncertainty is usually analyzed using the bootstrap procedure (see, e.g. Brouhns et al. 2005 and Wang and Lu 2005). Therefore, we implement the following procedure to obtain prediction intervals incorporating model and parameter uncertainty:

1. We generate S samples of the number of deaths by sampling from the Binomial distribution as in Eq. (1), that is $D_{x,t}^{(s)} \sim \text{Binomial}\left(E_{x,t}^{0}, q_{x,t}\right)$, $s = 1, \dots, S$;
2. For each bootstrapped sample we re-perform the estimation procedure in Sect. 2.2 obtaining the bootstrapped parameter estimates
$$\boldsymbol{\theta}^{(s)} = \left(\hat{a}_x^{(s)}, \hat{b}_x^{(l),(s)}, \hat{k}_t^{(l),(s)}, \hat{c}_x^{(s)}, \hat{\gamma}_{t-x}^{(s)}\right);$$
3. We collect S bootstrap estimates of the predictor in Eq. (2) and we simulate each of them forward by the following:
$$\log \frac{q_{x,t_n+h}^{(s)}}{1 - q_{x,t_n+h}^{(s)}} = \hat{a}_x^{(s)} + \sum_{l=1}^{L} \hat{b}_x^{(l),(s)} \kappa_{t_n+h}^{(l),(s)} + \hat{c}_x^{(s)} \gamma_{t_n+h-x}^{(s)}.$$
Hence, we obtain simulated trajectories accounting for both the forecast error in the period and cohort indexes and the error in the model fitting;
4. Prediction intervals are finally achieved calculating the empirical α-quantiles.

Concerning numerical experiments exposed in Sect. 4.1, the aforementioned procedure is executed for each population $i \in \mathcal{I}$.

Appendix B: Heatmaps for residuals

See Figs. 8 and 9.

