Life expectancy and lifespan disparity forecasting: a long short-term memory approach

Andrea Nigri, Susanna Levantesi & Mario Marino

Life expectancy and lifespan disparity forecasting: a long short-term memory approach

Andrea Nigri, Susanna Levantesi and Mario Marino

Department of Statistics, Sapienza University of Rome, Rome, Italy

ABSTRACT

After the World War II, developed countries experienced a constant decline in mortality. As a result, life expectancy has never stopped increasing, despite an evident deceleration in developed countries, e.g. England, USA and Denmark. In this paper, we propose a new approach for forecasting life expectancy and lifespan disparity based on the recurrent neural networks with a long short-term memory. This type of neural network leads to predicting future values of longevity indexes while maintaining the significant influence of the past trend, but at the same time adequately reproducing the recent trend into forecasting. The model was applied to five countries for two fitting periods focusing on the forecasting life expectancy and lifespan disparity, both independently and simultaneously at birth and age 65. The results were compared to the projections obtained by four different models, namely, the Double Gap, ARIMA, CoDa and Lee-Carter in the independent case and the first-order Vector Autoregression model in the simultaneous case. Our predictions seem to be coherent with historical trends and biologically reasonable, providing a more accurate portrait of the future life expectancy and lifespan disparity.

KEYWORDS

Life expectancy; lifespan disparity; forecasting; long short-term memory

1. Introduction

Since nineteenth century, developed countries have been experiencing a steady improvement in mortality level, as a result the impact of human longevity on population dynamics has become crucial in defining social and financial policies. Thus, mortality forecasting still represents a prominent research field for both demographers and actuaries (among others, Alho 1990, Lee & Carter 1992, Brouhns et al. 2002, Currie et al. 2004, Li & Lee 2005, Booth et al. 2006, Renshaw & Haberman 2006 , Cairns et al. 2006, Cairns et al. 2009, Bergeron-Boucher et al. 2017).

In particular, the investigation on human lifespan boundaries leads to new approaches focused on life expectancy, bringing new perspectives into mortality forecasting. A breakthrough has been posed by Oeppen & Vaupel (2002) who introduced the concept of 'best-practice life expectancy' (BPLE), i.e. the maximum life expectancy observed among national populations in a given calendar year. They underlined the absence of an impending limit in human life expectancy, disproving the historical estimates of the human life boundary. The BPLE approach introduced a new line of investigation in which researchers devote their efforts to modeling the life expectancy trend. Many promising studies have been carried out, starting from Lee (2006), who considers a stochastic behavior of changes in the life expectancy trend, assuming that the average changes are functions of the gap with the BPLE. Similarly, using data dating back to twentieth century, Torri & Vaupel (2012) bring forward a Geometric Brownian motion-based model, overcoming the main limitation of Lee's approach in which future

life expectancy can exceed the level of BPLE. On the other hand, Pascariu et al. (2018) state that 'the Torri-Vaupel approach is promising but has the drawback that populations that lag behind record life expectancy cannot become the record holder'. Moreover, the Torri and Vaupel model excludes the shock points in the life expectancy trend, replacing them with the long-term trend. Another significant contribution comes from Raftery et al. (2013) who consider several countries and propose a hierarchical Bayesian model. Finally, Pascariu et al. (2018) take up the forecasting approaches based on the BPLE gap by using the gap between country female life expectancy and BPLE for women, and the gap with female life expectancy in the same country for men.

Despite an evident deceleration in developed countries (Aburto et al. 2018; Hiam et al. 2018; Ho & Hendi 2018), empirical evidence does not show an impending limit on life expectancy, which is a measure of average mortality levels that hides variation in the lifespan of each individual. Populations with the same level of life expectancy may have age distribution of deaths with different degree of uniformity. A demographical indicator capturing the variation in lifespan, thus providing a measure of dispersion in the age-at-death at individual level, is the lifespan disparity. As pointed out by van Raalte et al. (2018), lifespan disparity can be considered an estimation of the heterogeneity at population level. Its understanding and adoption are crucial both in the insurance market and public system, as well as for modeling and forecasting mortality levels (Edwards & Tuljapurkar 2005). However, as stated by Aburto et al. (2020) 'few countries have begun to monitor and acknowledge the importance of disparities in age at death' and none has monitored the age-at-death diversity across those who have already survived to age 65 (lifespan disparity at age 65). So far, there are no scientific contributions providing projections for this indicator yet.

Our paper contributes to the literature by proposing a new method for forecasting life expectancy and life disparity (at birth and at age 65) based on a long short-term memory (LSTM) architecture. As pointed out by Nigri et al. (2019), LSTM is a recurrent neural network able to elaborate sequences of data preserving significant either short and long term dependencies. Specifically, LSTM allows to predict future values maintaining the noteworthy influence of the past trend and adequately reproducing it into forecasting. Therefore, the resulting future values of life expectancy and lifespan disparity should be more consistent with the historical dynamics and meet biological reasonableness criteria, first the non-linearity.

The use of neural networks is gradually spreading in the insurance literature, e.g. Richman & Wüthrich (2019a) apply deep neural networks, consisting of multiple layers, to extend the Lee-Carter model to multiple populations by automatically selecting the model features; Wüthrich (2019) and Gabrielli et al. (2020) use feed-forward neural networks to enhance the traditional generalized linear models, the former providing a motor insurance pricing application, the latter dealing with claim reserving problems. While, at the state of art, the LSTM networks have been considered in actuarial science only by Richman & Wüthrich (2019b), who have applied them to predict mortality rates. Finally, Nigri et al. (2019) use LSTM to improve the accuracy of the Lee-Carter model forecasts.

Our approach, based on the LSTM network, mainly consists of forecasting life expectancy and life disparity independently using an univariate network. The analysis of lifespan disparity may allow us to acquire further knowledge on the life expectancy future evolution. However, these indicators may be linked by a long-term relationship (Bohk-Ewald et al. 2017, Aburto & van Raalte 2018, Aburto et al. 2020), therefore the forecasting accuracy might take advantage by simultaneous modeling, exploiting the potential link between the dynamics of the two series. Within the recurrent neural network setting, the simultaneous forecasting of two time series requires the construction of a bivariate network. Thus, we also propose a bivariate LSTM framework aimed at forecasting life expectancy and lifespan disparity simultaneously.

We provide a numerical application to demonstrate the strong predictive power of univariate LSTM networks compared to the ARIMA model and the Double Gap model (DG) proposed by Pascariu et al. (2018), which applies to life expectancy but not to lifespan disparity. The ARIMA model can be considered as a benchmark for time series forecasting, while the DG model represents a prominent approach which might be seen as an improvement of ARIMA, allowing to consider the gender

gap in life expectancy trend. In addition, we provide a further comparison with two extrapolative models: Lee-Carter (the extension proposed by Brouhns et al. (2002)) and CoDa (Oeppen 2008), which is based on the principal component analysis. The bivariate LSTM is compared to the first-order Vector Autoregression model (VAR) that is often used as a benchmark for multivariate series forecasting.

This work is structured as follows. Section 2 introduces life expectancy and lifespan disparity. Section 3 describes the functioning of recurrent neural networks with a specific paragraph dedicated to the LSTM. Section 4 describes the life expectancy and lifespan disparity modeling from the LSTM networks' perspective, highlighting the connection between the concepts from neural networks and the input data in a demographic framework. Section 5 illustrates the numerical application carried out on five countries of the world: Australia, Italy, Japan, Sweden, and the USA. Finally, Section 6 provides conclusions.

2. Life expectancy and lifespan disparity

The achievement of longer lives has been driven by a decline in infant mortality, and by reductions in mortality at older ages after the World War II (Vaupel 1997, Rau et al. 2008). The constant improvement of BPLE suggests that the mortality reductions should not be viewed as a disconnected sequence of unrepeatable revolutions, but rather as a regular flow of continuous progress (Oeppen & Vaupel 2006). Indeed, mortality developments are linked to social progress in terms of health, nutrition, education, hygiene, and medicine (Riley 2001).

However, populations characterized by the same level of life expectancy could experience substantial differences in the time of death (Aburto et al. 2020), with different age-at-death distributions. As mentioned in the introduction, life expectancy is not likely to detect variations in lifespan, which are instead captured by lifespan disparity allowing to describe variations in lifespan distribution (Bohk-Ewald et al. 2017).[1] While life expectancy has been proved to hide heterogeneity in individual mortality paths, lifespan disparity measures the dispersion of observations around the time of death, evaluating from, respectively, a probabilistic and a descriptive point of view, uncertainty in age-at-death distribution and heterogeneity (van Raalte et al. 2018, Kaakai et al. 2019). When mortality is highly variable, some individuals will die at a much younger age than the expected age-at-death, contributing many lost years to life disparities; conversely, when mortality is highly concentrated around older ages or the modal age, life disparity decreases (Aburto & van Raalte 2018).

In the following, we provide the formal notation and definitions, first for life expectancy and then for lifespan disparity.

- *Life expectancy*
  Let $S(x,t)$ and $\mu(x,t)$ be two continuous functions with respect to age x and time t, respectively representing the survival function and the force of mortality of an individual aged x at time t in a given population. We denote by $e_{x,t}$ the life expectancy at age x and time t, that is defined as follows:
  $$e_{x,t} = \frac{\int_x^\infty S(y,t)\,\mathrm{d}y}{S(x,t)} \tag{1}$$
  where $S(x,t) = \exp\left(-\int_0^x \mu(a,t)\,\mathrm{d}a\right)$ is the survival function and $\mu(a,t)$ is the force of mortality at age a at time t.

- *Lifespan disparity*
  According to Vaupel (1986), the lifespan disparity is an indicator representing the life

[1] In addition to life disparity, other inequality measures have been proposed in literature, e.g. the Gini coefficient and the Keyfitz's entropy (Wilmoth & Horiuchi 1999, Shkolnikov et al. 2003, van Raalte & Caswell 2013) that appear to be linearly related and negatively correlated to life expectancy at birth (Colchero et al. 2016, Nemeth 2017, Aburto et al. 2020).

expectancy lost due to death by an individual aged x at time t. Formally, the lifespan disparity at birth is defined as follow:

$$e^{\dagger}_{0,t} = -\int_0^{\infty} S(a,t) \cdot \ln S(a,t)\, \mathrm{d}a, \tag{2}$$

where the term e-dagger, $e^{\dagger}$, was coined by Vaupel & Canudas-Romo (2003). A more intuitive expression was derived by Vaupel (1986) and Goldman & Lord (1986):

$$e^{\dagger}_{0,t} = \frac{\int_0^{\infty} e_{a,t} \cdot d(a,t)\, \mathrm{d}a}{S(0,t)}. \tag{3}$$

From Equation (3), the lifespan disparity above age x can be defined as:

$$e^{\dagger}_{x,t} = \frac{\int_x^{\infty} e_{a,t} \cdot d(a,t)\, \mathrm{d}a}{S(x,t)}, \tag{4}$$

where $e_{a,t}$ is the remaining life expectancy at age a at time t and $d(a,t)$ are the deaths at age a at time t.

Since the same information is involved in the calculation of both life expectancy and lifespan disparity, the relationship between these two indicators has been discussed by several researchers. For example, Bohk-Ewald et al. (2017) proposed to evaluate the performance of extrapolative mortality models by analyzing both the average lifespan and lifespan disparity, while Rabbi & Mazzuco (2020) to adjust the time component of the Lee-Carter model with the observed lifespan disparity. Aburto & van Raalte (2018) explored trends in lifespan disparity under periods of life expectancy decline by focusing on Central and Eastern Europe. They measured the relationship between life expectancy and lifespan disparity by their absolute and relative changes. Aburto et al. (2020) developed a mathematical framework to jointly explore the evolution over time of life expectancy at birth and lifespan equality analyzing three different indicators of lifespan equality: life table entropy, Gini coefficient, and coefficient of variation of the age-at-death distribution. They found a strong link between life expectancy and each life span equality indicator, especially when life expectancy is less than 70 years. These studies generally investigate life expectancy and lifespan variation since birth, without considering the dispersion in the time of death conditioned on survival at a specific age, as well as the forecasting.

