Annals of Operations Research
https://doi.org/10.1007/s10479-023-05808-2

ORIGINAL RESEARCH

Multi-population mortality modelling and forecasting with divergence bounds

Salvatore Scognamiglio[1]

Received: 5 September 2023 / Accepted: 19 December 2023

Abstract
Understanding the mortality dynamics and forecasting its future evolution is crucial for insurance companies and governments facing the risk that individuals might live longer than expected (the so-called *longevity risk*). This paper introduces a neural network model that allows an accurate modelling and forecasting of the mortality rates of many populations. The neural network model we propose is designed to present a fully explainable structure, allowing for understanding how predictions are formulated. Furthermore, the model addresses the problem of measuring and managing the divergence of the long-term forecasts of the mortality rates arising when one decides to model the mortality of two or more populations simultaneously. Indeed, for many models available in the literature, this divergence grows over time, resulting in an ever-increasing trend in the gap in life expectancy among countries that appear unrealistic and biologically unreasonable. The proposed model allows the construction of analytical bounds for this divergence and illustrates that these bounds can be exploited to analyse and measure the dissimilarities between two or more populations and identify opportunities for longevity risk diversification. Numerical experiments performed using all the data from the Human Mortality Database data show that our model produces more accurate mortality forecasts with respect to some well-known stochastic mortality models and allows us to obtain valuable insights about the mortality pattern of the population considered.

Keywords Multi-population Neural mortality modelling · Neural networks · Coherence mortality forecasting · Human Mortality Database

1 Introduction

The mortality rates of most countries worldwide are considerably decreasing in the last few decades, mainly due to the improvements in medical technologies, hygiene, lifestyle changes, and government regulation (Oeppen & Vaupel, 2002). Although these longevity improvements are generally perceived as a benefit for society, they can also represent a risk for insurance companies issuing life annuities and other products providing longevity

✉ Salvatore Scognamiglio
salvatore.scognamiglio@uniparthenope.it

[1] Department of Management and Quantitative Science, University of Naples - Parthenope, Via Generale Parisi 13, Naples 80133, Italy

insurance, and governments with social security pension obligations (Barrieu et al., 2012; Devolder et al., 2021). Numerous stochastic mortality models have been developed in recent years for capturing the longevity improvements. Among them, the Lee-Carter (LC) model proposed in Lee and Carter (1992) is the most popular one. It is an extrapolative method that describes the logarithm of the mortality rate as the sum of an age-specific base level and the product of a time-varying index (period effect) and an age-modulating parameter (age effect). This seminar paper estimates the parameters using the Singular Value Decomposition for the model calibration. Projections of the future mortality rates are obtained by keeping constant the age-specific parameters and extrapolating future values of the time index using ARIMA models. Many variants and alternatives to the LC model have appeared in the literature. Brouhns et al. (2002) embeds the LC model in a Poisson regression framework and suggests a maximum likelihood estimator to overcome the heteroskedasticity issues associated with the OLS estimator. Other authors identified a cohort effect in mortality data and recommended enhancing the LC model by introducing an additional term to effectively capture and model this phenomenon, as proposed in Renshaw and Haberman (2006). This effect pertains to the impact of shared experiences or characteristics among individuals born in the same year. However, introducing this term improves the forecasting performance only for some populations. Koissi and Shapiro (2006) introduces an extension based on the fuzzy logic that allows for the representation of uncertainty and imprecision in the observed data. We refer to (Basellini et al., 2022) for a detailed and recent review of the topic. One of the main criticisms of the LC model concerns the biological reasonableness of mortality forecasts when applied to multiple populations simultaneously. Tuljapurkar et al. (2000) shows that applying the LC method separately to the G7 countries produces divergent forecasts that highlight an ever-increasing trend in the gap in life expectancy among these countries that appear unrealistic. In this regard, Li and Lee (2005) proposes a multi-population extension of the LC model that ensures long-term not-divergent forecasts among the populations. This property is generally called *coherence* of the mortality forecasts, and it has been investigated in the literature, see (Hyndman et al., 2013; Shi, 2023). However, this assumption could be suitable only for specific groups of populations and over limited time windows and is generally perceived as too restrictive when one needs to model jointly the mortality of a large number of populations with different mortality patterns.

Recently, neural networks (NN) have been successfully applied to numerous fields, such as computer vision (Madhav et al., 2023), natural language processing (Ahmed et al., 2022), and operation research (Gupta et al., 2022). They have obtained notable results also in the context of mortality modelling and forecasting. Their ability to analyse large amounts of data and learn complex functional relationships make NNs promising tools, especially when a large number of populations are considered simultaneously with complex mortality patterns. Hainaut (2018) were pioneers in utilising neural networks for mortality forecasting. Since then, the literature on this topic has expanded rapidly. Subsequent studies such as Nigri et al. (2019, 2021) and Lindholm and Palmborg (2022) have further explored and refined the application of NN in the single population framework. Some years later, Richman and Wüthrich (2021), Perla et al. (2021), Scognamiglio (2022), Schnürch and Korn (2022) have extended the approach by to encompassing a broader array of populations with notable results.

While all these contributions show that neural networks generally outperform the other statistical models in the forecasting tasks, little attention has be paid to the divergence of the mortality forecasts and its measurement.

The contribution of this paper to the mortality modelling field is two-fold. First, we use NN to improve the model flexibility and produce accurate mortality forecasts for a large number of populations. Furthermore, the model allows to derive some bounds that measure

and control the divergence of the mortality forecasts among the population considered. We show that these bounds can also be used to measure the similarities of the mortality dynamics of different countries. In this way, we can identify longevity risk diversification opportunities. The model also presents an intuitive interpretation and allows to explain how the NN computes the predictions. This aspect is relevant to encourage the application of NNs in actuarial applications.

The agenda of the paper is the following: Sect. 2 two of the most popular mortality models, Sect. 3 describes the neural network building blocks, Sect. 4 presents the proposed mortality model and the NN architecture used for its calibration, Sect. 5 illustrates some numerical experiments, and Sect. 6 concludes.

2 Stochastic mortality models

Let $\mathcal{X} = \{x_0, x_1, \ldots, x_\omega\}$ be the set of (integer) ages, $\mathcal{T} = \{t_0, t_1, \ldots, t_n\}$ be the set of calendar years, and $\mathcal{I} = \{\texttt{pop}_1, \texttt{pop}_2, \ldots, \texttt{pop}_\texttt{I}\}$ be the set of the populations considered. $m_{x,t}^{(i)}$ indicates the mortality rate at age $x \in \mathcal{X}$ at time $t \in \mathcal{T}$ related to the population $i \in \mathcal{I}$. We discuss the models also in terms of coherence of the mortality forecasts. Coherence means that the ratio of the long-term predictions formulated at time t_0 of the mortality rates of two different populations maintains a constant ratio over time:

$$\lim_{H \to +\infty} \frac{m_{x,t_0+H}^{(i_1)}}{m_{x,t_0+H}^{(i_2)}} = c, \qquad c \in \mathbb{R}$$

for each age $x \in \mathcal{X}$ and two populations $i_1, i_2 \in \mathcal{I}$.

2.1 No-coherence: independent Lee-Carter models

The simplest approach for modelling the mortality of multiple populations consists of applying independent single-population mortality models to each populations. In the case of the LC model, the logarithm of the central death rate $\log m_{x,t}^{(i)} \in \mathbb{R}$ is defined as

$$\log m_{x,t}^{(i)} = a_x^{(i)} + b_x^{(i)} k_t^{(i)} + e_{x,t}^{(i)}, \qquad with\ i.i.d.,\ e_{x,t}^{(i)} \sim N\left(0, \left(\sigma_e^{(i)}\right)^2\right)$$

where $a_x^{(i)} \in \mathbb{R}$ is the average age- and population-specific pattern of mortality, $b_x^{(i)} \in \mathbb{R}$ represents the age- and population-specific sensitivity of the logarithm of the force of mortality at age x to variations in the time index $k_t^{(i)}$; $k_t^{(i)} \in \mathbb{R}$ is the population-specific time index summarising mortality trend and $e_{x,t}^{(i)} \in \mathbb{R}$ is the error term. The application of the LC model requires a two-step procedure. First, the parameters are estimated according to the OLS estimator by solving the optimisation problem:

$$\underset{(a_x^{(i)})_x, (b_x^{(i)})_x, (k_t^{(i)})_t}{\arg\min} \sum_{x \in \mathcal{X}} \sum_{t \in \mathcal{T}} \left(\log m_{x,t}^{(i)} - a_x^{(i)} - b_x^{(i)} k_t^{(i)} \right)^2 \qquad \forall i \in \mathcal{I}.$$

The authors propose a procedure based on the Singular Value Decomposition and suggest to apply the following contraints to avoid identifiability issues:

$$\sum_{x \in \mathcal{X}} b_x^{(i)} = 1 \qquad \sum_{t \in \mathcal{T}} k_t^{(i)} = 0.$$