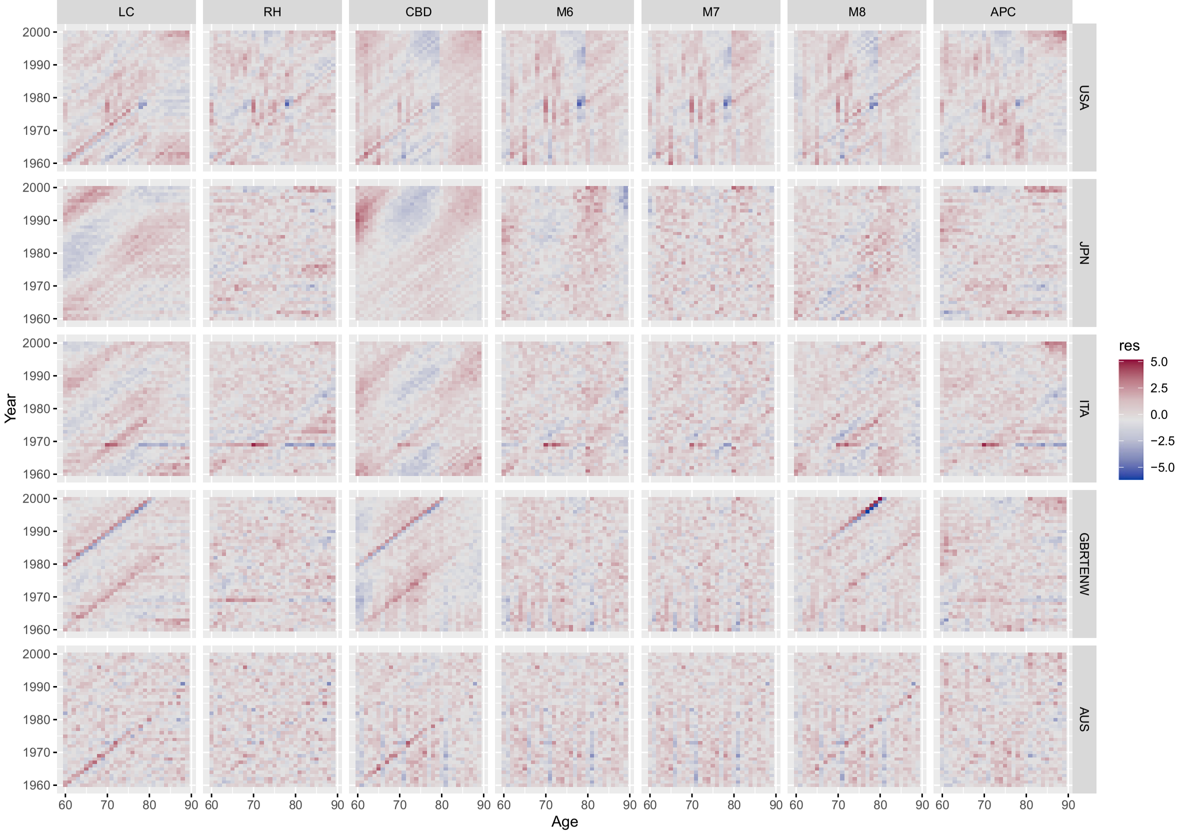


Fig. 8 Heatmaps of the binomial residuals for the different mortality models for all the male populations

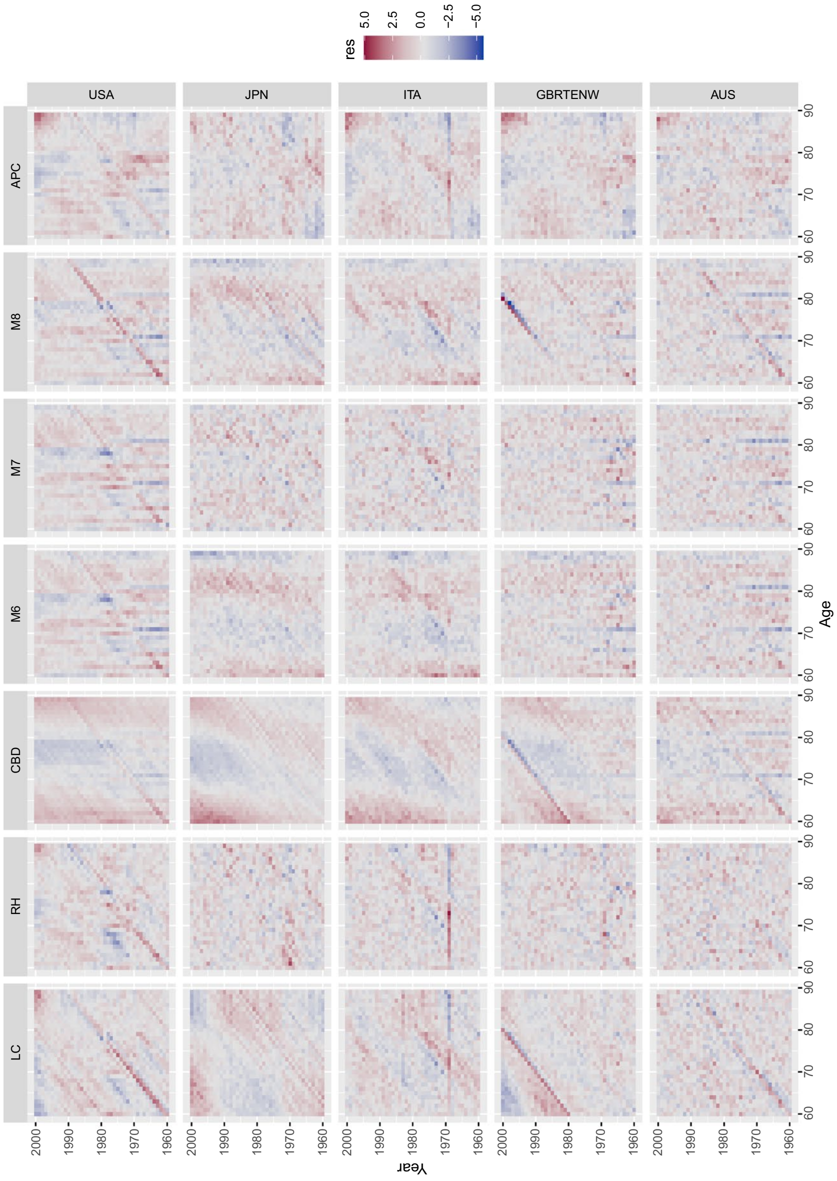


Fig. 9 Heatmaps of the binomial residuals for the different mortality models for all the female populations

Funding Open access funding provided by Università Parthenope di Napoli within the CRUI-CARE Agreement. Mario Marino was supported by the funded project Actuarial and Financial Risk Management Solutions in a Pandemic Mortality Framework (Sapienza University Funding No. RM120172B84AD5EC).