Both life expectancy and lifespan disparity might be understood as latent variables encompassing many factors that, directly or indirectly, affect mortality dynamics. This latent behavior should be emphasized in forecasting by incorporating both short term history and contribution from long term improvements in more recent periods. Bearing in mind the latter, we need models able to catch more in-depth the unobservable features in the historical observations. Our approach, based on the LSTM network, meets these needs, providing more accurate forecasts of life expectancy and lifespan disparity with respect to other well-established models, overcoming the above limitations.

3. Recurrent neural network

Neural networks with multiple hidden layers have recently become very popular for treating sequential data in a wide variety of tasks, such as automatic speech recognition, natural language processes, social network filtering and medical diagnosis (Rojas & Feldman 1996).

There are different types of neural networks and the number of their variants is growing exponentially; for instance: multilayer perceptron, convolutional, recursive, recurrent, long short-term memory and auto encoder.

In this paper, we consider the recurrent neural networks (RNNs) that are networks suitable for time series analysis. RNNs are characterized by units self-connected or connected to units of the

previous layers (in addition to the feedforward connections). The recurrence implies short-term memory in order to store information from the past inputs (see Rumelhart et al. 1986, Werbos 1988 and Elman 1990 for further details on RNNs) and allows to discover temporal correlations between events that may be far from each other. This latter feature is vital for time series learning. The neural network model and its architecture determine how a network transforms inputs into outputs. The fundamental unit is the neuron that receives the inputs, next the weights are applied to the inputs and transferred in an activation function together with the bias.

Given an observed input received by the network at time t, $\boldsymbol{x}_t \in \mathbb{R}^n, n \in \mathbb{N}$, and the associated target $\boldsymbol{y}_t \in \mathbb{R}^m, m \in \mathbb{N}$, we pursue the goal of learning the unknown temporal map $\psi : \boldsymbol{x}_t \mapsto \boldsymbol{y}_t$, using an RNN with a single hidden layer constituted by $\kappa \in \mathbb{N}$ neurons.

Let $\boldsymbol{h}_t \in \mathbb{R}^\kappa$ be the hidden layer containing the information at time t and f_h and f_y be the nonlinear functions of the hidden layer and output, respectively. $\boldsymbol{h}_t$ is a function of the input at the same time step and the hidden layer of the previous time step, $\boldsymbol{h}_{t-1}$, while the output at time t is a function of the hidden layer at the same time step. Therefore, this dynamic system is defined by the following two equations:

$$\boldsymbol{h_t} = f_h(\boldsymbol{x}_t, \boldsymbol{h_{t-1}}); \tag{5}$$

$$\boldsymbol{y}_t = f_y(\boldsymbol{h_t}). \tag{6}$$

Let define $\boldsymbol{W} = \{\boldsymbol{W}_{yh}, \boldsymbol{W}_{hh}, \boldsymbol{W}_{hx}\}$ the parameters characterizing the RNN, where:

- $\boldsymbol{W}_{hx} \in \mathbb{R}^{\kappa \times n}$ is the weight matrix between the input and the hidden layer;
- $\boldsymbol{W}_{hh} \in \mathbb{R}^{\kappa \times \kappa}$ is the recurrences weight matrix within hidden layers;
- $\boldsymbol{W}_{yh} \in \mathbb{R}^{m \times \kappa}$ is the weight matrix between the hidden layer and the output.

Moreover, the bias vectors $\boldsymbol{b}_h \in \mathbb{R}^\kappa$ and $\boldsymbol{b}_y \in \mathbb{R}^m$ are added to the parameters set of the hidden layer and the output layer, respectively. These additional parameters are necessary to govern the triggering value for the activation functions f_h and f_y, as they ensure an affine transformation of the input, $\boldsymbol{x}_t$, and the hidden layer, $\boldsymbol{h}_t$. Therefore, each bias component acts as a neuron activation threshold and allows for a reliable data elaboration during the neural networks training, avoiding a poorly fit and providing a major model flexibility.

Equation (5) and (6) can be written as a function of the weight matrices and bias vectors:

$$\boldsymbol{h}_t = f_h(\boldsymbol{W}_{xh}^T \cdot x_t + \boldsymbol{W}_{hh}^T \cdot \boldsymbol{h}_{t-1} + \boldsymbol{b}_h); \tag{7}$$

$$\boldsymbol{y}_t = f_y(\boldsymbol{W}_{yh}^T \cdot \boldsymbol{h}_t + \boldsymbol{b}_y). \tag{8}$$

An illustration of the RNN architecture is provided in Figure 1. Similarly to the feedforward neural networks, the RNNs are based on two main steps: the forward and the back propagation. The former calculates and stores the weighted inputs (i.e. $\boldsymbol{W}_{xh}^T \cdot \boldsymbol{x}_t$ and $\boldsymbol{W}_{hh}^T \cdot \boldsymbol{h}_{t-1}$) going in the forward direction from input layer to output layer. The latter is a technique that minimizes the error between predicted and actual values by updating weights and bias through gradient-based method. The error is back propagated in the network from output layer to input layer towards the hidden layers and used to adjust the network initial weights. The network output is then the result of an iterative optimization procedure. In particular, let $L_t(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t)$ be the loss function at time step t, where $\hat{\boldsymbol{y}}_t$ is the output estimated by the RNN. The overall loss function is defined as the sum of the losses over t:

$$L(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \sum_{t=1}^{T} L_t(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t). \tag{9}$$

The network parameters estimation stems from the minimization of Equation (9). It is worth noting that for long time series, the RNN optimization procedure shows some learning problems. As pointed

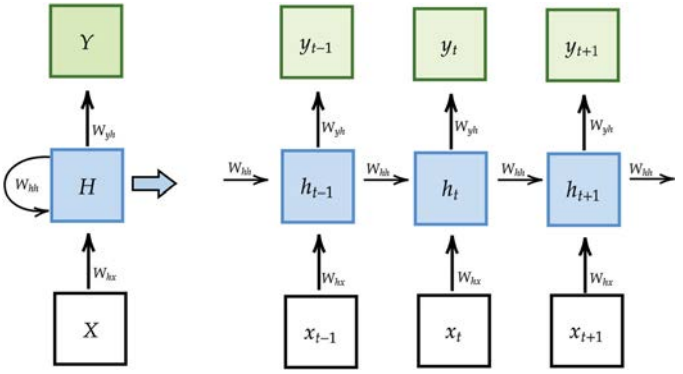

Figure 1. The architecture of a one-hidden layer RNN.

out by Bengio et al. (1994) and Pascanu et al. (2013), RNNs suffer from one main structural drawback affecting their performances: the vanishing or exploding gradient. Because of the recurrent structure, the issue refers to the large (or small) increase in the norm of the loss function gradient, since the long term components dominate over the short term ones. Further details on the vanishing/exploding gradient and the parameter optimization procedure are provided in Appendix 1.

3.1. Long short-term memory

Hochreiter & Schmidhuber (1997) came up with a solution to the vanishing (exploding) gradient problem introducing an innovative structure of the recurrent elaborations: the so called LSTM architecture (further developed by Gers et al. 1999, Gers & Schmidhuber 2000, Graves et al. 2009). The LSTM is an extension of RNN models providing a memory unit within neurons of the hidden layers, being able to handle long sequences of data.

The LSTM neuron, also called LSTM block, is made up of two fundamental parts. The first is the memory or cell unit, $\boldsymbol{c}_t$, which incorporates significant information over time, allowing long-term dependencies to be maintained by integrating them from time to time with the inputs of the current time step. Then we have the gates, that let current information to be elaborated through time. Gates are divided into:

(1) Input gate, $\boldsymbol{i}_t$, that transfers current information input to the memory unit by a sigmoid activation function, $\sigma(x) = \frac{1}{1+e^{-x}}$;
(2) Input modulation gate, $\boldsymbol{z}_t$, is auxiliary of the input gate $\boldsymbol{i}_t$, using a hyperbolic tangent as the activation function, $\phi(x) = \tanh(x)$;
(3) Forget gate, $\boldsymbol{f}_t$, necessary to reset the memory unit through a sigmoid function;
(4) Output gate, $\boldsymbol{o}_t$, that controls the block output through a sigmoid function.

The following set of equations describes the forward flow of an LSTM block:

$$\boldsymbol{f}_t = \sigma\left(\boldsymbol{W}_{fx}\boldsymbol{x}_t + \boldsymbol{W}_{fh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_f\right); \tag{10}$$

$$\boldsymbol{i}_t = \sigma\left(\boldsymbol{W}_{ix}\boldsymbol{x}_t + \boldsymbol{W}_{ih}\boldsymbol{h}_{t-1} + \boldsymbol{b}_i\right); \tag{11}$$

$$\boldsymbol{o}_t = \sigma\left(\boldsymbol{W}_{ox}\boldsymbol{x}_t + \boldsymbol{W}_{oh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_o\right); \tag{12}$$

$$\boldsymbol{z}_t = \phi\left(\boldsymbol{W}_{zx}\boldsymbol{x}_t + \boldsymbol{W}_{zh}\boldsymbol{h}_{t-1} + \boldsymbol{b}_z\right); \tag{13}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \boldsymbol{z}_t; \tag{14}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \phi(\boldsymbol{c}_t); \tag{15}$$

$$\boldsymbol{y}_t = \varphi\left(\boldsymbol{W}_{yh}\boldsymbol{h}_t + \boldsymbol{b}_y\right). \tag{16}$$

Therefore, at each time step t the LSTM block receives the current inputs $\boldsymbol{x}_t$ and $\boldsymbol{h}_{t-1}$ that are jointly elaborated by the forget, input, input modulation and output gates ($\boldsymbol{i}_t, \boldsymbol{z}_t, \boldsymbol{f}_t, \boldsymbol{o}_t$, respectively), whose outputs are defined by Equations (10)–(13). Afterwards, the current state of memory $\boldsymbol{c}_t$ is created by combining the previous one and adding the modulated current input, as described in Equation (14). In this way, the memory unit $\boldsymbol{c}_t$ preserves the significant long term correlations among the entire time series. Instead, the short term relations are expressed by the Equation (15) involving the output gate and the hyperbolic tangent transformation of the memory unit. Finally, the terminal output in Equation (16) is defined by means of a generic activation function φ.

4. Life expectancy and lifespan disparity modeling

In this section, we will describe the model used to forecast country-specific life expectancy and lifespan disparity, both independently and simultaneously, considering two ages: 0 and 65.

Let $\{e_{x,t}\}_{t=t_0}^{t_s}$ and $\{e^{\dagger}_{x,t}\}_{t=t_0}^{t_s}$, for $t_0 < t_s$, be the country-specific observed time series of life expectancy and lifespan disparity, respectively. Let $\{e_{x,t}, e^{\dagger}_{x,t}\}_{t=t_0}^{t_s}$ be the country-specific bivariate series we would like to model simultaneously. Following an appropriate rule, each series is split into a training set and a testing set, where the first one is used for fitting the model's parameters, while the second one to test the model's prediction and calculate the error. Let t_τ be the calendar year corresponding to the last realization on the training set. The training and testing sets for the life expectancy series are defined as follows:

$$\begin{aligned}
&\text{TRAINING SET (TR)} : \mathcal{TR}^{(e)} = \left\{e_{x,t}\right\}_{t=t_0}^{t_\tau} \\
&\text{TESTING SET (TS)} : \mathcal{TS}^{(e)} = \left\{e_{x,t}\right\}_{t=t_\tau+1}^{t_s}
\end{aligned}$$

Similarly, we can define training and testing sets for lifespan disparity, $\mathcal{TR}^{(e^\dagger)}$ and $\mathcal{TS}^{(e^\dagger)}$, and for the bivariate series, $\mathcal{TR}^{(e,e^\dagger)}$ and $\mathcal{TS}^{(e,e^\dagger)}$.

