

DEMOGRAPHIC RESEARCH

*A peer-reviewed, open-access journal of population sciences*

---

***DEMOGRAPHIC RESEARCH***

**VOLUME 47, ARTICLE 8, PAGES 199–232  
PUBLISHED 26 JULY 2022**

<http://www.demographic-research.org/Volumes/Vol47/8/>

DOI: 10.4054/DemRes.2022.47.8

*Research Article*

**Leveraging deep neural networks to estimate  
age-specific mortality from life expectancy at  
birth**

**Andrea Nigri**

**Susanna Levantesi**

**José Manuel Aburto**

© 2022 Nigri, Levantesi & Aburto.

*This open-access work is published under the terms of the Creative Commons Attribution 3.0 Germany (CC BY 3.0 DE), which permits use, reproduction, and distribution in any medium, provided the original author(s) and source are given credit.*

*See <https://creativecommons.org/licenses/by/3.0/de/legalcode>*

# Contents

1	Introduction	200
2	Models to derive age-specific mortality from life expectancy	201
2.1	Ševčíková and colleagues' model	202
2.2	Linear-link model	202
3	Data	203
4	Method: deep neural networks	203
4.1	Implementation	207
5	Results	209
5.1	Age-specific relative differences	210
5.2	Mean absolute error and root mean square error	212
6	Multi-population model	213
6.1	Multi-population model results	214
6.2	Network evaluation	217
7	Conclusion	218
8	Acknowledgements	220
	References	221
	Appendix	223

# Leveraging deep neural networks to estimate age-specific mortality from life expectancy at birth

Andrea Nigri<sup>1</sup>

Susanna Levantesi<sup>2</sup>

José Manuel Aburto<sup>3</sup>

## Abstract

### BACKGROUND

Life expectancy is one of the most informative indicators of population health and development. Its stability, which has been observed over time, has made the prediction and forecasting of life expectancy an appealing area of study. However, predicted or estimated values of life expectancy do not tell us about age-specific mortality.

### OBJECTIVE

Reliable estimates of age-specific mortality are essential in the study of health inequalities, well-being and to calculate other demographic indicators. This task comes with several difficulties, including a lack of reliable data in many populations. Models that relate levels of life expectancy to a full age-specific mortality profile are therefore important but scarce.

### METHODS

We propose a deep neural networks (DNN) model to derive age-specific mortality from observed or predicted life expectancy by leveraging deep-learning algorithms akin to demography's indirect estimation techniques.

### RESULTS

Out-of-sample validation was used to validate the model, and the predictive performance of the DNN model was compared with two state-of-the-art models. The DNN model

---

<sup>1</sup> Department of Social and Political Science, Bocconi University, Milan, Italy.

Email: [andrea.nigri@unibocconi.it](mailto:andrea.nigri@unibocconi.it).

<sup>2</sup> Sapienza, University of Rome, Rome, Italy.

<sup>3</sup> Leverhulme Centre for Demographic Science, Department of Sociology and Nuffield College, University of Oxford, Oxford, United Kingdom. Interdisciplinary Centre on Population Dynamics, University of Southern Denmark, Odense, Denmark. Department of Population Health, London School of Hygiene and Tropical Medicine, London, England.

provides reliable estimates of age-specific mortality for the United States, Italy, Japan, and Russia using data from the Human Mortality Database.

## **CONTRIBUTION**

We show how the DNN model could be used to estimate age-specific mortality for countries without age-specific data using neighbouring information or populations with similar mortality dynamics. We take a step forward among demographic methods, offering a multi-population indirect estimation based on a data driven-approach, that can be fitted to many populations simultaneously, using DNN optimisation approaches.

## **1. Introduction**

The rise in human longevity over the last two centuries has led to a growing interest in modelling and predicting death rates and life expectancy at birth (hereafter referred to as life expectancy). Reliable estimates of age-specific mortality are essential in the study of health inequalities and well-being between and within countries. This task comes with several difficulties, including a lack of reliable data or stochastic variation in death counts. Because of these difficulties in several countries and sub-populations, the regularities observed in trends of life expectancy have made this indicator appealing to model and predict. A key advantage of modelling life expectancy at birth, or at any age, is that the predictive model deals only with a single indicator that summarises the overall level of mortality over time, instead of modelling multiple time series of death rates for each age simultaneously.