Second, to derive the forecasts of the mortality rates in a future year $t_n + H$, $H \in \mathbb{N}$, the LC model assumes that the parameters $(a_x^{(i)})_x$ and $(b_x^{(i)})_x$ are constant over time, while the time indices $k_t^{(i)}$ are modelled as Random Walk with Drift (RWD):

$$k_t^{(i)} = k_{t-1}^{(i)} + \theta^{(i)} + \epsilon_t^{(i)} \qquad with\ i.i.d,\ \epsilon_t^{(i)} \sim N\left(0, \left(\sigma_\epsilon^{(i)}\right)^2\right)$$

where $\theta^{(i)} \in \mathbb{R}$ is the population-specific drift term. Once the model parameters are estimated, mortality forecasts are obtained as:

$$\log \hat{m}_{x,t_n+H}^{(i)} = \hat{a}_x^{(i)} + \hat{b}_x^{(i)} \left(\hat{k}_{t_n}^{(i)} + H\hat{\theta}^{(i)}\right).$$

The difference at time $t_n + H$ between the mortality rates at age x, i related to two populations $i_1, i_2 \in \mathcal{I}$ is:

$$\begin{aligned} \log \frac{m_{x,t_n+H}^{(i_1)}}{m_{x,t_n+H}^{(i_2)}} &= \left[a_x^{(i_1)} + b_x^{(i_1)} \left(k_{t_n}^{(i_1)} + H\theta^{(i_1)}\right)\right] - \left[a_x^{(i_2)} + b_x^{(i_2)} \left(k_{t_n}^{(i_2)} + H\theta^{(i_2)}\right)\right], \\ &= \left[a_x^{(i_1)} - a_x^{(i_2)}\right] + \left[b_x^{(i_1)} k_{t_n}^{(i_1)} - b_x^{(i_2)} k_{t_n}^{(i_2)}\right] + H\left[b_x^{(i_1)}\theta^{(i_1)} - b_x^{(i_2)}\theta^{(i_2)}\right]. \end{aligned}$$

We note that the divergence can be decomposed in two terms: the term $\left[a_{x,i_1} - a_{x,i_2}\right] + \left[b_{x,i_1} k_{t_n,i_1} - b_{x,i_2} k_{t_n,i_2}\right]$ is fixed, while the quantity $H\left[b_{x,i_1}\theta_{i_1} - b_{x,i_2}\theta_{i_2}\right]$ depends on H. Then, this divergence increases indefintively over time and this result persists also when we consider the male and female populations living in the same countries are considered.

2.2 Full coherence: the Li-Lee model

Li and Lee (LL) (Li & Lee, 2005) proposed a multi-population extension of the LC model where all the populations share the same age and period effects ($b_x^{(i)} = B_x \in \mathbb{R}$ and $k_t^{(i)} = K_t \in \mathbb{R}, \forall i \in \mathcal{I}$):

$$\log m_{x,t}^{(i)} = a_x^{(i)} + B_x K_t + \nu_{x,t}^{(i)}, \qquad with\ i.i.d\ \nu_{x,t}^{(i)} \sim N\left(0, \left(\sigma_\nu^{(i)}\right)^2\right). \tag{1}$$

In this case, $a_x^{(i)}$ are estimated for each population individually, while B_x and K_t are obtained by applying the ordinary LC method to the aggregate mortality rates of the whole group. Also in this case, a two-step procedure is required to obtain forecasts, and the authors assume that the dynamics of K_t is described with a RWD. Despite many other multi-population mortality models are available in the literature (Dong et al., 2020; Schnürch et al., 2021; Cardillo et al., 2022), we focus on the LL model since it requires the coherence of the mortality forecasts. The authors also provide an extension of the model that includes an additional bilinear term with population-specific parameters. In that case, the full model reads:

$$\log m_{x,t}^{(i)} = \alpha_x^{(i)} + B_x K_t + b_x^{(i)} k_t^{(i)} + \zeta_{x,t}^{(i)}, \qquad with\ i.i.d\ \zeta_{x,t}^{(i)} \sim N\left(0, \left(\sigma_\zeta^{(i)}\right)^2\right), \tag{2}$$

where the parameters $b_x^{(i)}$ and $k_t^{(i)}$ are estimated by performing the first-order SVD to the residuals matrix of the model for the different populations. The additional time-components $k_t^{(i)}$ is a stationary processes modelled with a first-order autoregressive $AR(1)$ model:

$$k_t^{(i)} = \psi_0^{(i)} + \psi_1^{(i)} k_{t-1}^{(i)} + o_t^{(i)} \qquad with\ i.i.d\ o_t^{(i)} \sim N\left(0, \left(\sigma_o^{(i)}\right)^2\right),$$

where $\psi_0^{(i)}, \psi_1^{(i)} \in \mathbb{R}, \forall i \in \mathcal{I}$ are the population-specific coefficients. We use the abbreviation LL to refer the model in (2). In the case of the LL model, the difference at time $t_n + H$ between the mortality rates at age x, related to two populations $i_1, i_2 \in \mathcal{I}$ has the following asymptotic behavior:

$$\lim_{H\to\infty} \log \frac{m_{x,t_n+H}^{(i_1)}}{m_{x,t_n+H}^{(i_2)}} = \left[a_x^{(i_1)} - a_x^{(i_2)}\right] + \left[b_x^{(i_1)}\psi_0^{(i_1)} - b_x^{(i_2)}\psi_0^{(i_2)}\right]$$

that doesn't depends on H.

3 Neural networks

NNs consist of interconnected computational units, called neurons in the NN jargon, arranged on different layers that learn from data using training algorithms. To illustrate the mechanism behind the NNs, let $\boldsymbol{u} = \left(u_1, u_2, \ldots, u_{q_0}\right) \in \mathbb{R}^{q_0}$ denote the vector of the input features. A Fully-Connected Network (FCN) layer with $q_1 \in \mathbb{N}$ units is a function that maps $\boldsymbol{u}$ to in a q_1-dimensional real-valued space:

$$\boldsymbol{z}^{(1)} : \mathbb{R}^{q_0} \to \mathbb{R}^{q_1}, \qquad \boldsymbol{u} \mapsto \boldsymbol{z}^{(1)}(\boldsymbol{u}) = \left(z_1^{(1)}(\boldsymbol{u}), z_2^{(1)}(\boldsymbol{u}), \ldots, z_{q_1}^{(1)}(\boldsymbol{u})\right)'.$$

The output of each unit can be interpreted as a new feature $z_j^{(1)}(\boldsymbol{u})$ depending in a non-linear fashion on $\boldsymbol{u}$:

$$z_j^{(1)}(\boldsymbol{u}) = \phi\left(w_{j,0}^{(1)} + \sum_{l=1}^{q_0} w_{j,l}^{(1)} x_l\right) \qquad j = 1, 2, ..., q_1,$$

where $\phi : \mathbb{R} \mapsto \mathbb{R}$, and $w_{j,l}^{(1)} \in \mathbb{R}$ represent the network weights. We also denote as $W^{(1)} = (w_{j,l}^{(1)})_{1\le j\le q_1, 0\le l\le q_0} \in \mathbb{R}^{q_1\times q_0}$ the full matrix of weights related to this layer[1].

Shallow neural networks present a single hidden layer and directly use the features for computing the quantity of interest $y \in \mathcal{Y}$. In the case of $\mathcal{Y} \subseteq \mathbb{R}$, the output of shallow NN is determined as:

$$y = g^{-1}\left(\beta_0 + \sum_{k=0}^{q_1} \beta_k z_k^{(1)}(\boldsymbol{u})\right)$$

where $\beta_k \in \mathbb{R}$, $(k = 0, \ldots, q_1)$ are the coefficients of the output layer, and $g : \mathbb{R} \to \mathbb{R}$ is a fixed strictly monotone and smooth link function.

If the network is deep, the vector $\boldsymbol{z}^{(1)}(\boldsymbol{u})$ is utilised as input in the subsequent layer to compute new features, and this holds true for the subsequent layers as well. Let $h \in \mathbb{N}$ be the number of hidden layers (depth of network), and $q_k \in \mathbb{N}$, for $1 \le k \le h$, be a sequence of integers that indicates the dimension of each FCN layer (widths of layers). A deep FCN can be described as follows:

$$\boldsymbol{u} \mapsto \boldsymbol{z}^{(h:1)}(\boldsymbol{u}) = \left(\boldsymbol{z}^{(h)} \circ \cdots \circ \boldsymbol{z}^{(1)}\right)(\boldsymbol{u}) \in \mathbb{R}^{q_h},$$

where the functions $\boldsymbol{z}^{(k)} : \mathbb{R}^{q_{k-1}} \to \mathbb{R}^{q_k}$ share the same structure Although they present different network weights $W^{(k)} \in \mathbb{R}^{q_k\times q_{k-1}}$, for $1 \le k \le h$. In the case of deep NN, the output layer uses the features extracted by the last hidden layer $\boldsymbol{z}^{(h:1)}(\boldsymbol{u})$ instead of those $\boldsymbol{z}^{(1)}(\boldsymbol{u})$. The NN coefficients have to be calibrated according to a chosen loss function. The gradient

[1] The superscript (1) underscores our reference to the first hidden layer.

descent algorithm or any of its extension is generally used for finding an approximation of the solution. See (Goodfellow et al., 2016) for more details.