4.1. LSTM model

In the LSTM network, aimed at forecasting life expectancy and lifespan disparity, we adopt a first-order autoregressive approach. Therefore, denoting $\psi^{(\cdot)}_{LSTM}$ as a composition of functions defined in Equation (16) according to the optimal number of LSTM blocks, the model is described by:

$$\begin{aligned}
e_{x,t} &= \psi^{(e)}_{LSTM}\left(e_{x,t-1}\right) + \varepsilon^{(e)}_t \quad \text{or} \\
e^{\dagger}_{x,t} &= \psi^{(e^\dagger)}_{LSTM}\left(e^{\dagger}_{x,t-1}\right) + \varepsilon^{(e^\dagger)}_t \quad \text{or} \\
\left[e_{x,t}, e^{\dagger}_{x,t}\right] &= \psi^{(e,e^\dagger)}_{LSTM}\left\{\left[e_{x,t-1}, e^{\dagger}_{x,t-1}\right]\right\} + \varepsilon^{(e,e^\dagger)}_t,
\end{aligned} \tag{17}$$

where $\varepsilon^{(\cdot)}_t$ is a zero mean error. The set of functions $\psi^{(\cdot)}_{LSTM}$ is the map linking life expectancy or lifespan disparity or both at an annual pace. In a first-order autoregressive approach, the network learns at each time step the relationship between consecutive values on the training set and, according to the same logic, predicts the future values on the testing set. This process is optimized using an L2

loss function.

$$\min_{W} \frac{1}{2} \sum_{t=t_0}^{t_\tau} \left(e_{x,t} - \hat{e}_{x,t}\right)^2 \quad \text{or}$$

$$\min_{W} \frac{1}{2} \sum_{t=t_0}^{t_\tau} \left(e^{\dagger}_{x,t} - \hat{e}^{\dagger}_{x,t}\right)^2 \quad \text{or} \tag{18}$$

$$\min_{W} \frac{1}{2} \sum_{t=t_0}^{t_\tau} \left\{ \left[e_{x,t}, e^{\dagger}_{x,t}\right] - \left[\hat{e}_{x,t}, \hat{e}^{\dagger}_{x,t}\right] \right\}^2,$$

where $\boldsymbol{W} = \{\boldsymbol{W}_{fx}, \boldsymbol{W}_{fh}, \boldsymbol{W}_{ix}, \boldsymbol{W}_{ih}, \boldsymbol{W}_{ox}, \boldsymbol{W}_{oh}, \boldsymbol{W}_{zx}, \boldsymbol{W}_{zh}, \boldsymbol{W}_{yh}\}$ is the LSTM parameters set.

4.1.1. LSTM in a demographical framework

We are now going to connect the concepts from RNN to the input data used in the application, aiming at creating a bridge between RNN and demography. In the following, we will only refer to life expectancy (the extension to life disparity and to the bivariate case is straightforward). In our model, the input received by the network at state t is life expectancy at a given age x, i.e. $\boldsymbol{x}_t \equiv \boldsymbol{e}_{x,t}$. The output of the network at state t is the life expectancy at time $t+1$, consistently with the first-order autoregressive pattern, that is $\boldsymbol{y}_t \equiv \boldsymbol{e}_{x,t+1}$. Therefore, following the Equation (16), $\boldsymbol{e}_{x,t+1} \equiv \varphi(\boldsymbol{W}_{eh}\boldsymbol{h}_t + \boldsymbol{b}_y)$ is the theoretical relationship defining the life expectancy at year $t+1$, given the life expectancy at year t, and the LSTM block processing. The final output of LSTM, after the estimation procedure, which implies to estimate the weights, becomes:

$$\hat{\boldsymbol{e}}_{x,t+1} = \varphi\left(\hat{\boldsymbol{W}}_{eh}\boldsymbol{h}_t + \hat{\boldsymbol{b}}_e\right). \tag{19}$$

where $\hat{\boldsymbol{e}}_{x,t+1}$ is the life expectancy estimation resulting from the application of the estimated parameters (weights matrix $\hat{\boldsymbol{W}}_{eh}$ and bias $\hat{b}_e$), obtained by the optimization procedure described in Appendix 1.

4.2. Other models

The actuarial and demographic literature provides a wide variety of mortality models. In this paper, the performance of the univariate LSTM network is compared to the ARIMA, DG, Lee-Carter and CoDa models. LSTM, ARIMA and DG models allow to directly work with the life expectancy and life disparity time series, without passing through an extrapolative stochastic model which provides the mortality rates used to calculate such demographical indicators. However, we also consider two extrapolative models: Lee-Carter, which is probably the most used by practitioners and CoDa, which forecasts the life table distributions of deaths ($d_{x,t}$) using principal component analysis in a compositional data pattern. While the performance of the bivariate LSTM is compared to the VAR model. A brief description of these models is reported in Appendix 2.

5. Numerical illustration

In the numerical application, we consider historical mortality data collected by gender from the Human Mortality Database (2018) for Australia, Italy, Japan, Sweden, and USA.

It is well known that mortality modeling is a process that should fulfill some qualitative criteria, robustness, among others. Thus, the forecast should not be too sensitive towards the selected period's choice, but it should be consistent with historical data. Therefore, in our analysis, we will carry out an out-of-sample test considering the same forecast horizon for two different overlapping estimation periods: 1938–1999 and 1947–1999. The time frame 2000–2014 is then used as evaluation chunk. In

this way, we obtain a sufficient size for training and testing sets in both the time frames, according to the common splitting rule: 80% and 20%. Finally, to assess the models' accuracy, we calculate the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Before training the LSTM for all countries and both genders, we will implement a preliminary fine-tuning to identify the optimal hyper-parameters, such as mini-batch size, epochs, and neurons number for each hidden layer (see Appendix 1 for further details). For this purpose, we will select a finite set for each hyper-parameter, exploring the specification minimizing the loss function. The best combination obtained in the training phase is used to calibrate LSTM in the forecasting one. The mini-batch size is equal to the number of training samples in one forward/backward pass before updating the model weights. In our case, the mini-batch size is equal to 1, as our input data have been arranged into a column vector, where each row represents the life expectancy at a generic time t. Therefore, we need to compute the weight's update for each one-time step. It is worth noting that a batch size greater than 1 is not consistent with our autoregressive framework based on one order of differentiation. Not least, the literature suggests that the use of small batch sizes improves the out-of-sample performance and the optimization convergence (LeCun & Muller 2012, Keskar et al. 2016) requiring small memory (then gaining efficiency) by exploiting memory locality. The architectures with a single hidden layer work better than others, and the number of neurons and epochs depends on the specific-country data. In our model, the loss function is minimized over the neural network weights using the Adadelta (Zeiler 2012), a variant of the Stochastic Gradient Descent (SGD) method. We use the Rectified Linear Unit (ReLU) (Glorot et al. 2011) as activation function φ involved in the terminal output (Equation (16)) that outperformed the other tested functions.

The LSTM performances are compared to the models presented in Section 4.2. Therefore, we will compare the univariate LSTM to the best ARIMA(p, d, q), DG, Lee-Carter and CoDa models, while

Table 1. Out-of-sample test for $e_{0,t}$: MAE and RMSE for LSTM, ARIMA, DG, LC and CoDa model by country and gender.

		Fitting period: 1938–1999				Fitting period: 1947–1999			
		Female		Male		Female		Male	
Country	Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	ARIMA	0.3118	0.4149	0.8111	0.8504	0.2167	0.2844	0.2901	0.3216
	DG	0.3139	0.4175	0.2693	0.2896	0.1945	0.2277	0.3219	0.3468
	LSTM	**0.1139**	**0.1412**	**0.1485**	**0.1895**	**0.1110**	**0.1362**	**0.1407**	**0.1804**
	LC	0.2525	0.2806	1.0740	1.2133	0.2869	0.3204	1.0368	1.1640
	CoDa	0.1347	0.1655	1.1936	1.2763	0.1304	0.1639	1.0022	1.0629
Italy	ARIMA	1.5759	1.8872	0.9157	1.0819	0.3434	0.4455	0.1768	0.2155
	DG	0.2986	0.3836	0.2355	0.2697	0.2314	0.2722	0.2209	0.2444
	LSTM	0.1914	0.2304	**0.1396**	**0.1767**	0.2104	0.2587	**0.1758**	**0.2124**
	LC	**0.1663**	**0.2068**	1.8194	1.9259	**0.1518**	**0.1969**	1.5136	1.6463
	CoDa	0.4275	0.5507	0.9763	1.0531	0.4156	0.5356	1.0985	1.1880
Sweden	ARIMA	0.4305	0.4659	0.4760	0.5484	0.4467	0.4672	0.2696	0.3058
	DG	0.4305	0.4659	0.1659	0.1888	0.4467	0.4671	0.3983	0.4232
	LSTM	**0.0773**	**0.0964**	**0.0574**	**0.0703**	**0.0752**	**0.1000**	**0.0598**	**0.0718**
	LC	0.1761	0.1973	0.9698	1.0815	0.0823	0.1149	1.0199	1.1245
	CoDa	0.4079	0.4574	0.9496	1.0627	0.6612	0.7449	0.8578	0.9571
USA	ARIMA	0.7358	0.8898	0.1892	0.2449	0.6822	0.8165	0.1455	0.1845
	DG	0.7358	0.8898	1.3553	1.5444	0.6821	0.8164	1.0669	1.2119
	LSTM	0.2466	0.2939	**0.1140**	**0.1381**	0.3522	0.4279	**0.1137**	**0.1375**
	LC	**0.1173**	**0.1451**	0.5017	0.5950	0.3847	0.4096	0.7549	0.8336
	CoDa	0.2390	0.2688	0.4505	0.5432	**0.1038**	**0.1266**	0.4529	0.5451
Japan	ARIMA	–	–	–	–	**0.1712**	**0.2291**	1.2220	1.4085
	DG	–	–	–	–	0.5569	0.5894	0.3721	0.4210
	LSTM	–	–	–	–	0.3342	0.3694	**0.2252**	**0.2662**
	LC	–	–	–	–	0.6543	0.7650	0.4330	0.5068
	CoDa	–	–	–	–	1.2086	1.5032	0.9961	1.2106

Notes: Years 2000–2014. Fitting periods: 1938–1999 (columns 3–6) and 1947–1999 (columns 7–10).

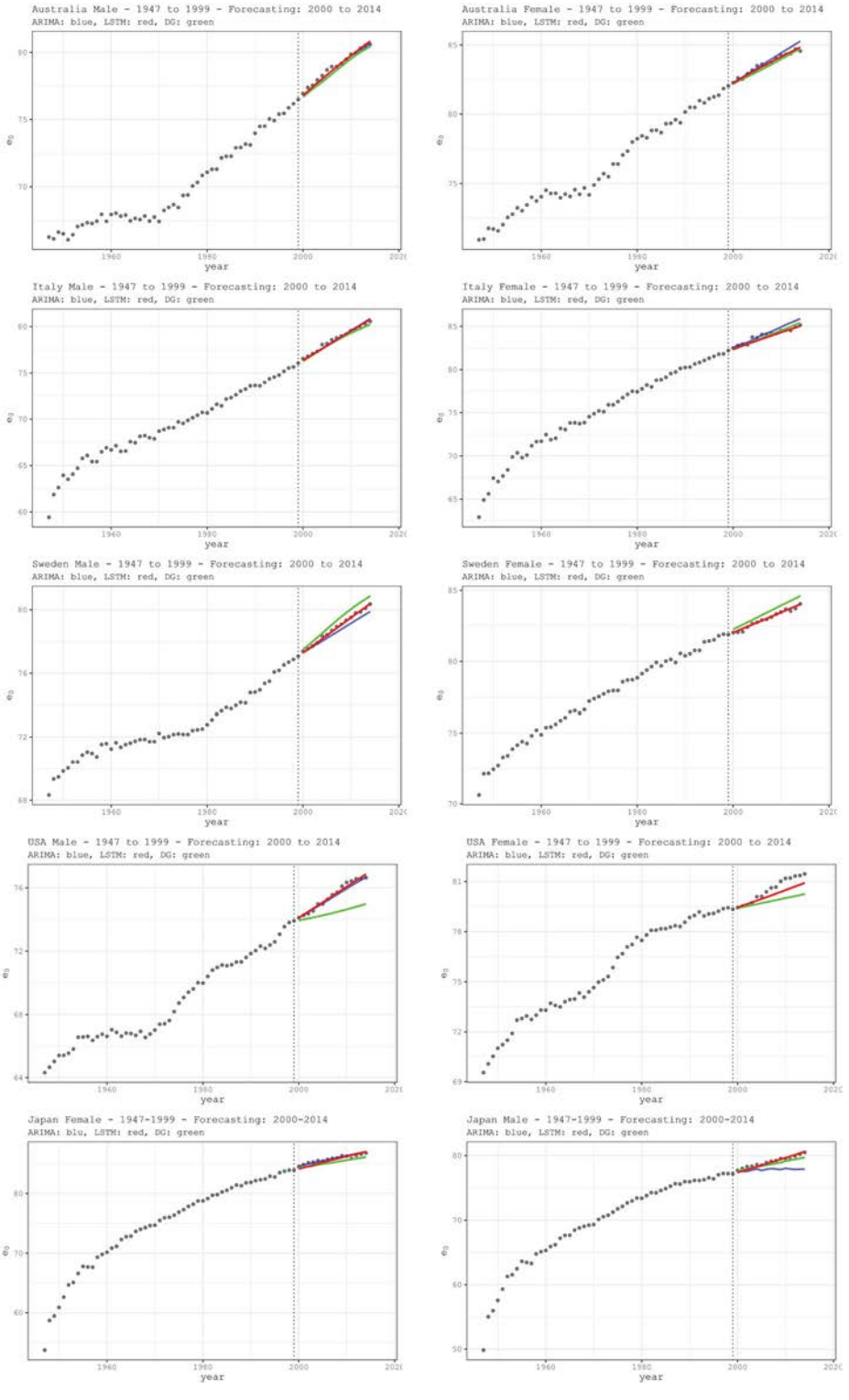

Figure 2. Historical and forecasted values of $e_{0,t}$ by country and gender (females on the left, males on the right).