Approaches that forecast life expectancy consider past trends in this indicator and its regularities, such as the linear increase of the best practice life expectancy (Oeppen and Vaupel 2002). Best practice life expectancy refers to the highest sex-specific national life expectancy observed in a given year. Lee (2006) exploits this regularity and models the changes in life expectancy as a linear function of the gap with the best practice trend, allowing countries to exceed the best practice levels. In contrast, Torri and Vaupel (2012) model life expectancy linearly by including a smooth function that accounts for the gap with the best practice life expectancy, which is constrained to not allow countries to overtake the best practice line. Raftery et al. (2013) introduce a Bayesian hierarchical model to obtain joint probabilistic projections of life expectancy in an international context. This model is currently used by the United Nations (2019). More recently, Nigri, Levantesi, and Marino (2021) and Levantesi, Nigri, and Piscopo (2022) propose forecasting life expectancy based on recurrent neural networks. These approaches forecast males and females independently. Pascariu, Canudas-Romo, and Vaupel (2018) further include the well-documented female advantage on longevity (Luy 2003) to forecast life expectancy

for both sexes simultaneously. Although the use of life expectancy as an indicator to forecast is appealing, estimating age-specific death rates is needed to analyse patterns of mortality at different ages and to calculate other indicators, such as lifespan inequality, as well as for estimating insurance pricing and pension liabilities. This has become even more important with the recent patterns of stalls in longevity improvements, or temporary reversals, observed in several countries (Nigri, Barbi, and Levantesi 2021), including the United States and the United Kingdom (Mehta, Abrams, and Myrskylä 2020; Aburto et al. 2021; Ho and Hendi 2018) but also around the globe in contexts where timely data is needed and often reported with significant delays, such as Mexico or Venezuela (Aburto et al. 2016; García and Aburto 2019).

Here, we propose a model to derive age-specific mortality from observed or predicted life expectancies. The model leverages deep-learning algorithms based on neural networks to uncover age-specific mortality based on past trends. Two approaches to deriving age-specific mortality from values of life expectancy that are more closely related to ours, which we describe in depth in the next section, were recently proposed by Ševčíková et al. (2016) and Pascariu et al. (2020). Ševčíková et al. (2016) adopts a reverting process based on the Lee-Carter model, while Pascariu et al. (2020) follow a similar strategy expressing the logarithm of age-specific deaths as a linear function of the logarithm of life expectancy.

In this article, we take advantage of deep neural network (DNN) models to derive the full age-specific mortality profile from values of life expectancy overcoming the linearity assumption and data requirements from past methods and provide new insights into the indirect approaches. We also offer a further step ahead by extending our DNN model to multiple populations (i.e., countries and both sexes). The resulting estimates would be useful for guiding public health interventions, informing about age-specific mortality dynamics in contexts with deficient data collection, as well as pension and social security schemes, which rely on longevity dynamics.

## 2. Models to derive age-specific mortality from life expectancy

We chose to benchmark our model using two recently proposed indirect models. Ševčíková et al. (2016) and Pascariu et al. (2020) propose two models aimed at deriving age-specific mortality from values of life expectancy at birth in line with the functional form of the well known Lee-Carter model (Lee and Carter 1992). Let  $\mathcal{A} = \{a_0, a_1, \dots, a_\omega\}$  and  $\mathcal{T} = \{t_0, t_1, \dots, t_n\}$  be the set of age and year categories, respectively. The Lee-Carter model describes the logarithm of the central death rate at age  $a \in \mathcal{A}$  and time  $t \in \mathcal{T}$ ,  $\log(m_{a,t})$ , as

$$\log(m_{a,t}) = \alpha_a + \beta_a \kappa_t + \epsilon_{t,a} \quad (1)$$

where  $\alpha_a$  captures the log mortality average by age,  $\kappa_t$  is the level of mortality in year  $t$ ,  $\beta_a$  is an age pattern of mortality change at age  $a$ , and  $\epsilon_{t,a}$  is the error term. The following constraints on  $\kappa_t$  and  $\beta_a$  avoid identifiability problems with the parameters

$$\sum_{t \in \mathcal{T}} \kappa_t = 0 \quad \sum_{a \in \mathcal{A}} \beta_a = 1.$$

Lee and Carter (1992) find that  $\kappa_t$  changes linearly and can be forecasted using a random walk with drift or other time series methods.