Declarations

Competing interests The authors have not disclosed any competing interests.

References

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

Biffi, P., Clemente, G.P.: Selecting stochastic mortality models for the Italian population. Decis. Econ. Finance **37**, 255–286 (2014)

Bjerre, D.S.: Tree-based machine learning methods for modeling and forecasting mortality. ASTIN Bulletin, 1–23 (2022). https://doi.org/10.1017/asb.2022.11

Booth, H., Maindonald, J., Smith, L.R.: Age–time interactions in mortality projection: applying lee–carter to Australia. Working Papers in Demography, The Australian National University, 25 (2002)

Brouhns, N., Denuit, M., Van Keilegom, I.: Bootstrapping the Poisson log-bilinear model for mortality forecasting. Scand. Actuar. J. **2005**(3), 212–224 (2005)

Brouhns, N., Denuit, M., Vermunt, J.K.: A Poisson log-bilinear regression approach to the construction of projected lifetables. Insur. Math. Econ. **31**(3), 373–393 (2002)

Cairns, A.J., Blake, D., Dowd, K.: A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. J. Risk Insur. **73**(4), 687–718 (2006)

Cairns, A.J., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A., Balevich, I.: A quantitative comparison of stochastic mortality models using data from England and wales and the United States. N. Am. Actuar. J. **13**(1), 1–35 (2009)

Cairns, A.J., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Khalaf-Allah, M.: Mortality density forecasts: an analysis of six stochastic mortality models. Insur. Math. Econ. **48**(3), 355–367 (2011)

Cairns, A.J., Kallestrup-Lamb, M., Rosenskjold, C., Blake, D., Dowd, K.: Modelling socio-economic differences in mortality using a new affluence index. ASTIN Bulletin **49**(3), 555–590 (2019)

Carfora, M., Cutillo, L., Orlando, A.: A quantitative comparison of stochastic mortality models on Italian population data. Comput. Stat. Data Anal. **112**, 198–214 (2017)

Clayton, D., Schifflers, E.: Models for temporal variation in cancer rates. II: age-period-cohort models. Stat. Med. **6**, 469–481 (1987)

Currie, I.D.: Smoothing and forecasting mortality rates with p-splines. Research talk slides (2006)

Currie, I.D.: On fitting generalized linear and non-linear models of mortality. Scand. Actuar. J. **2016**(4), 356–383 (2016)

Czado, C., Delwarde, A., Denuit, M.: Bayesian Poisson log-bilinear mortality projections. Insur. Math. Econ. **36**(3), 260–284 (2005)

Deprez, P., Shevchenko, P.V., Wüthrich, M.V.: Machine learning techniques for mortality modeling. Eur. Actuar. J. **7**, 337–352 (2017)

Dowd, K., Cairns, A.J., Blake, D., Coughlan, G.D., Epstein, D., Khalaf-Allah, M.: Backtesting stochastic mortality models: an ex-post valuation of multi-period-ahead density forecasts. N. Am. Actuar. J. **14**, 281–298 (2010a)

Dowd, K., Cairns, A.J., Blake, D., Coughlan, G.D., Epstein, D., Khalaf-Allah, M.: Evaluating the goodness of fit of stochastic mortality models. Insur. Math. Econ. **47**, 255–265 (2010b)

Haberman, S., Renshaw, H.: A comparative study of parametric mortality projection models. Insur. Math. Econ. **48**, 35–55 (2011)

Human Mortality Database University of California, Berkeley (USA), and Max Planck institute for demographic research (Germany). (2018). www.mortality.org