Table 2. Out-of-sample test for $e_{65,t}$: MAE and RMSE for LSTM, ARIMA, DG, LC and CoDa model by country and gender.

		Fitting period: 1938–1999				Fitting period: 1947–1999			
		Female		Male		Female		Male	
Country	Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	ARIMA	0.2928	0.3271	0.1501	0.1811	0.2277	0.2587	0.2846	0.3664
	DG	0.3205	0.3564	0.8552	0.9366	0.2817	0.3151	0.8163	0.8928
	LSTM	**0.0804**	**0.0999**	**0.0764**	**0.0998**	**0.0782**	**0.0975**	**0.0764**	**0.0996**
	LC	0.4688	0.4962	1.2422	1.3336	0.3583	0.3878	1.0972	1.189
	CoDa	0.1842	0.2143	0.9851	1.0834	0.1295	0.1569	0.8639	0.9467
Italy	ARIMA	0.2604	0.2954	1.0379	1.1296	0.2732	0.3059	1.0918	1.2212
	DG	0.2604	0.2954	0.5539	0.5946	0.2732	0.3059	0.5669	0.6100
	LSTM	**0.1578**	**0.1972**	**0.1529**	**0.1803**	**0.1591**	**0.2022**	**0.1576**	**0.1893**
	LC	0.4672	0.4936	1.4899	1.5551	0.3479	0.3798	1.2268	1.2999
	CoDa	0.2347	0.2765	0.7775	0.8372	0.2437	0.2878	0.8047	0.8681
Sweden	ARIMA	0.1361	0.1705	0.8900	0.9902	0.2786	0.3042	0.7178	0.8177
	DG	**0.1007**	0.1384	0.4020	0.4703	0.2786	0.3042	0.2381	0.2836
	LSTM	0.1058	0.1357	**0.0828**	**0.1015**	0.1147	0.1455	**0.0861**	**0.1032**
	LC	0.1095	**0.1248**	1.15	1.2278	**0.0541**	**0.0637**	0.9145	1.0011
	CoDa	0.2121	0.2575	0.8872	0.9903	0.1015	0.1181	0.7832	0.8718
USA	ARIMA	0.2529	0.2923	0.9051	1.0138	0.3112	0.3753	0.6753	0.7572
	DG	0.2616	0.2734	0.3081	0.3449	**0.1775**	**0.2047**	0.7755	0.8431
	LSTM	0.6146	0.7095	**0.2773**	**0.3109**	0.5283	0.6094	**0.2485**	**0.2963**
	LC	**0.2212**	**0.2512**	0.9987	1.093	0.2601	0.2979	0.9337	1.0345
	CoDa	0.1915	0.2226	0.9193	1.0245	0.2732	0.3267	0.8893	0.9908
Japan	ARIMA	–	–	–	–	0.2804	0.3762	**0.1815**	**0.2073**
	DG	–	–	–	–	**0.2590**	0.3287	0.3436	0.4494
	LSTM	–	–	–	–	0.2928	**0.3189**	0.2173	0.2392
	LC	–	–	–	–	0.5048	0.5775	0.3906	0.4240
	CoDa	–	–	–	–	0.5747	0.7193	0.4275	0.5262

Notes: Years 2000–2014. Fitting periods: 1938–1999 (columns 3–6) and 1947–1999 (columns 7–10).

the bivariate LSTM is compared to the VAR model. All these models are trained aiming at generating life expectancy and lifespan disparity projections on the testing set. The following goodness of fit measures are used to evaluate the forecasting quality:

- *Mean Absolute Error*

$$\mathbf{MAE} = \frac{\sum_{t=t_\tau+1}^{t_s} \mid e_{x,t} - \hat{e}_{x,t} \mid}{(t_s - t_\tau - 1)}, \tag{20}$$

- *Root Mean Square Error*

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{t=t_\tau+1}^{t_s} \left(e_{x,t} - \hat{e}_{x,t}\right)^2}{(t_s - t_\tau - 1)}}, \tag{21}$$

where $\hat{e}_{x,t}$ represents the future estimation of life expectancy produced by the models. These measures are also used to evaluate the forecasting of the lifespan disparity $e^{\dagger}_{x,t}$ and the bivariate series $[e_{x,t}, e^{\dagger}_{x,t}]$.

All the experiments were performed using the R packages: *keras* and *tensorflow* (version 1.13.1) for LSTM,[2] *forecast* for ARIMA, *MortalityGap* for DG model, *MortalityForecast* for CoDA model, *StMoMo* for Lee-Carter model and *vars* for VAR model.

[2] The program that was used to implement the LSTM can be obtained from: https://doi.org/10.5281/zenodo.3999994.

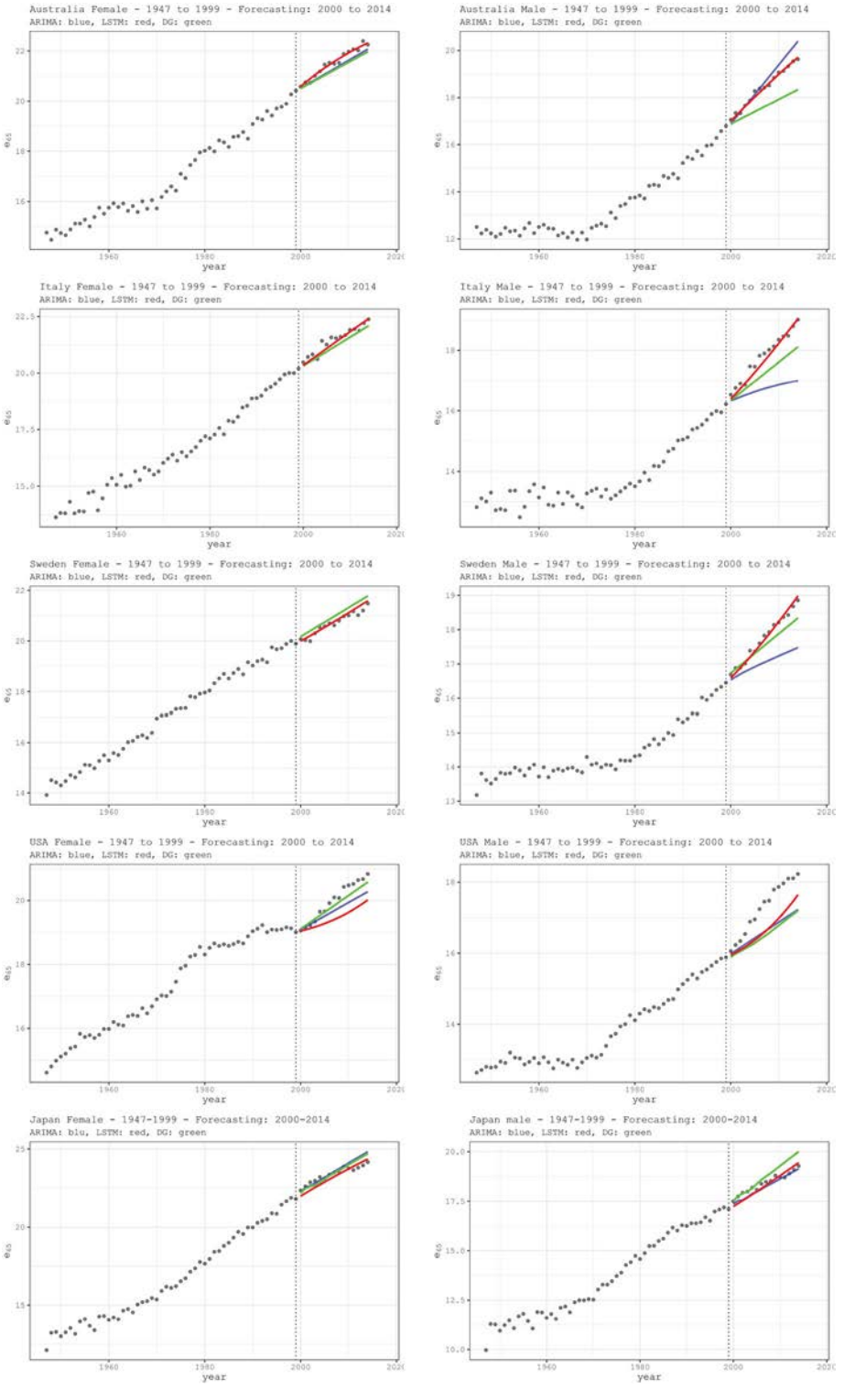

Figure 3. Historical and forecasted values of $e_{65,t}$ by country and gender (females on the left, males on the right).

5.1. Results of the out-of-sample test: independent modeling

This section will provide the estimated future life expectancy and lifespan disparity at birth and age 65 from separate modeling. The results are provided for five countries (Australia, Italy, Japan, Sweden, and USA) and both genders over the testing period. As already pointed out in the introduction, mortality models should also satisfy biological reasonableness criteria, with respect to both the short and the long term dynamics must be biologically consistent. Hence, we will perform the considered models on two different time windows, carrying out a sensitivity analysis based on two periods according to the historical demographic changes. The longest period (1938–1999) covers the World War II mortality shocks, which is excluded in the shortest one (1947–1999). Japan is not considered in the period starting from 1938 as data were made available starting from 1947.

5.1.1. Results for life expectancy: $e_{0,t}$ and $e_{65,t}$

Table 1 shows MAE and RMSE values of life expectancy at birth for both the estimation periods and each country by gender. Overall, the univariate LSTM provides remarkably high accuracy compared to the other models, overperforming in 72% of cases. Our model is only beaten in case of Japan females, Italy, and US females for both periods, however reaching the second-best performance. The historical and forecasted values of e_0 by country and gender are illustrated in Figure 2. We generally observe that when life expectancy does not experience any trend changes, the reduction of mortality compression does not provide any evidence of imminent interruption (Bohk-Ewald et al. 2017) as detected by lifespan disparity whose results are shown in Figure 4.

The results of the backtesting exercise for life expectancy at age 65 are reported in Table 2 for both the estimation periods and each country by gender. Also, by graphical analysis, the univariate LSTM seems to well catch the nonlinearity of the future mortality trend, showing its aptitude to better represent the decreasing dynamics of mortality at age 65. In this case, our model overperforms all the other models in 69% of cases. The historical and forecasted values of e_{65} by country and gender are illustrated in Figure 3. Overall, e_{65} shows a nonlinear behavior and irregular patterns, especially for

Table 3. Out-of-sample test for $e^{\dagger}_{0,t}$: MAE and RMSE for LSTM, ARIMA, LC and CoDa model by country and gender.