## 2.1 Ševčíková and colleagues' model

The first method proposed to estimate an age-specific mortality profile from a projected or forecasted value of life expectancy at birth was developed by Ševčíková et al. (2016). Their method consists of calibrating the parameter that reflects the level of mortality in the Lee-Carter model ( $\kappa_t$ ) to derive a desired level of life expectancy, similarly to the ideas proposed by Lee and Miller (2001) and Li, Lee, and Gerland (2013).

Let  $t \in \{1, \dots, T\}$  and  $\tau \in \{T + 1, \dots, T_p\}$  denote the observed and projected time periods, respectively. Ševčíková et al. (2016) estimate the Lee-Carter parameters  $\alpha_a$ ,  $\kappa_t$ , and  $\beta_a$  using the observed death rates  $m_{a,t}$  independently by sex. For a given value of projected life expectancy at birth  $e_0(\tau)$ , the method solves for future  $k_\tau$  based on the previously estimated parameters  $\hat{\alpha}_a$  and  $\hat{\beta}_a$  using life tables. Finally, the age-specific log-death rates are derived as follows:

$$\log(\hat{m}_{a,\tau}) = \hat{\alpha}_a + \hat{\beta}_a \hat{k}_\tau.$$

## 2.2 Linear-link model

The linear-link model proposed by Pascariu et al. (2020) derives specific death rates at time  $t$  and age  $a$ , with  $m_{a,t}$  as a linear function of the logarithm of life expectancy at birth ( $e_{0,t}$ ) and at time  $t$  given by

$$\log(m_{a,t}) = \beta_a \log(e_{0,t}) + \nu_a k + \varepsilon_{a,t} \quad (2)$$

The linear-link model is based on the least squares estimation of the slope  $\beta_a$  over the observation period.  $\beta_a$  can be regarded as an age-specific parameter and  $\varepsilon_{a,t}$  can denote a set of normally distributed errors with mean zero and variance  $\sigma^2$ . The model specification involves a second step to compute the singular value decomposition of the matrix of regression residuals to obtain the parameter  $\nu_a$ . To avoid projecting age-specific noise, Pascariu et al. (2020) smooth the parameters  $\beta_a$  and  $\nu_a$  using splines. Finally, parameter  $k$  is optimised to achieve the value of a projected life expectancy. The model can also be estimated by assuming that deaths follow a Poisson distribution with maximum likelihood estimation.

### 3. Data

We use high quality  $1 \times 1$  life tables from the Human Mortality Data base (HMD 2021) categorised by sex with a focus on Japan, the United States, Italy, and Russia from 1950 to 2015 to test the accuracy of our model. This set of countries covers a range of longevity trajectories with Japan having one of the highest life expectancies in the world, the United States with stagnation and slow improvements in life expectancy, Italy with its late demographic transition and rapid increase in life expectancy, and Russia with the highest mortality at younger ages and lower life expectancy within the HMD.

### 4. Method: deep neural networks

Deep-learning techniques, including DNNs, have become important in a wide range of applications, such as image classification or speech recognition with high predictive accuracy, often on par with human performance. Conventional machine-learning techniques were limited in their ability to process data, requiring careful engineering. DNNs provide higher flexibility, relying on the paradigm of representation learning, with multiple levels of representation, obtained by composing nonlinear modules. From the input data, they build layer by layer, new sets of features, to make optimal predictions of target variables (for more details see Lecun, Bengio, and Hinton (2015)). Recent contributions of deep-learning in longevity have been proposed in the field of actuarial science (see, e.g., Hainaut (2018), Richman and Wüthrich (2021), Perla et al. (2021), and Scognamiglio (2022)); however, its applications in demographic research are still scarce.

A DNN is a collection of neurons organised in a sequence of multiple layers, where