Mortality data typically include categorical features, such as the country or gender of the populations under consideration. One-hot encoding and dummy encoding procedures are standard for dealing with categorical data. However, in the case of which the data presents many categorical features, or one of them presents a high cardinality, these coding schemes produce high-dimensional sparse vectors, which often leads to computational issues. Categorical embedding layers represent an interesting alternative to these coding schemes. They have been introduced in the contexts of Natural Language Processing; see for example (Bengio et al., 2003).

An Embedding Network (EN) layer learns a low-dimensional representation of the levels of a categorical variable. Let $q_{\mathcal{L}} \in \mathbb{N}$ denote the hyperparameter that defines the size of the embedding. The levels of the categorical variable are mapped into a real-valued $\mathbb{R}^{q_{\mathcal{L}}}$-dimensional space. The coordinates of the level in the new space are parameters of the NN that have to be trained (Guo & Berkhahn, 2016). The distance of the levels in the new learned space reflects the similarity of levels concerning the target variable: similar levels will have a small euclidean distance, whereas very different categories will have a large one.

Formally, let $\mathcal{L} = \{l_1, l_2, \ldots, l_{n_{\mathcal{L}}}\}$ be the set of categories of the qualitative variable and $n_{\mathcal{L}}$ be its cardinality. An embedding layer is a mapping

$$z_{\mathcal{L}} : \mathcal{L} \to \mathbb{R}^{q_{\mathcal{L}}}.$$

The number of embedding weights that must be learned during training is $n_{\mathcal{L}} q_{\mathcal{L}}$, and the embedding size typically satisfies $q_{\mathcal{L}} \ll n_{\mathcal{L}}$.

The network's performance hinges on the calibration of weights across different layers, denoted as $w_{l,j}^{(k)}$. In the context of FCN layers, these weights take the form of matrices $W^{(k)}$ and a bias term $\boldsymbol{w}_0^{(k)}$ for $k = 1, \ldots, h$. Meanwhile, for EN layers, the weights are represented by the coordinates of levels in the new embedding space, denoted as $z_l(l)$ for all $l \in \mathcal{L}$. The training process involves unconstrained optimisation, wherein a suitable loss function $L(w_{l,j}^{(k)}, \cdot)$ is chosen, and its minimum is sought. NN training employs the Back-Propagation algorithm, updating weights based on the gradient of the loss function. The goal is to iteratively adjust the weights to minimize errors between the network outputs and reference values. The complexity of training escalates with the increasing number of layers and units per layer in the network architecture. For an in-depth exploration of neural networks and back-propagation, we refer to (Goodfellow et al., 2016).

4 A neural network mortality model with divergence bounds

A fully coherent modelling of the mortality rates prevents diverging long-term forecasts, which do not seem biologically reasonable. However, the constraints of full coherence may be considered overly stringent and are not always substantiated by empirical observations. The issue is especially relevant when populations with very different mortality patterns are modelled simultaneously. In this regard, we introduce a novel NN-based mortality model that relaxes this constraint, allowing divergences of the forecasts within some bounds analytically measured. We call this model as Multi-Population Neural Network model with Divegence Bounds (MPNNDB).

4.1 Mortality model formulation

Let $\mathcal{R} = \{\texttt{reg}_1, \ldots \texttt{reg}_\texttt{R}\}$ be the set of geographical regions, and $\mathcal{G} = \{\texttt{male}, \texttt{female}\}$ be the set of sexes. We consider a set of populations that differ among them for the region and sexes, such that $\mathcal{I} = \mathcal{R} \times \mathcal{G}$. We assume the existence of a small set of latent factors $\boldsymbol{\kappa}_t = (\kappa_t^{(1)}, \kappa_t^{(2)}, \ldots, \kappa_t^{(S)})' \in \mathbb{R}^S$, $S \ll |\mathcal{I}|$[2] describing the mortality dynamics of all the populations considered. The idea consists of expressing the time index driving the mortality of each country as a convex combination of the latent factor $\boldsymbol{k}_t$ with coefficients depending on the country r. These coefficients, denoted as $\boldsymbol{\gamma}^{(r)} = (\gamma_1^{(r)}, \gamma_2^{(r)}, \ldots, \gamma_S^{(r)}) \in (0, 1)^S$, represent points lying in a standard simplex. In such a case, the time-index can be expressed as:

$$k_t^{(i)} = \gamma_1^{(r)}\kappa_t^{(1)} + \gamma_2^{(r)}\kappa_t^{(2)} + \cdots + \gamma_S^{(r)}\kappa_t^{(S)} = \left\langle \boldsymbol{\gamma}^{(r)}, \boldsymbol{\kappa}_t \right\rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product in $\mathbb{R}^S$, and the time-index is the same for the male and female populations living in the same country ($k_t^{(i)} = k_t^{(r,\texttt{male})} = k_t^{(r,\texttt{female})}$). We then define the following model:

$$\log m_{x,t}^{(i)} = a_x^{(i)} + B_x \sum_{s=1}^{S} \gamma_l^{(r)} \kappa_t^{(l)} + \xi_{x,t}^{(i)} \qquad with\ i.i.d\ \xi_{x,t}^{(i)} \sim N(0, (\sigma_\xi^{(i)})^2).$$

where $a_x^{(i)}$ is the age and population-specific level of mortality; B_x is a global age-specific parameter expressing the sensitivity of the mortality rates at each age to variations of the time-index; $\kappa_t^{(s)}, s = 1, \ldots, S$ are time-indexes driving the mortality evolution of all the considered populations; $\gamma_s^{(r)}, s = 1, \ldots, S$ are coefficients representing the contribution of each time-index to the region-specific time-index. The proposed model can be seen as an advancement of the model proposed in Perla and Scognamiglio (2023). In that model, the mortality dynamics of each region were hooked to a single latent factor. On the opposite, in the proposed model, we assume that the mortality of each region depends on all the latent factors through the coefficients $\gamma_s^{(r)}$. A higher value of this coefficient indicates that the s-th latent factor contributes more than the others to drive the mortality of the r-th region.

The MPNNDB model also admits the following compact form:

$$\log \mathbf{m}_t^{(i)} = \mathbf{a}^{(i)} + \boldsymbol{B} \cdot \left\langle \boldsymbol{\gamma}^{(i)}, \boldsymbol{\kappa}_t \right\rangle + \boldsymbol{\xi}_t^{(i)}, \qquad \forall i \in \mathcal{I}, \tag{3}$$

where $\log \mathbf{m}_t^{(i)} = \{\log m_{x,t}^{(i)}\}_{x \in \mathcal{X}}$, $\mathbf{a}^{(i)} = \{a_x^{(i)}\}_{x \in \mathcal{X}}$, $\boldsymbol{B} = \{B_x\}_{x \in \mathcal{X}}$. The parameters in (3) can be expressed as functions of the data and can be replaced by some NNs specifically designed to mimic the model parameters. Denoting as $M_t = \{\log \boldsymbol{m}_t^{(i)}\}_{i \in \mathcal{I}} \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{I}|}$ the matrix containing the mortality rates for all the ages and population considered at time t. We define the following functions:

$$\begin{aligned} \boldsymbol{\alpha}_{W^{(\alpha)}}(i) &: \mathcal{I} \to \mathbb{R}^{|\mathcal{X}|} \\ \boldsymbol{\gamma}_{W^{(\gamma)}}(i) &: \mathcal{I} \to [0, 1]^S \\ \boldsymbol{\kappa}_{W^{(\kappa)}}(M_t) &: \mathbb{R}^{|\mathcal{X}| \times |\mathcal{I}|} \to \mathbb{R}^S \end{aligned} \tag{4}$$

which depend on some NN pararameters, respectively denoted as $W^{(\alpha)}$, $W^{(\gamma)}$ and $W^{(\kappa)}$. Replacing the mortality model parameters with these functions, the model can be written as:

[2] We employ the notation $|\mathcal{I}|$ to represent the cardinality of the set $\mathcal{I}$. Throughout the remainder of the paper, we will consistently use this notation for other sets as well.

$$\log \mathbf{m}_t^{(i)} = \boldsymbol{\alpha}_{W^{(\alpha)}}(i) + \boldsymbol{B} \cdot \left\langle \boldsymbol{\gamma}_{W^{(\gamma)}}(i); \boldsymbol{\kappa}_{W^{(\kappa)}}(M_t) \right\rangle + \boldsymbol{\xi}_t^{(i)}, \qquad \forall i \in \mathcal{I}.$$

The OLS estimation of the model involves training all three networks simultaneously by utilising the Mean Squared Error (MSE) as the loss function. Let $W = (W^{(\alpha)}, W^{(\gamma)}, W^{(\kappa)}, \boldsymbol{B})$ be the full set of network parameters, the calibration of the network arises the following optimisation:

$$\arg\min_{W} \sum_{i\in\mathcal{I}} \sum_{x\in\mathcal{X}} \sum_{t\in\mathcal{T}} \left(\log m_{x,t}^{(i)} - \alpha_{W^{(\alpha)},x}(i) - B_x \cdot \left\langle \boldsymbol{\gamma}_{W^{(\gamma)}}(i); \boldsymbol{\kappa}_{W^{(\kappa)}}(M_t) \right\rangle \right)^2.$$