		Fitting period: 1938–1999				Fitting period: 1947 – 1999			
		Female		Male		Female		Male	
Country	Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	ARIMA	0.0916	0.0985	0.1058	0.1348	**0.0718**	0.0994	0.1426	0.1618
	LSTM	0.0906	0.1018	**0.0631**	**0.0756**	0.0794	0.0929	0.0880	0.1103
	LC	**0.0729**	0.0949	0.0844	0.1150	0.1931	0.2041	**0.0764**	**0.0981**
	CoDa	0.0757	**0.0864**	0.2147	0.2270	0.0845	**0.0924**	0.1601	0.1785
Italy	ARIMA	0.3209	0.3709	0.9104	1.0565	0.5013	0.5810	0.3444	0.3955
	LSTM	**0.0545**	**0.0643**	**0.0702**	**0.0866**	**0.1362**	**0.1528**	**0.0646**	**0.0827**
	LC	0.2100	0.2222	0.3984	0.4641	0.1649	0.1792	0.3860	0.4513
	CoDa	0.2451	0.2833	0.2807	0.3247	0.2331	0.2702	0.3533	0.4073
Sweden	ARIMA	0.2438	0.2666	0.3020	0.3390	0.2944	0.3262	0.2166	0.2442
	LSTM	**0.0598**	**0.0736**	**0.0468**	**0.0550**	**0.0572**	**0.0669**	**0.0439**	**0.0565**
	LC	0.1204	0.1291	0.0559	0.0734	0.1025	0.1126	0.1798	0.1955
	CoDa	0.2195	0.2398	0.0612	0.0729	0.2379	0.2634	0.0771	0.0920
USA	ARIMA	0.5569	0.6499	0.8677	0.9733	0.4795	0.5508	0.5935	0.6626
	LSTM	**0.0457**	**0.0547**	**0.0497**	**0.0603**	**0.0517**	**0.0561**	**0.0529**	**0.0628**
	LC	0.1670	0.2006	0.3277	0.3742	0.3281	0.3514	0.1659	0.1931
	CoDa	0.3269	0.3885	0.4529	0.5027	0.1970	0.2433	0.4246	0.4715
Japan	ARIMA	–	–	–	–	0.0635	0.0868	0.2265	0.2863
	LSTM	–	–	–	–	**0.0573**	**0.0760**	**0.0726**	**0.0799**
	LC	–	–	–	–	1.1321	1.1351	0.7256	0.7276
	CoDa	–	–	–	–	0.9129	0.9958	0.5617	0.6017

Notes: Years 2000–2014. Fitting periods: 1938–1999 (columns 3–6) and 1947–1999 (columns 7–10).

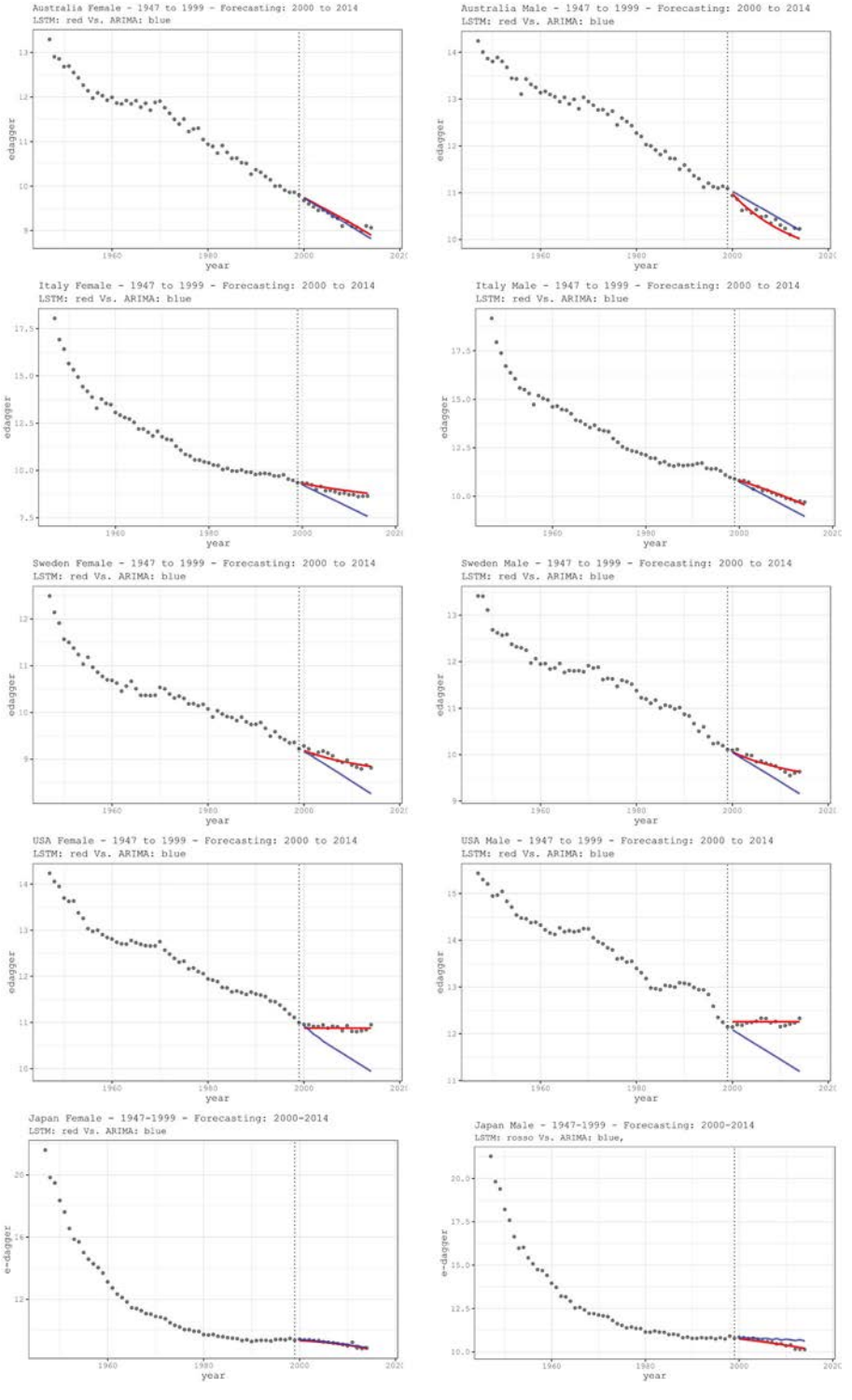

Figure 4. Historical and forecasted values of $e^{\dagger}_{0,t}$ by country and gender (females on the left, males on the right).

males, and the gain provided by LSTM is more evident if compared to the other models. Indeed, one of the main features of LSTM is to reproduce in the projections the irregular patterns of a phenomenon observed in the past. In particular, in the case of US females, we speculate that the historical periods 1973–1979 and 1989–1992 seem to strongly affect the LSTM weights, by reproducing in the forecasts the sudden longevity growth after the stagnation following the World War II.

5.1.2. Results for lifespan disparity: $e^{\dagger}_{0,t}$ and $e^{\dagger}_{65,t}$

The results of the out-of-sample test for $e^{\dagger}_{0}$ are shown in Table 3 for both the estimation periods and each country by gender. We can observe that for lifespan disparity at birth, the univariate LSTM outperforms the other models in 83% of the cases. Our model does not reach the best performance only for Australia females for both periods, where however, the prediction errors are incredibly low. The most remarkable out-of-sample result for $e^{\dagger}_{0}$ is provided by US females. Such a result shows a decreasing trend periodically interrupted by stagnation periods. In this case, the LSTM weights are probably influenced by the two short periods of stagnation, 1960–1970 and 1985–1990, that are reproduced in the projections, allowing to reach a high level of accuracy (see Figure 4). The same speculation holds for US males where the stagnation periods are more evident. We assume that a similar forecast behavior is challenging to be achieved by a canonical model that could ignore the long-short term dynamics.

The results for $e^{\dagger}_{65}$ are shown in Table 4 for both the estimation periods and each country by gender. The MAE and RMSE values highlight the LSTM ability to detect the hidden patterns of noisy time series, outperforming the other models in 89% of the cases. Indeed, $e^{\dagger}_{65}$ is characterized by a high variability level, since it summarizes disparity across individuals who have already survived to age 65. In case of US male (Figure 5), the LSTM prediction is not consistent with the historical values and might be influenced by short-term stagnation dynamics.

Table 4. Out-of-sample test for $e^{\dagger}_{65,t}$: MAE and RMSE for LSTM, ARIMA, LC and CoDa model by country and gender.

		Fitting period: 1938–1999				Fitting period: 1947–1999			
		Female		Male		Female		Male	
Country	Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	ARIMA	0.2077	0.2265	0.1743	0.1883	0.2119	0.2304	0.0905	0.1048
	LSTM	**0.0466**	**0.0525**	**0.0399**	**0.0525**	**0.0435**	**0.0539**	**0.0469**	**0.0596**
	LC	0.3093	0.3176	0.1887	0.1962	0.2895	0.2976	0.1354	0.1450
	CoDa	0.1148	0.1285	0.0771	0.0932	0.1128	0.1263	0.0646	0.0789
Italy	ARIMA	0.1284	0.1499	0.2735	0.3149	0.1405	0.1636	0.2755	0.3111
	LSTM	**0.0505**	**0.0605**	**0.0441**	**0.0583**	**0.0498**	**0.0595**	**0.0425**	**0.0532**
	LC	0.0988	0.1154	0.4352	0.4586	0.0727	0.0856	0.3720	0.3943
	CoDa	0.0566	0.0743	0.2217	0.2633	0.0659	0.0841	0.2174	0.2592
Sweden	ARIMA	0.0733	0.0836	0.1412	0.1548	0.1379	0.1494	0.1408	0.1554
	LSTM	**0.0286**	**0.0342**	**0.0307**	**0.0388**	**0.0290**	**0.0353**	**0.0308**	**0.0388**
	LC	0.0705	0.0788	0.2767	0.2831	0.0541	0.0637	0.2316	0.2377
	CoDa	0.0814	0.0937	0.0668	0.0821	0.1015	0.1182	0.0701	0.0860
USA	ARIMA	0.0733	0.0826	**0.1033**	**0.1341**	0.0599	0.0693	**0.1074**	**0.1411**
	LSTM	**0.0439**	**0.0539**	0.1221	0.1613	**0.0446**	**0.054**	0.1153	0.1532
	LC	0.1872	0.1948	0.1673	0.1879	0.2361	0.2407	0.1295	0.1510
	CoDa	0.0634	0.0860	0.1092	0.1462	0.0470	0.0585	0.1158	0.1543
Japan	ARIMA	–	–	–	–	0.0896	0.1050	0.1923	0.2363
	LSTM	–	–	–	–	**0.0647**	**0.0765**	**0.0773**	**0.0912**
	LC	–	–	–	–	0.2086	0.2174	0.1232	0.1519
	CoDa	–	–	–	–	0.3542	0.3695	0.1402	0.1461

Notes: Years 2000–2014. Fitting periods: 1938–1999 (columns 3–6) and 1947-1999 (columns 7–10).