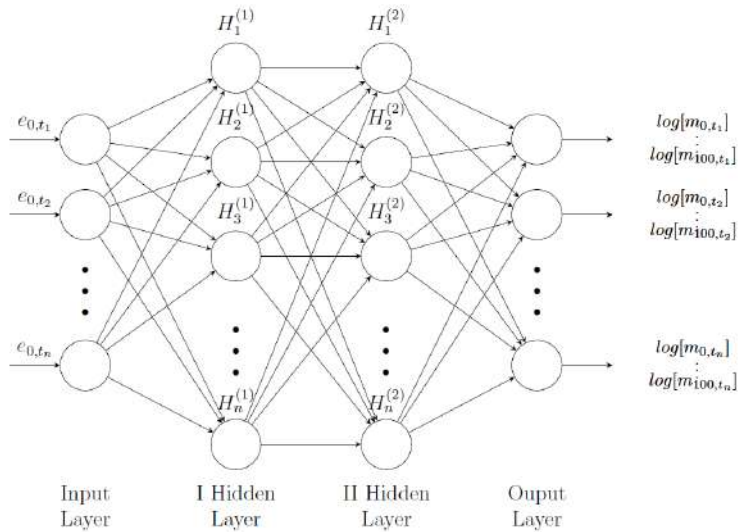
the input is the neuron activation from the previous layer that performs a weighted sum of the input followed by a nonlinear activation (Montavon, Samek, and Müller 2018). The neurons then implement complex nonlinear mapping from the input to the output. This mapping is learned from the data by adapting the weights of each neuron performing a technique known as error back-propagation (Rumelhart, Hinton, and Williams 1986). The general idea is that for a given set of training data  $\{(x_1, y_1) \dots (x_n, y_n)\}$  sampled according to an unknown probability distribution  $P(\mathbf{x}, \mathbf{y})$ , we find a function  $f(\cdot)$  that minimises the expected error on a new test set of data:

$$\int L(y, f(\mathbf{x}))P(\mathbf{x}, \mathbf{y})d\mathbf{x}d\mathbf{y},$$

where  $L(y, f(\mathbf{x}))$  is the loss function that measures the prediction error for a given  $x$  against the actual value  $y$ . We propose a model based on DNNs that assigns to life expectancy at birth at a generic time  $t_i$  a vector of age-specific death rates with the structure shown in Figure 1, where the input is the vector of life expectancy at birth over time  $t \in \mathcal{T}$ ,  $\mathbf{e}_0 = (e_{0,t_1}, e_{0,t_2}, \dots, e_{0,t_n})$ .



**Figure 1:** Graphical representation of our DNN model. The input is a vector of life expectancies over time  $e_0$ , which passes through the neurons and all multiple layers. The output in this diagram is a set of log-death rates at each age that correspond to each value of life expectancy, trained with observed age-specific data.



For a hidden layer  $\mathbf{H}^{(k)}$ , the specific neural network structure illustrated in Figure 1 is given by the following:

$$\mathbf{H}^{(k)} = f^{(k)} \left( \underbrace{\begin{bmatrix} w_{1,1}^{(k)} & w_{1,2}^{(k)} & w_{1,3}^{(k)} & \cdots & w_{1,n}^{(k)} \\ w_{2,1}^{(k)} & w_{2,2}^{(k)} & w_{2,3}^{(k)} & \cdots & w_{2,n}^{(k)} \\ w_{3,1}^{(k)} & w_{3,2}^{(k)} & w_{3,3}^{(k)} & \cdots & w_{3,n}^{(k)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n,1}^{(k)} & w_{n,2}^{(k)} & w_{n,3}^{(k)} & \cdots & w_{n,n}^{(k)} \end{bmatrix}}_{\mathbf{W}^{(k)}} \underbrace{\begin{bmatrix} H_1^{(k-1)} \\ H_2^{(k-1)} \\ H_3^{(k-1)} \\ \vdots \\ H_n^{(k-1)} \end{bmatrix}}_{\mathbf{H}^{(k-1)}} + \underbrace{\begin{bmatrix} b_1^{(k)} \\ b_2^{(k)} \\ b_3^{(k)} \\ \vdots \\ b_n^{(k)} \end{bmatrix}}_{\mathbf{b}^{(k)}} \right) \quad (3)$$

where  $f^{(k)}$  is the activation function,  $\mathbf{W}^{(k)}$  is the matrix of weights,  $\mathbf{H}^{(k-1)}$  is the hidden layers, and  $\mathbf{b}^{(k)}$  is the bias used to control the triggering value of the activation function.

We use the rectified linear unit (ReLU) function given by  $f(z) = \max(z, 0)$  (Glorot, Bordes, and Bengio 2011). This function ensures faster learning in networks with many layers. In the feed-forward architecture, each hidden layer involves the previous one as shown below:

$$\mathbf{H}^{(k)} = f^{(k)}(\mathbf{W}^{(k)} \mathbf{H}^{(k-1)} + \mathbf{b}^{(k)}),$$

where  $\mathbf{H}^{(k-1)}$  can be expressed as a function of a vector with life expectancy values as follows:

$$\mathbf{H}^{(k-1)} = f^{(k-1)}(\dots f^{(1)}(\mathbf{W}^{(1)} \mathbf{e}_0 + \mathbf{b}^{(1)}) \dots).$$