4.2 Network architecture

$\boldsymbol{\alpha}_{W^{(\alpha)}}(i)$ depends on two categorical features, and we learn it by using an NN consisting of two embedding layers and a fully-connected layer with size equal to the number of age categories considered ($|\mathcal{X}|$). More specifically, the two embeddings, denoted by $\boldsymbol{z}_{\mathcal{R}}^{(\alpha)}$ and $\boldsymbol{z}_{\mathcal{G}}^{(\alpha)}$, process region and gender respectively. Then, an FCN layer, indicated as $\boldsymbol{f}^{(\alpha)}$, is applied to their output. Let $q_{\mathcal{R}}^{(\alpha)}, q_{\mathcal{G}}^{(\alpha)} \in \mathbb{N}$ be hyper-parameters defining the size of the two EN layers, which can be formalised as two mappings:

$$\begin{aligned} \boldsymbol{z}_{\mathcal{R}}^{(\alpha)} &: \mathcal{R} \to \mathbb{R}^{q_{\mathcal{R}}^{(\alpha)}}, \qquad r \mapsto \boldsymbol{z}_{\mathcal{R}}^{(\alpha)}(r) = \left(z_{\mathcal{R},1}^{(\alpha)}(r), z_{\mathcal{R},2}^{(\alpha)}(r), \ldots, z_{\mathcal{R},q_{\mathcal{R}}^{(\alpha)}}^{(\alpha)}(r) \right)^\top, \\ \boldsymbol{z}_{\mathcal{G}}^{(\alpha)} &: \mathcal{G} \to \mathbb{R}^{q_{\mathcal{G}}^{(\alpha)}}, \qquad g \mapsto \boldsymbol{z}_{\mathcal{G}}^{(\alpha)}(g) = \left(z_{\mathcal{G},1}^{(\alpha)}(g), z_{\mathcal{G},2}^{(\alpha)}(g), \ldots, z_{\mathcal{G},q_{\mathcal{G}}}^{(\alpha)}(g) \right)^\top. \end{aligned} \tag{5}$$

Denoting as $\boldsymbol{z}_{\mathcal{I}}^{(\alpha)} = \boldsymbol{z}_{\mathcal{I}}^{(\alpha)}(r,g) = \left(\left(\boldsymbol{z}_{\mathcal{R}}^{(\alpha)}(r)\right)^\top, \left(\boldsymbol{z}_{\mathcal{G}}^{(\alpha)}(g)\right)^\top \right)^\top \in \mathbb{R}^{q_{\mathcal{I}}^{(\alpha)}}$, the vector obtained concatenating the output of these two EN layers, it is further processed by a FCN layer which provides as many units as the age considered. This layer maps $\boldsymbol{z}_{\mathcal{I}}$ in a new $|\mathcal{X}|$-dimensional real-valued space

$$\boldsymbol{f}^{(\alpha)} : \mathbb{R}^{q_{\mathcal{I}}^{(\alpha)}} \to \mathbb{R}^{|\mathcal{X}|}, \qquad \boldsymbol{z}_{\mathcal{I}} \mapsto \boldsymbol{f}\left(\boldsymbol{z}_{\mathcal{I}}^{(\alpha)}\right) = \left(f_{x_0}^{(\alpha)}\left(\boldsymbol{z}_{\mathcal{I}}^{(\alpha)}\right), f_{x_1}^{(\alpha)}\left(\boldsymbol{z}_{\mathcal{I}}^{(\alpha)}\right), \ldots, f_{x_\omega}^{(\alpha)}\left(\boldsymbol{z}_{\mathcal{I}}^{(\alpha)}\right) \right)^\top.$$

Each new feature $f_x^{(\alpha)}(\boldsymbol{z}_{\mathcal{I}}^{(\alpha)})$ is a age-specific function of the vector $\boldsymbol{z}_{\mathcal{I}}^{(\alpha)}$

$$\boldsymbol{z}_{\mathcal{I}}^{(\alpha)} \;\mapsto\; f_x^{(\alpha)}(\boldsymbol{z}_{\mathcal{I}}^{(\alpha)}) = \phi^{(\alpha)}\left(w_{x,0}^{(\alpha)} + \sum_{l=1}^{q_{\mathcal{I}}^{(\alpha)}} w_{x,l}^{(\alpha)} z_{\mathcal{I},l}^{(\alpha)} \right) = \phi^{(\alpha)}\left(w_{x,0}^{(\alpha)} + \left\langle \boldsymbol{w}_x^{(\alpha)}, \boldsymbol{z}_{\mathcal{I}}^{(\alpha)} \right\rangle \right), \quad x \in \mathcal{X},$$

where $\phi^{(\alpha)} : \mathbb{R} \to \mathbb{R}$ is a (non-linear) activation function, and $w_{x,l}^{(\alpha)} \in \mathbb{R}$ are the network parameters. In matrix form, the output of the layer can be written as:

$$\boldsymbol{\alpha}_{W^{(\alpha)}}(i) = \boldsymbol{f}\left(\boldsymbol{z}_{\mathcal{I}}^{(\alpha)}\right) = \phi^{(\alpha)}\left(\boldsymbol{w}_0^{(\alpha)} + W^{(\alpha)} \boldsymbol{z}_{\mathcal{I}}^{(\alpha)} \right).$$

The subnets $\boldsymbol{\kappa}_{W^{(\kappa)}}(M_t)$ is a two-layered network which extracts the latent factors from the matrix M_t. The first layer is a one-dimensional FCN layer applied to each matrix column individually. It compresses the information of the log-mortality curves arranged in the column of M_t in single real-valued numbers. It can be written as the following mapping:

$$z^{(\kappa_1)} : \mathbb{R}^{|\mathcal{X}|} \to \mathbb{R}, \qquad \log \boldsymbol{m}_t^{(i)} \mapsto z^{(\kappa_1)}\left(\log \boldsymbol{m}_t^{(i)}\right).$$

The output of the layer can be formalised as follows:

$$\log \boldsymbol{m}_t^{(i)} \;\mapsto\; z^{(\kappa_1)}\left(\log \boldsymbol{m}_t^{(i)}\right) = \phi^{(\kappa_1)}\left(w_0^{(\kappa_1)} + \left\langle \boldsymbol{w}^{(\kappa_1)}, \log \boldsymbol{m}_t^{(i)} \right\rangle \right),$$

where $w_0^{(\kappa_1)} \in \mathbb{R}$, $\boldsymbol{w}^{(\kappa_1)} \in \mathbb{R}^{|\mathcal{X}|}$, and $\phi^{(\kappa_1)} : \mathbb{R} \mapsto \mathbb{R}$. The vector collecting the output of the layer of each population is denoted as $z^{(\kappa_1)}(M_t) = \{z^{(\kappa_1)}\big(\log \boldsymbol{m}_t^{(i)}\big)\}_{i\in\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$ contains the information related to the mortality trend of all the populations considered at time t. It is further processed with a S-dimensional FCN layer that maps:

$$z^{(\kappa_2)} : \mathbb{R}^{|\mathcal{I}|} \to \mathbb{R}^S, \qquad z^{(\kappa_1)}\big(M_t\big) \mapsto z^{(\kappa_2)}\big(z^{(\kappa_1)}(M_t)\big).$$

Each unit of this layer extracts a latent factor from $z^{(\kappa_1)}$:

$$z^{(\kappa_1)} \;\mapsto\; z_s^{(\kappa_2)}\big(z_\kappa^{(\kappa_1)}(M_t)\big) = \phi^{(\kappa_2)}\left(w_{s,0}^{(\kappa_2)} + \left\langle \boldsymbol{w}_s^{(\kappa_2)}, z^{(\kappa_1)}(M_t)\right\rangle\right), \quad l = 1, \ldots, S,$$

where $w_{0,s}^{(\kappa_2)} \in \mathbb{R}$, $\boldsymbol{w}_s^{(\kappa_2)} \in \mathbb{R}^{|\mathcal{I}|}$, and $\phi^{(\kappa_2)} : \mathbb{R} \mapsto \mathbb{R}$. The output of the network is obtained as $\boldsymbol{\kappa}(M_t) = z^{(\kappa_2)}\big(z^{(\kappa_1)}(M_t)\big)$.