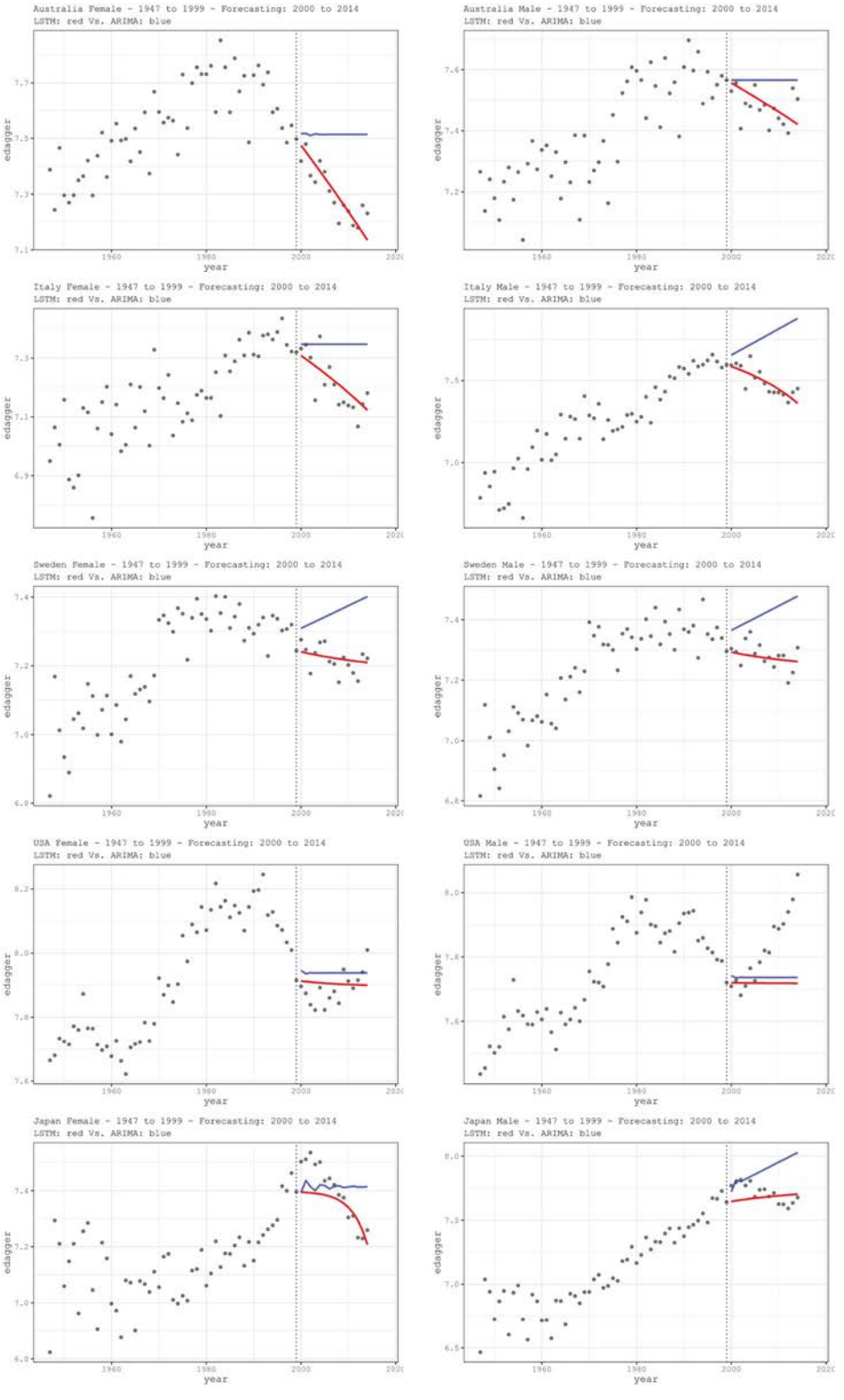

Figure 5. Historical and forecasted values of $e^{\dagger}_{65,t}$ by country and gender (females on the left, males on the right).

5.2. Results of the out-of-sample test: simultaneous modeling

The estimates of future life expectancy and lifespan disparity at birth and age 65 given by the out-of-sample test, resulting from the simultaneous modeling (namely LSTM-2D) are shown in the following tables, compared with the first-order VAR model that is used as a benchmark for multivariate series forecasting. The results for e_0 and e_{65} are respectively reported in Tables 5 and 6 for both the estimation periods and each country by gender. We note that the LSTM-2D outperforms the VAR model for life expectancy at birth in 86% of the cases, while this percentage drops to 47% at age 65. Similar behavior can be observed for lifespan disparity (Tables 7 and 8), where LSTM-2D obtains the best performance in 78% of the cases at birth and 58% at age 65. In some few cases, the bivariate network provides lower errors if compared to the other models (univariate and bivariate), especially for life expectancy at birth: for example, Italy females in the fitting period 1947–1999, Sweden males in the fitting period 1938–1999 and for e_{65} Japan females in the fitting period 1947–1999.

Our empirical analysis shows that the simultaneous modeling of life expectancy and lifespan disparity may be not suitable, however, it leads us to speculate that only life expectancy at birth projections may take advantage of a simultaneous forecasting with life disparity.

Table 5. Out-of-sample test for $e_{0,t}$: MAE and RMSE for LSTM-2D and VAR, by country and gender.

		Fitting period: 1938–1999				Fitting period: 1947-1999			
		Female		Male		Female		Male	
Country	Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	LSTM-2D	**0.1331**	**0.1674**	**0.4488**	**0.5152**	**0.1497**	**0.1916**	0.5602	0.6343
	VAR	0.6999	0.7269	0.8535	0.8658	0.2134	0.2520	**0.4687**	**0.5386**
Italy	LSTM-2D	**0.2442**	**0.3019**	**0.4417**	**0.4991**	**0.1235**	**0.1646**	**0.1392**	**0.1654**
	VAR	0.2957	0.3409	2.4970	2.6610	0.2957	0.3409	2.3200	2.5000
Sweden	LSTM-2D	**0.2437**	**0.2851**	**0.0488**	**0.0605**	**0.0909**	**0.1180**	**0.1001**	**0.1176**
	VAR	0.8205	0.9675	1.9560	2.1530	0.4442	0.5284	0.8585	0.9164
USA	LSTM-2D	**0.1786**	**0.2257**	0.2413	0.2754	**0.1216**	**0.1488**	**0.1960**	**0.2262**
	VAR	0.5654	0.6964	**0.1471**	**0.1824**	0.6170	0.7516	0.2121	0.2993
Japan	LSTM-2D	–	–	–	–	**0.3225**	0.3734	**0.5602**	**0.6343**
	VAR	–	–	–	–	0.3320	**0.3685**	1.2200	1.3300

Notes: Years 2000–2014. Fitting periods: 1938–1999 (columns 3–6) and 1947–1999 (columns 7–10).

Table 6. Out-of-sample test for $e_{65,t}$: MAE and RMSE for LSTM-2D and VAR, by country and gender.

		Fitting period: 1938–1999				Fitting period: 1947–1999			
		Female		Male		Female		Male	
Country	Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	LSTM-2D	**0.0986**	**0.1144**	**0.3142**	**0.3618**	**0.0826**	**0.1127**	**0.2934**	**0.3339**
	VAR	1.1765	1.3661	1.2694	1.6620	1.1546	1.3594	1.2624	1.6553
Italy	LSTM-2D	0.4061	0.4615	**0.3784**	**0.4227**	0.6179	0.6510	**0.4195**	**0.4622**
	VAR	**0.2985**	**0.3436**	1.3782	1.4669	**0.2219**	**0.2586**	1.2280	1.2955
Sweden	LSTM-2D	0.6780	0.7144	0.4292	0.4767	0.5264	**0.5452**	0.6009	0.6173
	VAR	**0.6617**	**0.7070**	**0.2310**	**0.2425**	**0.5137**	0.5524	**0.3113**	**0.4375**
USA	LSTM-2D	0.3981	0.4398	0.6644	0.7180	0.8393	0.9477	0.8614	0.9118
	VAR	**0.2985**	**0.3436**	**0.1725**	**0.2435**	**0.3174**	**0.3729**	**0.3627**	**0.3983**
Japan	LSTM-2D	–	–	–	–	**0.1317**	**0.1365**	**0.2139**	**0.2490**
	VAR	–	–	–	–	0.4882	0.5486	0.6607	0.6680

Notes: Years 2000–2014. Fitting periods: 1938–1999 (columns 3–6) and 1947–1999 (columns 7–10).

Table 7. Out-of-sample test for $e^{\dagger}_{0,t}$: MAE and RMSE for LSTM-2D and VAR, by country and gender.

		Fitting period: 1938–1999				Fitting period: 1947–1999			
		Female		Male		Female		Male	
Country	Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	LSTM-2D	0.1247	0.1363	**0.1406**	**0.1620**	0.1450	0.1569	**0.1846**	**0.2054**
	VAR	**0.0929**	**0.1027**	0.1627	0.2000	**0.1084**	**0.1150**	0.2076	0.2670
Italy	LSTM-2D	**0.2178**	**0.2416**	0.5282	0.5353	**0.2840**	**0.3237**	**0.0630**	**0.0782**
	VAR	0.7315	0.7936	**0.0830**	**0.1004**	0.7315	0.7936	0.2318	0.2737
Sweden	LSTM-2D	**0.1868**	**0.2190**	**0.0393**	**0.0445**	**0.0950**	**0.1194**	**0.0584**	**0.0701**
	VAR	0.3118	0.3414	0.1111	0.1371	0.2304	0.2496	0.2177	0.2375
USA	LSTM-2D	**0.1192**	**0.1436**	**0.2004**	**0.2306**	**0.1199**	**0.1406**	**0.2228**	**0.2615**
	VAR	0.1478	0.1557	0.5467	0.6241	0.2365	0.2424	0.5103	0.5956
Japan	LSTM-2D	–	–	–	–	0.2758	0.2902	**0.0875**	**0.0995**
	VAR	–	–	–	–	**0.1818**	**0.2353**	0.0897	0.1107

Notes: Years 2000–2014. Fitting periods: 1938–1999 (columns 3–6) and 1947–1999 (columns 7–10).

Table 8. Out-of-sample test for $e^{\dagger}_{65,t}$: MAE and RMSE for LSTM-2D and VAR, by country and gender.

		Fitting period: 1938–1999				Fitting period: 1947–1999			
		Female		Male		Female		Male	
Country	Model	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	LSTM-2D	**0.1781**	**0.1875**	**0.3494**	**0.3727**	**0.1628**	**0.1713**	**0.2981**	**0.3175**
	VAR	0.5819	0.6067	0.5209	0.5642	0.5427	0.5669	0.4879	0.5262
Italy	LSTM-2D	**0.2813**	0.3151	0.5826	0.6568	**0.2225**	**0.2483**	0.4985	0.5704
	VAR	0.2895	**0.2940**	**0.1536**	**0.1842**	0.2695	0.2990	**0.1417**	**0.1694**
Sweden	LSTM-2D	**0.1042**	**0.1118**	0.2810	0.3086	**0.1074**	**0.1151**	0.2179	0.2393
	VAR	0.2735	0.2841	**0.2687**	**0.2943**	0.2519	0.2624	**0.1438**	**0.1645**
USA	LSTM-2D	**0.1459**	**0.1582**	0.2301	0.2328	**0.1476**	**0.1574**	0.1277	0.1346
	VAR	0.2895	0.2940	**0.0667**	**0.0788**	0.2735	0.2783	**0.1048**	**0.1310**
Japan	LSTM-2D	–	–	–	–	**0.1174**	**0.1131**	0.1348	0.1573
	VAR	–	–	–	–	0.1442	0.1281	**0.1210**	**0.1384**

Notes: Years 2000–2014. Fitting periods: 1938–1999 (columns 3–6) and 1947–1999 (columns 7–10).

6. Conclusions

Mortality improvements are linked to social progress in terms of health, nutrition, education, hygiene, and medicine (Riley 2001). Thus, the lack of a model capable of adequately represent the future trend of life expectancy is evident. Indeed, a more plausible forecasting has a real social-economic impact, considering the nature of life expectancy. Life expectancy is not merely a time-trend index, but rather a 'latent factor'incorporating different unobserved latent variables. It implicitly encompasses economic fluctuations, medical innovation and many other variables that directly (or indirectly) influenced the mortality trend.

Our paper proposes a new approach based on LSTM neural network to forecast longevity indexes both independently and simultaneously at birth and age 65, catching either short and long term factors on mortality improvements. As for the LSTM applied to life expectancy, we observe that without imposing model restrictions (like the gender gap and the BPLE), we can obtain predictions coherent with historical trends and biological criteria. The univariate LSTM outperforms all the models analyzed (ARIMA, DG, LC and CoDa), especially for life expectancy at age 65, where e.g.the BPLE shows some weaknesses as the linear assumption.

A substantial part of this work has been devoted to forecasting future values of lifespan disparity, for which, to the best of our knowledge, the literature has not provided any contribution yet.

The wide discussion in the literature on the relationship between life expectancy and lifespan disparity suggests that projections of life expectancy and lifespan disparity may benefit from simultaneous forecasting. Accordingly, we introduce a bivariate LSTM, which represents a novelty in the demographic panorama, by simultaneously forecasting life expectancy and lifespan disparity in the RNN framework. Our simultaneous model obtains higher levels of accuracy compared to the first-order VAR model used as a benchmark for multivariate series forecasting. Our empirical analysis, based on five countries, two fitting periods and both genders, shows that the simultaneous forecasting of life expectancy and lifespan disparity is less adequate than independent modeling. Nevertheless, our results lead to speculating that only life expectancy at birth projections take advantage of simultaneous forecasting with life disparity. Extrapolative models, e.g. the Lee-Carter model, may also benefit from a parameter adjustment consistent not only with lifespan disparity as in Rabbi & Mazzuco (2020) but with both observed life expectancy and life disparity.