Let  $\mathbf{M} = \log(m_{a,t})_{a \in \mathcal{A}, t \in \mathcal{T}}$  be a matrix with death rates, where rows denote age and columns calendar years. For a standard architecture consisting of three hidden layers  $k = 3$  (input, hidden and output layer, respectively), the theoretical relationship defining the matrix of mortality  $\mathbf{M}$  given the vector of life expectancy at birth  $\mathbf{e}_0$  is represented by:

$$\mathbf{M} = f^{(3)}\left(\mathbf{W}^{(3)} f^{(2)}\left(\mathbf{W}^{(2)} f^{(1)}\left(\mathbf{W}^{(1)} \mathbf{e}_0 + \mathbf{b}^{(1)}\right) + \mathbf{b}^{(2)}\right) + \mathbf{b}^{(3)}\right), \quad (4)$$

where  $f^{(1)}(\mathbf{W}^{(1)} \mathbf{e}_0 + \mathbf{b}^{(1)}) = \mathbf{H}^{(1)}$  is the first hidden layer that accepts the vector  $\mathbf{e}_0$  as input.

The DNN model is based on a training algorithm that involves an unconstrained optimisation problem aiming to minimise the prediction error. The idea is to adjust the weights of the network connections to minimise a measure of the difference between the actual and desired output ( $\mathbf{M}$  and  $\hat{\mathbf{M}}$ ), respectively, known as the loss function  $\mathcal{L}$ . We use the mean square error (MSE) as a loss function given by

$$\mathcal{L}[\mathbf{M}, \hat{\mathbf{M}}] = \frac{1}{|\mathcal{A}| \cdot |\mathcal{T}|} \sum_{a,t} [\log(m_{a,t}) - \log(\hat{m}_{a,t})]^2$$

We chose the MSE because it is the benchmark in neural network regression problems (Lecun, Bengio, and Hinton 2015) and was the best performer compared to other suitable loss functions with our dataset. To minimise the loss function, we use gradient descent optimisation. Gradient descent is one of the most popular algorithms used to perform optimisation and is the most common way to optimise neural networks. It consists

of minimising the loss function by updating the weights in the opposite direction of the gradient ( $\nabla\mathcal{L}$ ) with respect to the weights. For a generic set of weights  $w_{n,n}^{(k)}$  and the  $k^{th}$  layer, using the chain rule, the gradient is given by

$$\nabla\mathcal{L} = \frac{\partial\mathcal{L}[\mathbf{M}, \hat{\mathbf{M}}]}{\partial w_{n,n}^{(k)}} = \frac{\partial\mathcal{L}[\mathbf{M}, \hat{\mathbf{M}}]}{\partial H_n^{(k)}} \frac{\partial H_n^{(k)}}{\partial z_n^{(k)}} \frac{\partial z_n^{(k)}}{\partial w_{n,n}^{(k)}}, \quad (5)$$

where  $z_n^{(k)} = w_n^{(k)} H_n^{(k-1)} + b_n^{(k)}$ . The gradient encodes the relative importance of each weight and bias. The algorithm for efficiently computing the gradient in Equation (5) is known as back-propagation. Back-propagation consists of a recursive algorithm. In the forward step, the prediction is computed by fixing the weights; subsequently, in the backward step, the weights are adjusted by back-propagating the gradient of the loss function to reduce the error. As a result of these adjustments, the internal hidden layers, which are not part of the input or output, are able to represent and capture important features of age-specific mortality. To update the weights ( $\tilde{\mathbf{W}}$ ), the gradient of the loss function is multiplied by a scalar,  $\eta$ , often called the learning rate, according to the following scheme:

$$\tilde{\mathbf{W}} = \mathbf{W} - \eta \nabla\mathcal{L} [\mathbf{M}, \hat{\mathbf{M}}]. \quad (6)$$