$\boldsymbol{\gamma}(r)$ aims to determine the optimal contributions of the S latent factors to the population-specific time-indexes. We assume it depends only on the region to guarantee the coherence of the forecasts for male and female populations living in the same country. The network output is a vector containing a set of values that sum 1 and express the contribution of each latent factor to the creation of the region-specific time index. This network consists of an EN and a FCN layer. The embedding is mapping with the same structure in eq. (5):

$$z_{\mathcal{R}}^{(\gamma_1)} : \mathcal{R} \to \mathbb{R}^{q_{\mathcal{R}}^{(\gamma_1)}}, \qquad r \mapsto z_{\mathcal{R}}^{(\gamma_1)}(r) = \left(z_{\mathcal{R},1}^{(\gamma_1)}(r), z_{\mathcal{R},2}^{(\gamma_1)}(r), \ldots, z_{\mathcal{R},q_{\mathcal{R}}^{(\gamma_1)}}^{(\gamma_1)}(r)\right)^\top.$$

The output of this layer is further processed with a S-dimensional FCN layer that can be formalised as:

$$z^{(\gamma_2)} : \mathbb{R}^{q_{\mathcal{R}}^{(\gamma_1)}} \to (0,1)^S, \quad z_{\mathcal{R}}^{(\gamma_1)} \mapsto z^{(\gamma_2)}(z_{\mathcal{R}}^{(\gamma_1)}) = \left(z_1^{(\gamma_2)}(z_{\mathcal{R}}^{(\gamma_1)}), \ldots, z_S^{(\gamma_2)}(z_{\mathcal{R}}^{(\gamma_1)})\right)^\top.$$

Each component is obtained as

$$z_s^{(\gamma_2)}\big(z_{\mathcal{R}}^{(\gamma_1)}\big) = \frac{\exp\left(w_{s,0}^{(\gamma_2)} + \left\langle \boldsymbol{w}_s^{(\gamma_2)}, z_{\mathcal{R}}^{(\gamma_1)}\right\rangle\right)}{\sum_{l=1}^{L}\exp\left(w_{l,0}^{(\gamma_2)} + \left\langle \boldsymbol{w}_l^{(\gamma_2)}, z_{\mathcal{R}}^{(\gamma_1)}\right\rangle\right)} \qquad s = 1, \ldots, S.$$

where $w_{s,0}^{(\gamma_2)} \in \mathbb{R}$, $\boldsymbol{w}_s^{(\gamma_2)} \in \mathbb{R}^{q_{\mathcal{R}}^{(\gamma_1)}}$. The output of this subnet is obtained as $\boldsymbol{\gamma}(r) = z^{(\gamma_2)}\big(z^{(\gamma_1)}(r)\big)$.

4.3 Forecasting and divergence bounds

Once the training is completed and the estimate of the optimal weights $\hat{W}$ is obtained, the mortality model's parameters can be computed via Forward Propagation:

$$\begin{aligned}
\hat{\boldsymbol{a}}_{NN}^{(i)} &= \boldsymbol{\alpha}_{W^{(\alpha)}}(i), \qquad i \in \mathcal{I} \\
\hat{\boldsymbol{\gamma}}_{NN}^{(r)} &= \boldsymbol{\gamma}_{W^{(\gamma)}}(r) \qquad r \in \mathcal{R} \\
\hat{\boldsymbol{\kappa}}_{t,NN}^{(i)} &= \kappa_{W^{(\kappa)}}(M_t^{(i)}), \qquad t \in \mathcal{T}, i \in \mathcal{I}.
\end{aligned}$$

We remark that the ouput of the networks can be interpreted as NN estimates of the mortality model's parameters in eq. (4.1). Since $\boldsymbol{B}$ is constant along the populations and it concides with

the NN parameters, an estimate can be directly obtained from the solution of the optimisation in eq. (4.1). Similar to the other mortality models, we apply the following constraints:

$$\sum_{x\in\mathcal{X}} B_x = 1, \qquad \sum_{t\in\mathcal{T}} \kappa_t^{(s)} = 0, \quad s = 1,\ldots,S.$$

Similarly to the other stochastic models, mortality projections are obtained by keeping the age-specific parameters constant over time, and extrapolating future values of the time indexes the $\kappa_t^{(l)}$ using univariate ARIMA(0,1,0) models:

$$\kappa_t^{(s)} = \theta^{(s)} + \kappa_{t-1}^{(s)} + \upsilon_t^{(s)}, \qquad with\ i.i.d\ \upsilon_t^{(s)} \sim N\left(0, \left(\sigma_\upsilon^{(s)}\right)^2\right)$$

with $\theta^{(s)} \in \mathbb{R}$. The time index related to the population i is obtained through a linear combination of by combining with coefficients $\boldsymbol{\gamma}(r)$. More specifically, mortality forecasts related to the population $i = (r, g)$ are obtained as:

$$\log m_{x,t_n+H}^{(i)} = a_x^{(i)} + B_x \sum_{s=1}^{S} \gamma_s(r) \cdot \left(\kappa_{t_n}^{(s)} + H\theta^{(s)}\right)$$

Since we model $\boldsymbol{\gamma}$ as function of r only, our model ensures no-divergent mortality forecasts for male and female populations living in the same countries. Furthermore, the divergence in the mortality forecasts related to the age x, at time time $t_n + H$ between two populations $i_1 = (r_1, g_1), i_2 = (r_2, g_2) \in \mathcal{I}$ is:

$$\begin{aligned}
\log \frac{m_{x,t_n+H}^{(i_1)}}{m_{x,t_n+H}^{(i_2)}} &= \left[a_x^{(i_1)} + B_x \sum_{s=1}^{S} \gamma(r_1)_s \cdot \left(\kappa_{t_n}^{(s)} + H\theta^{(s)}\right)\right] - \left[a_x^{(i_2)} + B_x \sum_{s=1}^{S} \gamma_s(r_2) \cdot \left(\kappa_{t_n}^{(s)} + H\theta^{(s)}\right)\right], \\
&= \left[a_x^{(i_1)} - a_x^{(i_2)}\right] + B_x\left[\sum_{s=1}^{L} \gamma_s(r_1) \cdot \kappa_{t_n}^{(s)} - \sum_{s=1}^{L} \gamma_s(r_2) \cdot \kappa_{t_n}^{(s)}\right] + \\
&\quad H B_x\left[\sum_{l=1}^{L} \gamma_s(r_1) \cdot \theta^{(s)} - \sum_{l=1}^{L} \gamma_s(r_2) \cdot \theta^{(s)}\right] \\
&= \left[a_x^{(i_1)} - a_x^{(i_2)} + B_x \sum_{s=1}^{L} \kappa_{t_n}^{(s)}(\gamma_s(r_1) - \gamma_s(r_2))\right] + H \sum_{s=1}^{L} \theta^{(s)}(\gamma_s(r_1) - \gamma_s(r_2)).
\end{aligned}$$

Also in this case, the following decomposition can be done: the term $\left[a_x^{(i_1)} - a_x^{(i_2)} + B_x \sum_{s=1}^{L} \kappa_{t_n}^{(s)}(\gamma_s(r_1) - \gamma_s(r_2))\right]$ doesn't change over time, while $H \sum_{s=1}^{L} \theta^{(s)}(\gamma_s(r_1) - \gamma_s(r_2))$ increases linearly in H. However, since $\gamma_s(r) \in (0, 1), s = 1,\ldots,S, r \in \mathcal{R}$, we can derive the following bounds

$$-B_x \sum_{s=1}^{S} \mid \theta^{(s)} \mid < B_x \sum_{s=1}^{S} \theta^{(s)}(\gamma(r_1)_s - \gamma(r_2)_s) < B_x \sum_{s=1}^{S} \mid \theta^{(s)} \mid,$$

which depend on the global age-specific parameter B_x, and the sum the drift terms of the ARIMA(0,1,0) models used for the time-indexes $\kappa^{(s)}$. After estimating the parameters of the MPNNDB model, the model allows for the analytical derivation of the upper and lower bounds for the spread between mortality rates at the same age and time in two populations. If the value $\sum_{s=1}^{S} \mid \theta^{(s)} \mid$ is sufficiently small, we can ensure that the divergence between the mortality rates of all the population considered is relatively small.

5 Numerical experiments

This section presents some numerical experiments performed using real data for validating our MPNNDB model. We investigate the results in terms of forecasting accuracy, the divergence of the mortality forecasts and model explainability. We consider the data from all the countries available in the Human Mortality Database (HMD) (Wilmoth & Shkolnikov, 2021),it is the most popular open-access data source collecting high-quality mortality data. We take into account the set of ages $\mathcal{X} = \{x \in \mathbb{N} : 0 \leq x < 100\}$, the set of calendar years $\mathcal{T} = \{t \in \mathbb{N} : 1960 \leq t < 2020\}$, and the set of populations $\mathcal{I}$ containing the male and female populations of all the countries for which the HMD reports the data from 1960 onwards. We preprocessed the dataset following the procedure suggested in Perla et al. (2021). The NN training is performed using the ADAptive Moment estimation (ADAM) algorithm Kingma and Ba (2014) which is an extension of the SGD algorithm Goodfellow et al. (2016).