We show that both independent and simultaneous forecasts of life expectancy and lifespan disparity provide new insights for a comprehensive evaluation of the mortality forecasts, representing a useful tool to capture irregular mortality trajectories. Our findings support the decrease of lifespan disparity among developed countries, for which the evolution of age-at-death distribution assumes more compressed tails over time. Besides, our approach based on the long-short term enables to consider the entire time series, without excluding short-term shocks from the analysis. Using two different periods, we show that the LSTM provides robust forecasts to the unexpected mortality changes. This aspect sounds coherent with the *modus operandi* behind the LSTM architecture, where the neuron cell manages the time series noise, combining the long and short-term past information. Unfortunately, LSTM does not make explicit how the algorithm handles the long and short-term trade-off.

Since the main purpose of this paper is to focus on the LSTM predictive ability to capture the central trend of life expectancy and lifespan disparity, we do not assess the prediction intervals. Confidence intervals can be provided using Bayesian neural networks, which present a different structure from our model. In our case, we could estimate prediction intervals through an iterative simulation in the spirit of MCMC methods, but as mentioned above, this is beyond the scope of our investigation.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

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

Aburto, J. M., Wensink, M., van Raalte, A. & Lindahl-Jacobsen, R. (2018). Potential gains in life expectancy by reducing inequality of lifespans in Denmark: an international comparison and cause-of-death analysis. *BMC Public Health* **18**(1), 831.

Aburto, J. M., Villavicencio, F., Basellini, U., Kjærgaard, S. & Vaupel, J. W. (2020). Dynamics of life expectancy and life span equality. *PNAS* **117**(10), 5250–5259.

Alho, J. M. (1990). Stochastic methods in population forecasting. *International Journal of Forecasting* **6**(4), 521–530.

Colchero, F., Rau, R., Jones, O. R., Barthold, J. A., Conde, D. A., Lenart, A., Nemeth, L., Scheuerlein, A., Schoeley, J., Torres, C. & Zarulli, V. (2016). The emergence of longevous populations. *Proceedings of the National Academy of Sciences (PNAS)* **113**(48), E7681–E7690.

Bergeron-Boucher, M.-P., Canudas-Romo, V., Oeppen, J. & Vaupel, J. W. (2017). Coherent forecasts of mortality with compositional data analysis. *Demographic Research* **37**, 527–566.

Brouhns, N., Denuit, M. & Vermunt, J. (2002). A Poisson log-bilinear approach to the construction of projected life tables. *Insurance: Mathematics and Economics.* **31**, 373–393.
Bengio, Y., Simard, P. & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* **5**(2), 157–166.
Bohk-Ewald, C., Ebeling, M. & Rau, R. (2017). Lifespan disparity as an additional indicator for evaluating mortality forecasts. *Demography* **54**, 1559. doi: 10.1007/s13524-017-0584-0
Booth, H., Hyndman, R. J., Tickle, L. & De Jong, P. (2006). Lee-carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research* **15**, 289–310.
Brouhns, N., Denuit, M. & Vermunt, J. K. (2002). A Poisson log-bilinear approach to the construction of projected life tables. *Insurance: Mathematics and Economics* **31**, 373–393.
Cairns, A. J. G., Blake, D. & Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance* **73**, 687–718.
Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A. & Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal,* **13**, 1–35.
Currie, I. D., Durban, M. & Eilers, P. H. C. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling* **4**, 279–298.
Edwards, R. D. & Tuljapurkar, S. (2005). Inequality in life spans and a new perspective on mortality convergence across industrialized countries. *Population and Development Review* **31**(4), 645–674.
Elman, J. (1990). Finding structure in time. *Cognitive Science* **14**(2), 179–211.
Gabrielli, A., Richman, R. & Wüthrich, M. (2020). Neural network embedding of the over-dispersed Poisson reserving model. *Scandinavian Actuarial Journal* **1**, 1–29. DOI: 10.1080/03461238.2019.1633394
Gers, F. A. & Schmidhuber, J. (2000). Recurrent nets that time and count. Paper presented at IEEE-INNS-ENNS International Joint Conference on Neural Networks, Como, Italy, 24–27 July. P. 189–194.
Gers, F. A., Schmidhuber, J. & Cummins, F. (1999). Learning to forget: Continual prediction with LSTM. *Artificial Neural Networks* **470**, 850–855.
Glorot, X., Bordes, A. & Bengio, Y. (2011). Deep sparse rectifier neural networks. In Gordon, G., Dunson, D., & Dudk, M. (eds.), *JMLR W&CP: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*, 15. P. 315–323.
Goldman, N. & Lord, G. (1986). A new look at entropy and the life table. *Demography* **23**, 275–282.
Graves, A., Liwicki, M., Bunke, H. & Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *2009 Transactions on Pattern Analysis and Machine Intelligence* **31**(5), 855–868.
Hiam, L., Harrison, D., McKee, M. & Dorling, D. (2018). Why is life expectancy in England and Wales 'stalling'? *Journal of Epidemiology and Community Health* **72**(5), 404–408.
Ho, J. Y. & Hendi, A. S. (2018). Recent trends in life expectancy across high income countries: retrospective observational study. *BMJ* **362**, k2562.
Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation* **9**, 1735–1780.
Human Mortality Database. (2018). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Data downloaded on 1 January 2019. Available at: https://www.mortality.org
Kaakai, S., Hardy, H. L., Arnold, S. & El Karoui, N. (2019). How can a cause-of-death reduction be compensated for by the population heterogeneity? A dynamic approach. *Insurance: Mathematics and Economics* **89**, 16–37.
Keskar, N. S., Mudigere, D. & Nocedal, J. (2016). On large-batch training for deep learning: generalization gap and sharp minima. arXiv preprint arXiv:1609.04836 [cs.LG].
LeCun, Y. & Muller, K. R. (2012). Efficient backprop. In *Neural Networks: Tricks of the Trade*. Berlin: Springer. P. 9–48.
Lee, R. D. (2006). Mortality forecasts and linear life expectancy trends. In *Perspectives on Mortality Forecasting*. Social Insurance Studies, 3. The Linear Rise in Life Expectancy: History and Prospects.
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.
Nemeth, L. (2017). Life expectancy versus lifespan inequality: a smudge or a clear relationship?. *PloS One* **12**(9), e0185702.
Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S. & Perla, F. (2019). A deep learning integrated Lee-Carter model. *Risks* **7**(1), 33.
Oeppen, J. (2008). Coherent forecasting of multiple-decrement life tables: a test using Japanese cause of death data. Compositional Data Analysis Conference.
Oeppen, J. & Vaupel, J. W. (2002). Broken limits to life expectancy. *Science (New York, N.Y.)* **296**(5570), 1029–1031.
Oeppen, J. & Vaupel, J. W. (2006). The linear rise in the number of our days. Social Insurance Studies, 3. The Linear Rise in Life Expectancy: History and Prospects.
Pascanu, R., Tomas Mikolov, T. & Bengio, Y. (2013). On the difficulty of training recurrent neural networks.

Pascariu, M. D., Canudas-Romo, V. & Vaupel, W. J. (2018). The double-gap life expectancy forecasting model. *Insurance: Mathematics and Economics,* **78**, 339–350.
Rabbi, A. M. F. & Mazzuco, S. (2020). Mortality forecasting with the Lee-Carter method: adjusting for smoothing and lifespan disparity. *European Journal of Population* 1–24. doi:10.1007/s10680-020-09559-9
Raftery, A. E., Chunn, J. L., Gerland, P. & Ševcíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography* **50**(3), 777–801.
Rau, R., Soroko, E., Jasilionis, D. & Vaupel, J. W. (2008). Continued reductions in mortality at advanced ages. *Population and Development Review* **34**, 747–768.
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. (2019). A neural network extension of the Lee-Carter model to multiple populations. *Annals of Actuarial Science*, 1–21. doi:10.1017/S1748499519000071
Richman, R. & Wüthrich, M. (2019b). Lee and Carter go machine learning: recurrent neural networks. Version: August 22, 2019. Available at SSRN: https://ssrn.com/abstract = 3441030
Riley, J. (2001). *Rising life expectancy: A global history*. Cambridge: Cambridge University Press.
Rojas, R. & Feldman, J. (1996). *Neural networks: A systematic introduction*. Heidelberg: Springer.
Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by backpropagating errors. *Nature* **323**(6088), 533–536.
Shkolnikov, V. M., Andreev, E. M. & Begun, A. Z. (2003). Gini coefficient as a life table function: computation from discrete data, decomposition of differences and empirical examples. *Demographic Research* **8**(article 11), 305–358.
Torri, T. & Vaupel, J. W. (2012). Forecasting life expectancy in an international context. *International Journal of Forecasting* **28**(2), 519–531.
van Raalte, A. A. & Caswell, H. (2013). Perturbation analysis of indices of lifespan variability. *Demography* **50**, 1615–1640.
van Raalte, A., Sasson, I. & Martikainen, P. (30 Nov 2018). The case for monitoring life-span inequality. *Science* 362(6418), 1002–1004.
Vaupel, J. W. (1986). How change in age-specific mortality affects life expectancy. *Population Studies* **40**, 147–157.
Vaupel, J. W. (1997). The remarkable improvements in survival at older ages. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* **352**, 1799–1804.
Vaupel, J. W. & Canudas-Romo, V. (2003). Decomposing change in life expectancy: a bouquet of formulas in honor of Nathan Keyfitz's 90th birthday. *Demography* **40**(2), 201–216.
Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks* **1**(4), 339–356.
Wilmoth, J. R. & Horiuchi, S. (1999). Rectangularization revisited: variability of age at death within human populations. *Demography* **36**(4), 475–495.
Wüthrich, M. (2019). From generalized linear models to neural networks, and back. Version: December 11, 2019. Available at SSRN: https://ssrn.com/abstract = 3491790
Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. Available at: http://arxiv.org/abs/1212.5701

Appendices

Appendix 1. Learning optimization

The estimation of the network parameters is obtained by minimizing the overall loss function, $L(\hat{\boldsymbol{y}}, \boldsymbol{y})$, in Equation (9). The first step is therefore differentiating the loss function with respect to the weights $\boldsymbol{W}$:

$$\frac{\partial L}{\partial \boldsymbol{W}} = \sum_{t=1}^{T} \frac{\partial L_t}{\partial \boldsymbol{W}}. \tag{A1}$$

For each time step, the updating process is founded on the following rule:

$$\frac{\partial L_t}{\partial \boldsymbol{W}} = \frac{\partial L_t}{\partial \hat{\boldsymbol{y}}_t} \frac{\hat{\boldsymbol{y}}_t}{\partial \boldsymbol{h}_t} \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{W}}. \tag{A2}$$

By multiplying and dividing by $\boldsymbol{h}_t$ and after some algebra, we obtain:

$$\frac{\partial L_t}{\partial \boldsymbol{W}} = \sum_{i=1}^{t} \frac{\partial L_t}{\partial \hat{\boldsymbol{y}}_t} \frac{\hat{\boldsymbol{y}}_t}{\partial \boldsymbol{h}_t} \prod_{j=i}^{t-1} \frac{\partial \boldsymbol{h}_{j+1}}{\partial \boldsymbol{h}_j} \frac{\partial \boldsymbol{h}_i}{\partial \boldsymbol{W}}. \tag{A3}$$

Hence, the partial derivative of the overall loss function with respect to the weights' matrix $\boldsymbol{W}$ is given by:

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{W}} &= \sum_{t=1}^{T}\sum_{i=1}^{t}\frac{\partial L_t}{\partial \hat{\boldsymbol{y}}_t}\frac{\hat{\boldsymbol{y}}_t}{\partial \boldsymbol{h}_t}\prod_{j=i}^{t-1}\frac{\partial \boldsymbol{h}_{j+1}}{\partial \boldsymbol{h}_j}\frac{\partial \boldsymbol{h}_i}{\partial \boldsymbol{W}} = \\ &= \sum_{t=1}^{T}\sum_{i=1}^{t}\frac{\partial L_t}{\partial \hat{\boldsymbol{y}}_t}\frac{\hat{\boldsymbol{y}}_t}{\partial \boldsymbol{h}_t}\prod_{j=i}^{t} diag\left(\phi'\left(\boldsymbol{W}_{hh}^T\boldsymbol{h}_j + \boldsymbol{W}_{xh}^T\boldsymbol{x}_j + \boldsymbol{b}_h\right)\right)\boldsymbol{W}_{hh}\frac{\partial \boldsymbol{h}_i}{\partial \boldsymbol{W}}\end{aligned}, \tag{A4}$$

where ϕ' is the first derivative of the hyperbolic tangent activation function in the hidden layer $\boldsymbol{h}$. Therefore, applying the delta rule, the following weights updating holds:

$$\Delta \boldsymbol{W} = -\eta\frac{\partial L}{\partial \boldsymbol{W}}, \tag{A5}$$

where η is learning rate, a hyper-parameter of the network that we can choose arbitrarily.[3] The forward-backward process can be repeated more times. The total number of times exploring the entire training dataset in the forward-backward sense is called epoch, representing another hyper-parameter of the network to be tuned. Therefore, to minimize the loss function we will update $d = 1, \ldots, N$ times the weights as in Equation (A5), where N is the maximum number of epochs before falling into overfitting. Since RNN suffers from the so called vanishing or exploding gradient (Bengio et al. 1994, Pascanu et al. 2013), the gradient dynamic is affected by the weights and the derivatives of the activation function[4] that the gradient passes through. Therefore, the vanishing (or exploding) gradient derives from the recursive derivative (Equation (A3)). When the time steps increase, the derivative in Equation (A3) is affected by short-term dependencies. Indeed, if the number of recurrences increases, the multiplication in Equation (A3) rapidly converges to 0 when the eigenvalues of the matrix in Equation (A4), λ_t, are less than 1. As a result, weights are less and fewer updated. This effect is called the 'vanishing gradient'. On the other hand, when eigenvalues λ_t are greater than 1, we obtain an opposite effect called 'exploding gradient'.