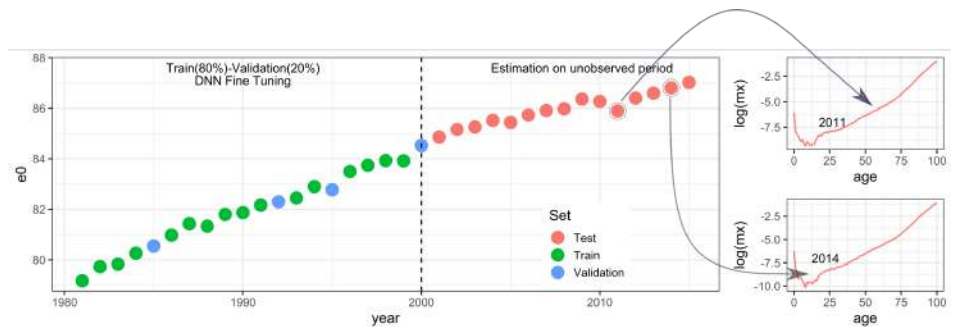
The learning rate  $\eta$  determines the size of the step taken to reach a global or local minimum. In other words, gradient descent is similar to ‘climbing down a hill’ until a global or local minimum is reached. For this stage, we implement the root mean square propagation algorithm proposed by Hinton, Srivastava, and Swersky (2013).

#### 4.1 Implementation

Testing the model accuracy on unseen data is a crucial phase when you parametrise (or train) the DNN model. Indeed, even traditional machine learning models require the choice of hyperparameters, which cannot be calibrated directly from the data. In the DNNs, they usually refer to the choices of model structure, such as the number of neurons, hidden layers, and epochs, and it is common practice to perform a fine-tuning phase according to the training error minimisation. This choice depends on the type of data that remains a heuristic problem in the field of neural networks. Therefore a fundamental procedure when applying machine- and deep-learning models is to test the performance of these models on unseen data. To this aim, our model requires an input vector with the time series of life expectancy at birth and a matrix with the corresponding age-specific death rates over columns and time periods over rows. Each data series is split into a training-

validation set with which the network is trained and a test set to check the accuracy of the model's prediction. The scheme is presented in Figure 2.

**Figure 2: Implementation scheme of the DNN model. The model is trained and validated with observed age-specific death rates that are consistent with life expectancy levels from the train and validation period (green and blue dots). The results from this training phase are then applied to estimate a full age-specific mortality profile for a given value of (projected or forecasted) life expectancy (orange dots).**



The process consists of feeding the model the training set and subsequently assessing its accuracy on the validation set. The test set stands for the unobserved time horizon, on which the model comparison will be performed. In practice, the test set would be the forecasted or projected life expectancy, for which the age-specific mortality profile is unknown. Formally, let  $t_\tau$ , with  $t_0 < t_\tau < t_s$ , be the calendar year that corresponds to the last realisation in the train-validation set. The values of life expectancy in the period  $(t_0, t_\tau)$ ,  $(e_{0,t})_{t \in [t_0, t_\tau]}$ , represent the input for train-validation, while the corresponding output is  $\log(\hat{m}_{a,t})_{a \in \mathcal{A}, t \in [t_0, t_\tau]}$ . During the train-validation phase, the neural network weights are estimated and subsequently used in the test phase concerning the period  $[t_{\tau+1}, t_s]$ , which starts from  $t_{\tau+1}$ .

The values of life expectancy over a subsequent period,  $(e_{0,t})_{t \in [t_{\tau+1}, t_s]}$ , represent the input for the test set, while the corresponding output is  $\log(\hat{m}_{a,t})_{a \in \mathcal{A}, t \in [t_{\tau+1}, t_s]}$ . Thereby, denoting  $\psi_{nn}$  as a composition of functions defined on the basis of the DNN architecture, the model can be described by:

$$\log(\hat{m}_{a,t}) = \psi_{nn} \left\{ (e_{0,t}) \middle| \hat{W} \right\}; \quad \forall a \in \mathcal{A}; \quad \forall t \in [t_{\tau+1}, t_s] \quad (7)$$

where  $(\hat{m}_{a,t})_{a \in \mathcal{A}, t \in [t_{\tau+1}, t_s]}$  is the matrix of death rates in the test set obtained by  $\psi_{nn}$  that involves the DNN weights  $\hat{W}$  estimated during the network training, using the MSE as a loss function and the ReLU as an activation function. Considering our data, which we used the 80%–20% splitting rule randomly sampled as the train-validation sets, the use of the validation set drastically reduces the chances to incur overfitting. In order to minimise these possibilities, we strengthen the model by introducing two regularisation techniques, drop out and early stopping, respectively. Early stopping acts on the neural network epochs. One epoch is defined as one full cycle through the training data by the network and modifying the weights. The epochs number corresponds to the total number of times the full training dataset is explored. Early stopping allows stopping training once the model performance stops improving on the validation dataset. Furthermore, to prevent the neurons' co-adaptation and reduce overfitting, the dropout regularisation technique is often used. Such method consists on randomly removing neurons along with their incoming and outgoing connections during training. For a detailed description of the selected hyperparameter, see Table A-3 in the Appendix.