We compare the results of the MPNNDB model against the LC model (no coherence) the LL model (full coherence). We assess the out-of-sample accuracy of the different models using the Mean Squared Error (MSE) and the Mean Absolute Error (MAE). Let $\hat{m}_{x,t}^{(i,j)}$ be the estimate of the j-th model $j \in \{MPNNDB, LC, LL\}$), they are respectively defined as:

$$MSE_j = \frac{1}{n} \sum_{i \in \mathcal{I}} \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}} (m_{x,t}^{(i)} - \hat{m}_{x,t}^{(i,j)})^2,$$

$$MAE_j = \frac{1}{n} \sum_{i \in \mathcal{I}} \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}} \mid m_{x,t}^{(i)} - \hat{m}_{x,t}^{(i,j)} \mid .$$

We remark that the MSE is the l_2-norm of the residuals and penalises more larger deviation from the actual values, while the MAE is the l_1-norm of the residuals (Giacalone et al., 2018).

5.1 Out-of-sample accuracy

We conduct a backtesting exercise to assess the out-of-sample accuracy of the proposed model. For a chosen observation period $T \in \mathcal{T}$, the data is split into two parts. The first part, $\mathcal{T}_1$, includes mortality data from calendar years preceding T and serves for model calibration. The second part, $\mathcal{T}_2$, comprises mortality rates from calendar years equal to or following T and is used to evaluate the accuracy of predictions. In this set of experiments, T is fixed at 2000. The use of the MPNNDB model requires to choose the number of latent factors S driving the mortality of the different populations. Since there is no golden rule in choosing this hyperparameter, we start the discussion by analysing the sensitivity of the model performance with respect to S. In particular, we fit the MPNNDB model for different values of $S = 1, 2, \ldots, 5$ and measure the out-of-sample accuracy. Table 1 reports the out-of-sample performance in terms of MSE and MAE and the number of parameters to optimise for each model.

As expected, we note that the forecasting performance of the MPNNDB model improves as S grows. A larger number of latent factors increases the flexibility of the model and allows to better describes the heterogeneity of the mortality pattern of the analysed populations. More specifically, when $S = 1$, the model presents the same functional form of the LL model (in eq. (1)). We observe that it overperforms the LC but is less accurate than the full LL model augumented with the additional population-specific period and time effects. This result emphasizes that a single factor is insufficient to describe well the mortality of

Table 1 Out-of-sample MSE and MAE, and number of network parameters to optimise of the LC, LL and MPNNDB with $S = 1, 2, \ldots, 5$

Model	MSE	MAE	# parameters
LC	4.8221	71.5219	17.280
LL	3.5398	60.6907	17.420
MPNNDB (S =1)	4.0633	71.4386	1.933
MPNNDB (S =2)	3.3687	64.5398	2.016
MPNNDB (S =3)	3.0182	58.8131	2.099
MPNNDB (S =4)	2.9880	58.1529	2.182
MPNNDB (S =5)	3.0170	58.9195	2.265

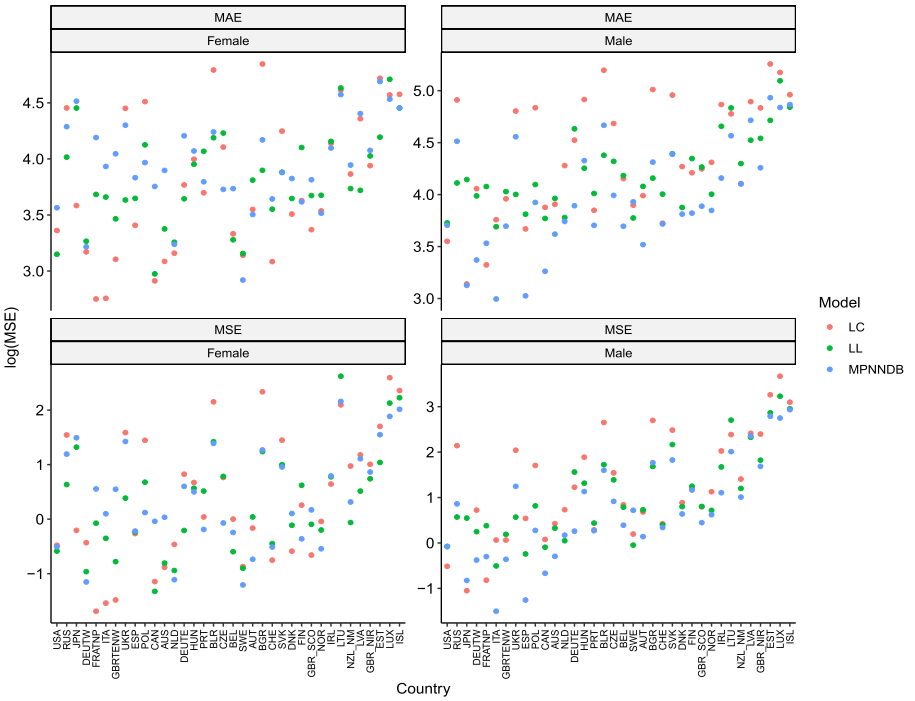

Fig. 1 Forecasting MSE (on log-scale) for the LC, the LL, and the MPNNDB models for the different populations

all 72 populations and that a certain heterogeneity is present. The results notably improve when $S = 2, 3$, where we observe that the MPNNDB model oveperforms in terms of both MSE and MAE. The performance improves further when the number of factors increase to $S = 4$, while no-improvements are registered for $S = 5$. For this reason, we consider the MPNNDB model with $S = 4$ for further comparisons. Furthermore, for all the considered cases, the number of parameters remain much lower than the number of parameters required by the independent LC models and the LL model. A more detailed comparison among the models considered is presented in Fig. 1 which reports the MSE and the MAE obtained in each population.

Looking at the female populations, we observe that there is not a clear winner. On the other side, we observe that the MPNNDB model oveperforms LC and LL on the most of the male

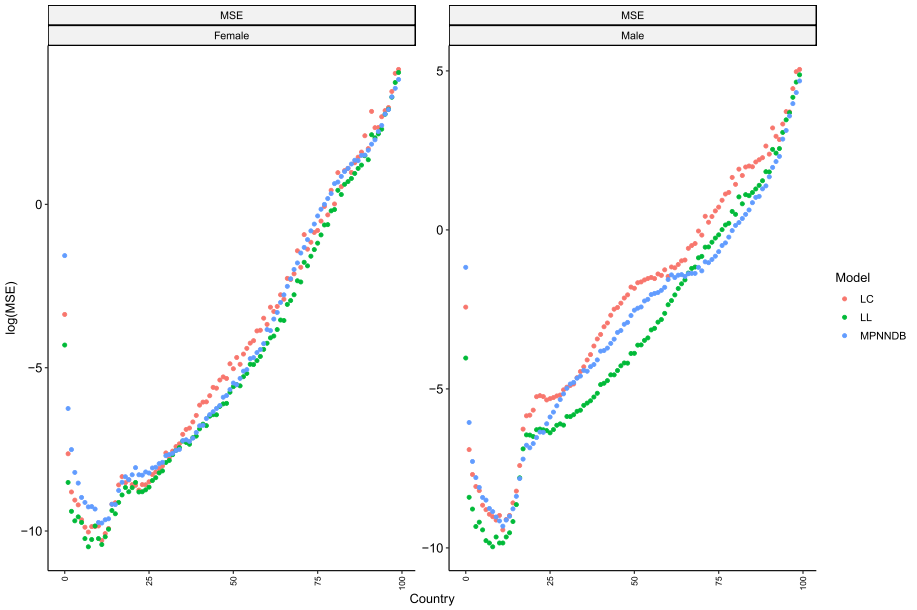

Fig. 2 Forecasting MSE (on log-scale) for the LC, the LL, and the MPNNDB models different ages distinguished by sex

populations. One plausible explanation for this observation is that the mortality dynamics of males tend to show greater stability and less complexity compared to those among females. This can be attributed to the fact that the female population demonstrates more heterogeneity across countries, influenced by sociological factors such as the degree of emancipation, the availability of social security benefits like maternity leave, and their active participation in the labour market. These aspects contribute to a more intricate and varied landscape in female mortality dynamics, distinguishing it from the comparatively more stable patterns observed in the male populations.

In summary, our model produces the best MSE (MAE) results on 27/36 (25/36) populations, the LC on 3/36 (2/36), while the LL on 6/36 (9/36). Fig. 2 depicts the out-of-sample Mean Squared Error (MSE) generated by the different models. In this instance, a logarithmic scale is employed to enhance readability. Once again, concerning females, all three models exhibit comparable performance, with no discernible method demonstrating significant outperformance over the others. Conversely, in the case of males, it is evident that for ages ranging from 25 to 60, the LL model overperforms the others, while the MPNNDB models produce more accurate results for older ages, which is especially relevant from the perspective of longevity risk.

5.2 Parameter estimates

Figure 3 shows the estimates related to the age-specific parameters, while Figure 4 graphically illustrates the estimates related to the four latent factors and their contribution to the country-specific time indexes.