We can note that any change in the hidden state $\Delta \boldsymbol{h}_{j+1}$ has a multiplicative effect. This means that, if the largest eigenvalue is less than 1, the gradient will vanish, otherwise it explodes. Consider $\lambda_{\max}^{\langle \boldsymbol{h}\rangle}$ and $\lambda_{\max}^{\langle \boldsymbol{W}\rangle}$ the largest eigenvalues associated with $||diag(\phi'(\boldsymbol{W}_{hh}^T\boldsymbol{h}_j + \boldsymbol{W}_{xh}^T\boldsymbol{x}_j + \boldsymbol{b}_h))||$ and $||\boldsymbol{W}_{hh}||$, respectively. Then:

$$\left|\left|\frac{\partial \boldsymbol{h}_{j+1}}{\partial \boldsymbol{h}_j}\right|\right| \leq \left|\left|diag\left(\phi'\left(\boldsymbol{W}_{hh}^T\boldsymbol{h}_j + \boldsymbol{W}_{xh}^T\boldsymbol{x}_j + \boldsymbol{b}_h\right)\right)\right|\right| \cdot \left|\left|\boldsymbol{W}_{hh}\right|\right| \leq \lambda_{max}^{\langle \boldsymbol{h}\rangle}\lambda_{max}^{\langle \boldsymbol{W}\rangle}; \tag{A6}$$

$$\left|\left|\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_i}\right|\right| = \left|\left|\prod_{j=i}^{t-1}\frac{\partial \boldsymbol{h}_{j+1}}{\partial \boldsymbol{h}_j}\right|\right| \leq \left(\lambda_{max}^{\langle \boldsymbol{h}\rangle}\lambda_{max}^{\langle \boldsymbol{W}\rangle}\right)^{t-i}. \tag{A7}$$

As the sequence becomes longer (i.e the distance between t and i increases), the eigenvalues will determine if the gradient either becomes exceptionally large (explodes) or very small (vanishes).

As mentioned in Section 3.1, the vanishing problem has been overcome by managing the recurrent hidden units through the so called gates. Looking at the LSTM architecture, these gates allow to transform both the current input, $\boldsymbol{x}_t$, and the short term output, $\boldsymbol{h}_{t-1}$, in order to update the current memory unit information. The backward flow involved in the optimization within each LSTM block is computed as follows:

$$\frac{\partial \boldsymbol{c}_t}{\partial \boldsymbol{c}_{t-1}} = \frac{\partial \boldsymbol{c}_t}{\partial \boldsymbol{f}_t}\frac{\partial \boldsymbol{f}_t}{\partial \boldsymbol{h}_{t-1}}\frac{\partial \boldsymbol{h}_{t-1}}{\partial \boldsymbol{c}_{t-1}} + \frac{\partial \boldsymbol{c}_t}{\partial \boldsymbol{i}_t}\frac{\partial \boldsymbol{i}_t}{\partial \boldsymbol{h}_{t-1}}\frac{\partial \boldsymbol{h}_{t-1}}{\partial \boldsymbol{c}_{t-1}} + \frac{\partial \boldsymbol{c}_t}{\partial \boldsymbol{z}_t}\frac{\partial \boldsymbol{z}_t}{\partial \boldsymbol{h}_{t-1}}\frac{\partial \boldsymbol{h}_{t-1}}{\partial \boldsymbol{c}_{t-1}}. \tag{A8}$$

We notice that at each time step the algorithm back-propagates the error through both loss function and memory unit of the next time step. Hence, we observe that if the terms $\frac{\partial \boldsymbol{c}_t}{\partial \boldsymbol{c}_{t-1}}$ start to converge towards zero, we can set higher gates values to reach the value close to 1, thus preventing the gradients from vanishing.

Appendix 2. Other models

ARIMA model is a well-established approach that can be considered as the reference model for the forecast of mortality. This model has three parameters p, d, q, representing respectively the auto-regressive, the differencing and the moving

[3] Tipically $\eta \in (0, 1)$.

[4] For $f_h = \phi_h$, the domain of the first derivative is $0 < \phi_h' < 1$. For $f_y = \sigma_y$, the domain is $0 < f_y' < 1/4$.

average order. The generic ARIMA(p, d, q) for life expectancy takes the following form:

$$\prod^{d} e_{x,t} = \delta + \sum_{i=1}^{p} \phi_i \prod^{d} e_{x,t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}. \tag{A9}$$

where δ is the drift process, ϕ_i are the autoregressive parameters, ϵ_t the error terms normally distributed with zero mean and variance σ_ϵ^2 and θ_j are the moving average parameters.

Double Gap model is one of the most recent and most prominent approaches in forecasting life expectations. It provides the life expectancy forecasts for both the genders by modeling first the gap between country-specific female life expectancy, e^f, and female BPLE (the female world record level), e^{bp}, and then the gap between male life expectancy, e^m, and female life expectancy, e^f, in a given country. Therefore, the future female life expectancy at age x and time t for a given country is calculated as the difference between the future $e^{bp}_{x,t}$ and the predicted values of the gap, $D_{x,t}$, between the country-specific female life expectancy and the female best-practice trend: $e^f_{x,t} = e^{bp}_{x,t} - D_{x,t}$. While, the future male life expectancy is calculated as the difference between the future female life expectancy and the predicted values of the gap, $G_{x,t}$, between the country-specific female and male life expectancy: $e^m_{x,t} = e^f_{x,t} - G_{x,t}$. The first gap, $D_{x,t}$, is modeled according to a traditional ARIMA(p, d, q):

$$\prod^{d} D_{x,t} = \delta^{(1)} + \sum_{i=1}^{p} \phi_i^{(1)} \prod^{d} D_{x,t-i} + \epsilon_t^{(1)} + \sum_{j=1}^{q} \theta_j^{(1)} \epsilon_{t-j}^{(1)}, \tag{A10}$$

where $\delta^{(1)}$ is the drift process, $\phi_i^{(1)}$ are the autoregressive parameters, $\epsilon_t^{(1)}$ the error terms normally distributed with zero mean and variance $\sigma^2_{\epsilon^{(1)}}$ and $\theta_j^{(1)}$ are the moving average parameters. The second gap, $G_{x,t}$, is modeled by a linear model and a random walk without drift:

$$G^*_{x,t} = \begin{cases} \beta_0 + \beta_1 \cdot G_{x,t-1} + \beta_2 \cdot G_{x,t-2} + \beta_3 \cdot \left(e^f_{x,t} - \tau\right)^+ + \epsilon_t^{(2)} & \textit{if } e^f_{x,t} < A, \\ G_{x,t-1} + \epsilon_t^{(3)} & \textit{otherwise} \end{cases}$$

where τ and A are fixed levels calculated on historical data by maximizing the resulting maximum likelihoods of the linear model over integer values of τ and A (see Pascariu et al. 2018 for further details on the estimation procedure). The algorithm is implemented by the function available in the R package *MortalityGap*.

The DG model is not applied in the case of lifespan disparity due to the non-existence of a best practice for disparity measures.

Lee-Carter model works with the linear extrapolations of age-specific mortality rates on the logarithmic scale. Its first formulation (Lee & Carter 1992) based on the latent approach using SVD has been widely improved across time. We use the extension proposed by Brouhns et al. (2002) that exploits Poisson log-likelihood estimation.

$$\log(m_{\mathbf{s}}) = \alpha_x + \beta_x \kappa_t, \tag{A11}$$

where α_x is the age-specific parameter providing the average age profile of mortality; $\beta_x \cdot \kappa_t$ is the age-period term describing the mortality trends (κ_t is the time index and β_x modifies the effect of κ_t across ages). The following constraints on κ_t and β_x avoid identifiability problems with the parameters: $\sum_{t\in\mathcal{T}} \kappa_t = 0 \quad \sum_{x\in\mathcal{X}} \beta_x = 1$. Mortality forecasting is obtained by modeling the time index κ_t by an autoregressive integrated moving average (ARIMA) process. In general, a random walk with drift properly fits the data:

$$\kappa_t = \kappa_{t-1} + \delta + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_k^2), \tag{A12}$$

where δ is the drift parameter and ϵ_t are the error terms, normally distributed with null mean and variance σ_k^2.

CoDa model was proposed by Oeppen (2008) and suggests forecasting $d_{x,t}$ using principal component analysis in a compositional data pattern, following the Lee and Carter's original approach:

$$clr(d_{x,t} \ominus \alpha_x) = \kappa_t \beta_t + \epsilon_{x,t}, \tag{A13}$$

where *clr* is one of the log-ratio representations of compositional data. According to Bergeron-Boucher et al. (2017) it is defined as the logarithm of the composition divided by its geometric mean: $clr(d_{x,t} = ln(\frac{d_{x,t}}{g_t}))$, where g_t is the geometric mean of the age-composition at time t. The $\ominus$ operator represents the standard operation in compositional data analysis consisting in perturbing a composition by the inverse element of another composition. It is used to center the matrix while retaining the constant sum. The parameter, obtained by SVD, are $\kappa_t = u_t s$ and $\beta_t = v_x$, where s is the leading singular value, u_t and v_x refer to period and age components that are respectively the first left and the first right-singular vectors, and the α_x is the age-specific geometric mean of $d_{x,t}$ over time. Then, the model provides the age at death distribution through the closing procedure $C[\cdot]$ used to transform the estimates into compositional data

summing up to the initial constant:

$$d_{x,t} = \alpha_x \otimes C[e^{\kappa_t \beta_x + \epsilon_{x,t}}]. \tag{A14}$$

Vector autoregression model, also known as VAR, is one of the most applied models in empirical economics and finance for the analysis of multivariate time series. It is a multivariate stochastic process that can be used to model the joint evolution of two or more series over time. We refer to the first-order VAR model which consists in jointly modeling life expectancy and life disparity as follows:

$$e_{x,t} = \phi_0 + \phi_1 e_{x,t-1} + \phi_2 e^{\dagger}_{x,t-1} + \epsilon^{e}_{x,t}; \tag{A15}$$

$$e^{\dagger}_{x,t} = \theta_0 + \theta_1 e_{x,t-1} + \theta_2 e^{\dagger}_{x,t-1} + \epsilon^{e^{\dagger}}_{x,t}, \tag{A16}$$

where ϕ_i and θ_i (for $i = 0, 1, 2$) are the model parameters, and the errors $\epsilon^{e}_{x,t}$ and $\epsilon^{e^{\dagger}}_{x,t}$ follow a bivariate normal distribution with a zero mean vector and a constant covariance matrix.