Hunt, A., Blake, D.: A general procedure for constructing mortality models. N. Am. Actuar. J. **18**, 116–138 (2014)

Keilman, N.: Erroneous population forecasts. In: Bengtsson, T., Keilman, N. (eds.) Old and New Perspectives on Mortality Forecasting. Demographic Research Monographs. Springer, Cham (2019)

Khomenko, S., Cirach, M., Pereira-Barboza, E., Mueller, N., Barrera-Gómez, J., Rojas-Rueda, D., de Hoogh, K., Hoek, G., Nieuwenhuijsen, M.: Premature mortality due to air pollution in European cities: a health impact assessment. Lancet Planet. Health **5**(3), 121–134 (2021)

Khosravi, A., Nahavandi, S., Creighton, D., Atiya, A.F.: Comprehensive review of neural network-based prediction intervals and new advances. IEEE Trans. Neural Netw. **22**(9), 1341–1356 (2011)

Kleinow, T.: A common age effect model for the mortality of multiple populations. Insur. Math. Econ. **63**, 147–152 (2015)

Lee, R.D., Carter, L.R.: Modeling and forecasting US mortality. J. Am. Stat. Assoc. **87**(419), 659–671 (1992)

Levantesi, S., Nigri, A.: A random forest algorithm to improve the Lee–Carter mortality forecasting: impact on q-forward. Soft Comput. **24**, 8553–8567 (2020)

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

Lovász, E.: Analysis of Finnish and Swedish mortality data with stochastic mortality models. Eur. Actuar. J. **1**, 259–289 (2011)

Marino, M., Levantesi, S., Nigri, A.: A neural approach to improve the Lee–Carter mortality density forecasts. N. Am. Actuar. J. (2022). https://doi.org/10.1080/10920277.2022.2050260

Mehta, N.K., Abrams, L.R., Myrskyla, M.: US life expectancy stalls due to cardiovascular disease, not drug deaths. Proc. Natl. Acad. Sci. **117**(13), 6998–7000 (2020)

Nigri, A., Barbi, E., Levantesi, S.: The relay for human longevity: country-specific contributions to the increase of the best-practice life expectancy. Qual. Quant. (2022). https://doi.org/10.1007/s11135-021-01298-1

Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., Perla, F.: A deep learning integrated Lee–Carter model. Risks **7**(1), 33 (2019)

Perla, F., Richman, R., Scognamiglio, S., Wüthrich, M.V.: Time-series forecasting of mortality rates using deep learning. Scand. Actuar. J. **7**, 572–598 (2021)

Plat, R.: On stochastic mortality modeling. Insur. Math. Econ. **45**(3), 393–404 (2009)

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2021)

Renshaw, A.E., Haberman, S.: Lee–Carter mortality forecasting with age-specific enhancement. Insur. Math. Econ. **33**(2), 255–272 (2003)

Renshaw, A.E., Haberman, S.: A cohort-based extension to the Lee–Carter model for mortality reduction factors. Insur. Math. Econ. **38**(3), 556–570 (2006)

Scognamiglio, S.: Calibrating the Lee–Carter and the Poisson Lee–Carter models via neural networks. ASTIN Bull. J. IAA **52**(2), 519–561 (2022)

Vaupel, J.W., Villavicencio, F., Bergeron-Boucher, M.-P.: Demographic perspectives on the rise of longevity. Proc. Natl. Acad. Sci. **118**(9), e2019536118 (2021)

Villegas, A., Kaishev, V.K., Millossovich, P.: StMoMo: an R package for stochastic mortality modelling. J. Stat. Softw. **84**(3), 1–38 (2018)

Wang, D., Lu, P.: Modelling and forecasting mortality distributions in England and Wales using the Lee–Carter model. J. Appl. Stat. **32**(9), 873–885 (2005)

Wilcoxon, F.: Individual comparisons by ranking methods. In: Breakthroughs in Statistics, pp. 196–202. Springer, New York (1992)

Zarulli, V., Sopina, E., Toffolutti, V., Adam, L.: Health care system efficiency and life expectancy: a 140-country study. PLoS ONE **16**(7), e0253450 (2021)