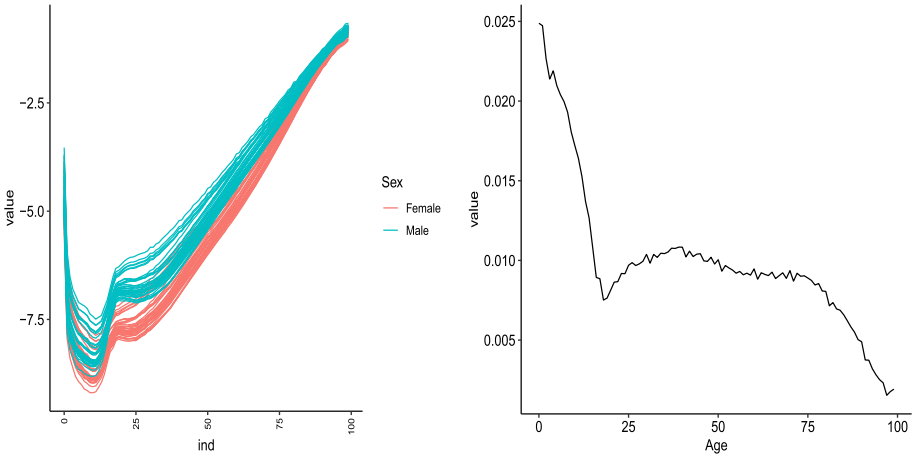

Fig. 3 Estimates of $(a_x^{(i)})_{x\in\mathcal{X}}$, $i \in \mathcal{I}$ (left); $(B_x)_{x\in\mathcal{X}}$ (right)

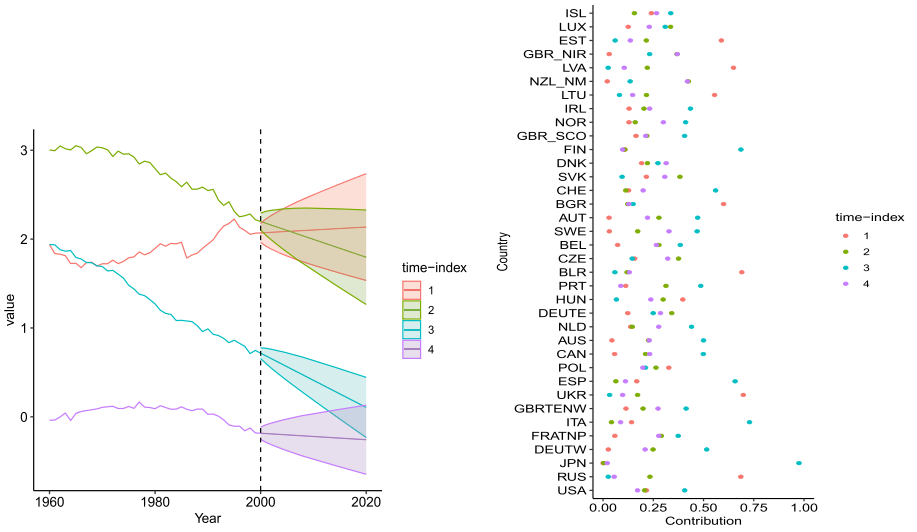

Fig. 4 Estimates and forecasts of $(\kappa_t^{(s)})_{t\in\mathcal{T}}$, $s = 1, \ldots, 4$ (left); estimates of $\boldsymbol{\gamma}^{(r)}$, $r \in \mathcal{R}$ (right)

More specifically, Figure 3 (left) shows the $(a_x)_{x\in\mathcal{X}}$ for the different populations. The curves related to the male populations are in blue, while those related to the female ones are in red. As expected, we observe that all the blue curves lie above to the red ones since the mortality rates of the male populations are generally higher the female ones. Furthermore, we note that the curves are rather smooth over the age dimension compared to other mortality models such as the LC and the LL. Indeed, especially when we consider small populations, the mortality data are noisy due to the presence of sampling error, affecting also the estimates of the parameters. Conversely, our model is estimated using an NN that simultaneously processes all the mortality data and presents some cross-population parameters that allow for information sharing among the different populations. These elements permit to mitigate the impact of the noise on the estimates and improve the robustness of the model. In Fig. 3

Table 2 Parameters for the ARIMA(0,1,0) models corresponding to the four time indices

Time-index	$\theta^{(s)}$	σ^2
1	0.0032	0.0030
2	−0.0201	0.0023
3	−0.0306	0.0009
4	−0.0037	0.0012

(right), we present the $(B_x)_{x\in\mathcal{X}}$ estimates that are the same for all the populations considered. Also, in this case, we obtain a quite smooth curve due to the huge amount of data used for the model calibration. Figure 4 (left) presents the estimates related to the fourth time indexes and the projections related to the years in $\mathcal{T}_2$. We also report the drift and standard deviation of the individual ARIMA(0,1,0) models used to describe these indexes' evolution over time in Table 2.

We note that three of the four indexes present declining trends, highlighting that the mortality rates of most of the 72 populations considered are decreasing over time. The 3rd time-index presents the largest (in absolute value) drift and the steepest death rate decline. Looking at Fig. 4, we note that the index especially contributes to drive the mortality with notable longevity improvements. Indeed, the countries presenting the largest $\gamma^{(3)}$ values are `JPN` (0.9749),`ITA` (0.7287), `FIN` (0.6959), `ESP` (0.6570) that are recognise as countries with high life expectancy.

Only one time-index has an increasing trend and looking at Table 2, one can also note that it presents the largest standard deviation. This index especially contributes to describing the mortality of most countries of Eastern Europe such as Russia (`RUS`), Belarus (`BLR`), Lithuania (`LTU`), Estonia (`EST`), Latvia (`LVA`),Ukraine (`UKR`). This result suggests that there is more uncertainty on the future mortality of those countries. It is unsurprising that these countries exhibit a similar mortality pattern due to their shared cultural background. The political instabilities witnessed in these nations over the past 30 years can be traced back to the dissolution of the USSR. This historical event has undoubtedly left a lasting impact on their political landscapes, contributing significantly to the fluctuations observed in the mortality rates of the populations in these countries.

5.3 Mortality divergences

In this subsection, we analyse the divergences in the mortality forecasts produced by the considered models. Figure 5 illustrates the difference $\log m^{(i_1)}_{x,t+H} - \log m^{(i_2)}_{x,t+H}$ at age $x = 65$ for two different couples of populations over a forecasting horizon $H = 1, \ldots, 200$.

Figure 5 (left) shows the difference between the Italian male and female populations. As expected, the LC model produces divergences that increase indefinitely over time, while the LL model presents a short-term divergence that converges to its asymptotic value when H increases. In this case, the MPNNDB model presents a small fixed divergence since it requires that the coherence property holds between male and female populations living in the same country. Figure 5 (right) shows the divergence between the Russian male and Italian Male populations (right). Also in this case, we note that the LC model divergence explodes, and the LL converges to fixed values when H increases. Conversely, the MPNNDB model presents a divergence that increases slowly over time, avoiding non-reasonable differences in the life expectancy between the two populations. Indeed, even considering the case $H = 200$, the

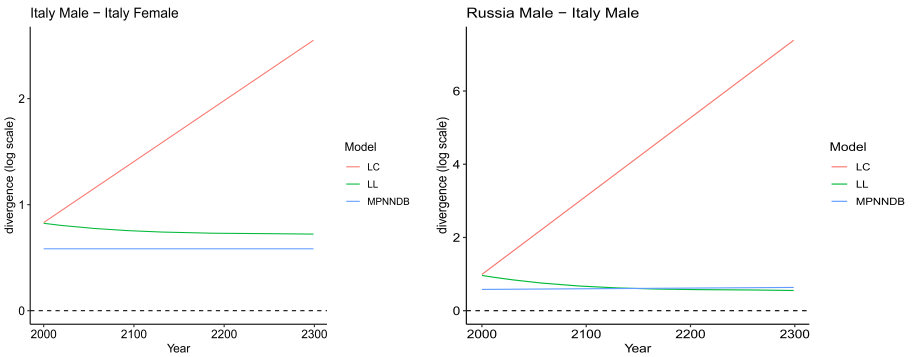

Fig. 5 Divergence in the forecasted mortality rates at age $x = 65$ over the forecasting horizon $H = 0, \ldots, 200$ produced by the different models between Italian males and females (left) and between Russian males and Italian males (right)

divergence produced by the MPNNDB model is very similar to that produced by the LL model.

As discussed in the previous section, the MPNNDB model allows to derive bounds for the divergence in the mortality forecasts between populations. The width of these bounds for two countries $r_1, r_2 \in \mathcal{R}$, depends on the difference $\mid \boldsymbol{\gamma}(r_1) - \boldsymbol{\gamma}(r_2) \mid$. In this setting, we define the following index:

$$D(r_1, r_2) = \|\boldsymbol{\gamma}(r_1) - \boldsymbol{\gamma}(r_2)\|_2$$

that summarises the divergence of the mortality forecasts between the two countries r_1 and r_2. It can be also seen as a measure of the dissimilarity of the mortality patterns. Figure 6 reports the $D(r_1, r_2)$ values for all the countries considered.

We note some horizontal and vertical marked lines corresponding to the Eastern Europe countries, see `RUS`, `BLR` and `UKR` for instance. These lines emphasise that a notable dissimilarity between the mortality pattern on these countries and the other ones. On the other side, we observe that the distance between the Eastern Europe countries is quite small highlighiting that, despite these countries have similar mortality pattern among them. The largest distance is obtained registrered between Japan (`JPN`) and Russia (`RUS`), while the smallest one is between France (`FRATNP`) and Belgium (`BEL`).