## 5. Results

To assess the robustness of our method, we performed an out-of-sample test over three time windows (1950–1980, 1960–1990, and 1970–2000) used as train-validation sets, and the subsequent years of each time window (1981–1995, 1991–2005, and 2001–2015, respectively) as test sets. The model was applied to data from Italy, Japan, Russia, and the United States by sex. We use a six-hidden-layer architecture following the fine-tuning and compared the results from the DNN model with those obtained in Ševčíková et al. (2016)'s and the linear-link models. To ensure comparability, the results were smoothed using P-splines (He and Ng 1999).<sup>4</sup> We focus on the results pertaining to females during the period from 2001–2015 in this section; the results related to males in the same years and in the other study periods for both sexes are reported in the Appendix.

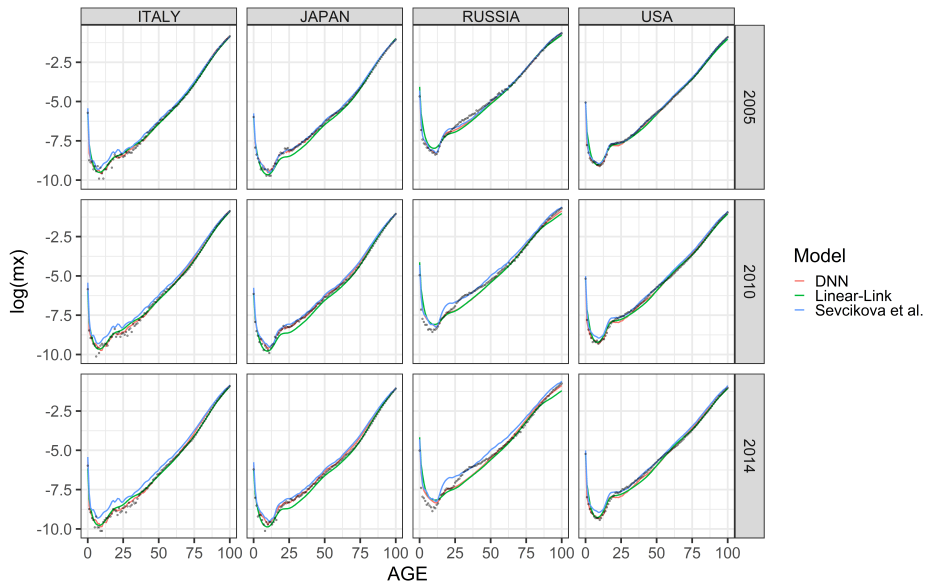
### Age-specific mortality estimates

Figure 3 shows age-specific death rates (in log scale) for females in Russia, Japan, Italy, and the United States. The observed (target) profile is shown with dots and estimated values from the models using the training period 1970–2000, which correspond to DNN (red line), linear-link (green line) and Ševčíková et al. (2016) (blue line). The three models

<sup>4</sup> Unsmoothed results are shown in the additional material. The smoothing step does not affect the models' accuracy ranking: the root mean square error (RMSE) and mean absolute error (MAE) improvements after the smoothing are on average 1.55% and 1.57%, respectively.

capture the general pattern of mortality, with a decreasing trend from birth to around age 15, and increasing linearly from around age 30. For Italy and a recent period in Russia, the Ševčíková et al. (2016) model tends to overestimate mortality at young ages, while the opposite is true for the linear-link in the case of Japan. The DNN model adequately captures the mortality patterns. However, the three models fail to accurately capture the sharp decrease from infancy in the cases of Italy and Russia.

**Figure 3:** Estimated age-specific female log-mortality rates  $\log(m_{a,t})$  for three models: DNN, linear-link and Ševčíková et al. (2016), by country, for 2005, 2010, and 2014 based on the training period 1970–2000. The black dots are the observed log-mortality rates.