6 Conclusion

This paper introduces a new multi-population mortality model that allows us to describe the mortality of a large number of populations simultaneously. The calibration of the model is carried out using neural networks since they allow to process large amounts of data and learn the non-linearities that could affect the evolution of the mortality rates over time. The network architecture is specifically designed to keep a straightforward model interpretation that is crucial for the safe introduction of deep learning models in high-regulated sectors such as the financial and insurance ones. The model also allows us to measure the divergence of the mortality forecasts it produces. Indeed, some differences between the mortality rates of the two populations are constrained within two bounds that could be analytically derived after the parameter estimations. We show that the proposed model can be used to measure

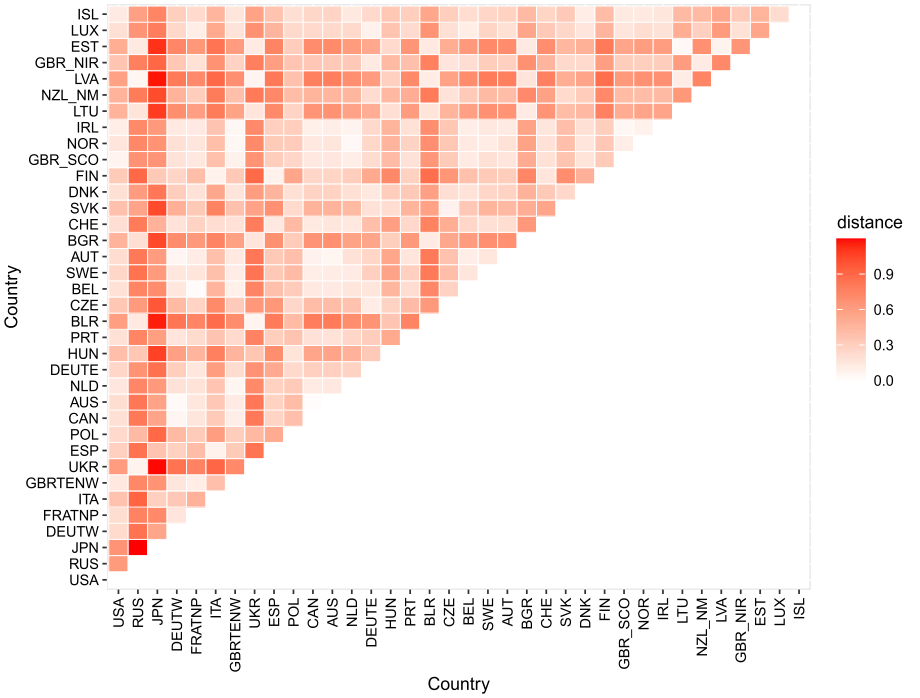

Fig. 6 $D(r_1, r_2)$ values for all the countries considered

the similarity between the mortality patterns of two populations, making the proposed model an interesting tool for identifying longevity risk diversification opportunities. A large set of numerical experiments performed using all the available data in the Human Mortality Database validate the proposed model. It produces more accurate forecasts than some well-known stochastic models, such as the Lee-Carter and the Li-Lee models, in terms of both mean squared error and mean absolute error. In addition, the MPNNDB model allows to obtain a massive reduction in the number of parameters to estimate. This feature reduces the model complexity and the risk of overfitting the population-specific data, which could be the case for the single-population mortality models. Furthermore, the analysis of the similarities among the countries considered shows some heterogeneity derived from geographical, cultural, and historical factors.

Future research will be devoted to analyse sub-national data such as the United States Mortality Database. It could be useful to detect differences in the mortality pattern of populations living in different geographical areas of the same countries. Furthermore, we want to study the use of modern deep learning models based on self-attention-based mechanisms (Vaswani et al., 2017) to further improve the forecasting accuracy of the proposed model.

Acknowledgements The authors thank the two anonymous referees whose comments helped to improve the manuscript.

Funding Open access funding provided by Università Parthenope di Napoli within the CRUI-CARE Agreement.

Declarations

Conflict of interest authors declare that they have no conflict of interest.

Ethical approval this article does not contain any studies with human participants or animals performed by any of the authors.

References

Ahmed, A., Sivarajah, U., Irani, Z., Mahroof, K., & Charles, V. (2022). Data-driven subjective performance evaluation: An attentive deep neural networks model based on a call centre case. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04874-2

Barrieu, P., Bensusan, H., El Karoui, N., Hillairet, C., Loisel, S., Ravanelli, C., & Salhi, Y. (2012). Understanding, modelling and managing longevity risk: key issues and main challenges. *Scandinavian Actuarial Journal, 2012*(3), 203–231.

Basellini, U., Camarda, C. G., & Booth, H. (2022). Thirty years on: A review of the lee-carter method for forecasting mortality. *International Journal of Forecasting, 39*(3), 1033–1049.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137–1155.

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

Cardillo, G., Giordani, P., Levantesi, S., & Nigri, A. (2022). A tensor-based approach to cause-of-death mortality modeling. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-05042-2

Devolder, P., Levantesi, S., & Menzietti, M. (2021). Automatic balance mechanisms for notional defined contribution pension systems guaranteeing social adequacy and financial sustainability: an application to the italian pension system. *Annals of Operations Research, 299*, 765–795.

Dong, Y., Huang, F., Yu, H., & Haberman, S. (2020). Multi-population mortality forecasting using tensor decomposition. *Scandinavian Actuarial Journal, 2020*(8), 754–775.

Giacalone, M., Panarello, D., & Mattera, R. (2018). Multicollinearity in regression: an efficiency comparison between l p-norm and least squares estimators. *Quality & Quantity, 52*, 1831–1859.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT press.

Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. arXiv preprint arXiv:1604.06737

Gupta, S., Modgil, S., Bhattacharyya, S., & Bose, I. (2022). Artificial intelligence for decision support systems in the field of operations research: review and future scope of research. *Annals of Operations Research, 308*, 215–274.

Hainaut, D. (2018). A neural-network analyzer for mortality forecast. *ASTIN Bulletin: The Journal of the IAA, 48*(2), 481–508.

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

Kingma, D.P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Koissi, M.-C., & Shapiro, A. F. (2006). Fuzzy formulation of the Lee-Carter model for mortality forecasting. *Insurance Mathematics and Economics, 39*(3), 287–309.

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

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

Lindholm, M., & Palmborg, L. (2022). Efficient use of data for LSTM mortality forecasting. *European Actuarial Journal, 12*(2), 749–778.

Madhav, M., Ambekar, S. S., & Hudnurkar, M. (2023). Weld defect detection with convolutional neural network: an application of deep learning. *Annals of Operations Research*. https://doi.org/10.1007/s10479-023-05405-3

Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., & Perla, F. (2019). A deep learning integrated lee-carter model. *Risks, 7*(1), 33.

Nigri, A., Levantesi, S., & Marino, M. (2021). Life expectancy and lifespan disparity forecasting: a long short-term memory approach. *Scandinavian Actuarial Journal, 2021*(2), 110–133.

Oeppen, J., & Vaupel, J.W. (2002). Broken limits to life expectancy. American Association for the Advancement of Science.

Perla, F., & Scognamiglio, S. (2023). Locally-coherent multi-population mortality modelling via neural networks. *Decisions in Economics and Finance, 46*(1), 157–176.

Perla, F., Richman, R., Scognamiglio, S., & Wüthrich, M. V. (2021). Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal, 2021*(7), 572–598.

Renshaw, A. E., & Haberman, S. (2006). A cohort-based extension to the lee-carter model for mortality reduction factors. *Insurance Mathematics and Economics, 38*(3), 556–570.

Richman, R., & Wüthrich, M. V. (2021). A neural network extension of the lee-carter model to multiple populations. *Annals of Actuarial Science, 15*(2), 346–366.

Schnürch, S., & Korn, R. (2022). Point and interval forecasts of death rates using neural networks. *ASTIN Bulletin: The Journal of the IAA, 52*(1), 333–360.

Schnürch, S., Kleinow, T., & Korn, R. (2021). Clustering-based extensions of the common age effect multi-population mortality model. *Risks, 9*(3), 45.

Scognamiglio, S. (2022). Calibrating the lee-carter and the Poisson Lee-Carter models via neural networks. *ASTIN Bulletin: The Journal of the IAA, 52*(2), 519–561.

Shi, Y. (2023). Coherent mortality forecasting with a model averaging approach: Evidence from global populations. *North American Actuarial Journal*. https://doi.org/10.1080/10920277.2023.2185260

Tuljapurkar, S., Li, N., & Boe, C. (2000). A universal pattern of mortality decline in the g7 countries. *Nature, 405*(6788), 789–792.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems **30**.

Wilmoth, J.R., & Shkolnikov, V. (2021). Human mortality database. University of California, Berkeley (US), and Max Planck Institute for Demographic Research (Germany).