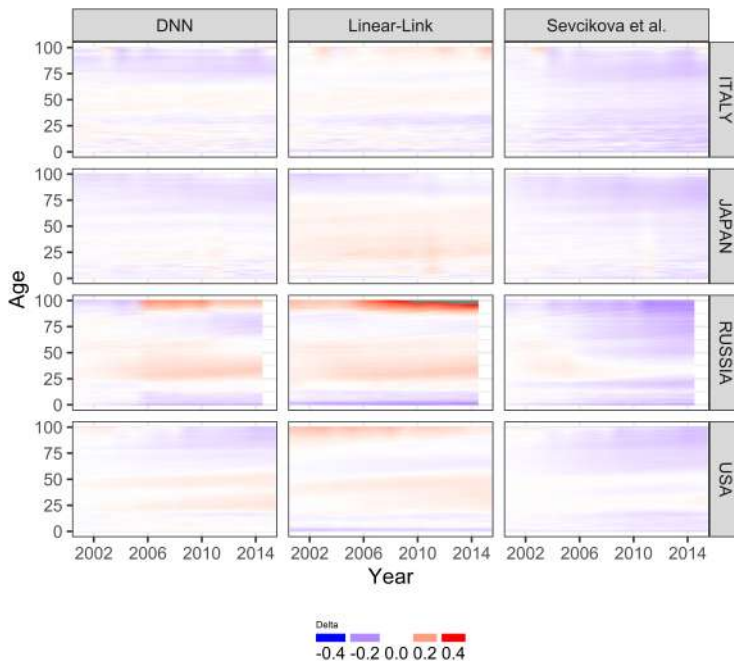
### 5.1 Age-specific relative differences

We further analyse the accuracy of the models with the relative differences ( $\Delta_{a,t}$ ) between estimates and the observed death rate by age for each model (see Figure 4) defined as

$$\Delta_{a,t} = \frac{\log(\hat{m}_{a,t}) - \log(m_{a,t})}{\log(m_{a,t})}.$$

Because the relative difference is calculated from the log-death rates, red hues indicate that the model underestimates mortality, while blue hues indicate overestimation. Differences with the observed mortality profile are small in general across models and countries. However, it is observed a systematic underestimation in working-ages (20 to 50) for Russia and the United States from the DNN and linear-link models in the time window from 2001–2015 (the other two time windows are shown in the Appendix Figures A-4, A-5, A-6, and A-7). Similarly, there is increased deviation at very old ages for recent periods from both models. In contrast, the Ševčíková et al. (2016) model tends to overestimate mortality at older ages for all countries, especially for recent periods in Russia.

**Figure 4:** Relative differences ( $\Delta_{a,t}$ ) between estimates and the observed death rate by age for each model. Red hues indicate that the model underestimates mortality, while blue hues indicate overestimation. The female test period took place 2001–2015.



Among males (Figure A-1), the DNN and linear-link models tend to underestimate mortality at working-ages, which is compensated with increased mortality at younger and

older ages below the age of 90. In contrast, the Ševčíková et al. (2016) model, as in the case for females, tends to overestimate mortality across all countries and is best used to capture the working-age pattern for Russia and the United States. Notably, deviations from the observed mortality increase with time across all time windows.

## 5.2 Mean absolute error and root mean square error

To summarise the performance of the methods and to evaluate their accuracy, we report the mean absolute error (MAE) and root mean square error (RMSE) on the test sets given by the following:

$$\text{MAE} : \sum_a \sum_t \frac{|\log(m_{a,t}) - \log(\hat{m}_{a,t})|}{|\mathcal{A}| \cdot |\mathcal{T}|},$$

$$\text{RMSE} : \sqrt{\frac{\sum_a \sum_t (\log(m_{a,t}) - \log(\hat{m}_{a,t}))^2}{|\mathcal{A}| \cdot |\mathcal{T}|}}.$$

Tables 1 and A-1 summarise the MAE and RMSE for the three models and four countries over the three time windows that we studied for females and males, respectively. For females, the DNN is the best performer in most cases, although the linear-link model showed the lowest MAE for Italy in the earliest and latest periods. The Ševčíková et al. (2016) model exhibited the lowest RMSE for Russia in the period 1991–2005. The results for males (see Table A-1) show less consistent results. For Italy and the United States, the DNN model consistently showed the lowest errors, but the Ševčíková et al. (2016) model performed the best for Japanese males.

Among males, for Italy and the United States the DNN model was the best performer in terms of MAE and RMSE. For Japan, as noted in the age-specific figures, the Ševčíková et al. (2016) model showed the lowest errors. For Russia, it was a mix between the three models depending on the time window and summary measure. For both sexes, while the DNN model showed in the majority of cases the lowest departures from age-specific mortality, the linear-link model performed the best in capturing the life expectancy level.



**Table 1: Out-of-sample test: MAE and RMSE for DNN, linear-link, and Ševčíková et al. (2016) by country and sex. The estimation period for females took place 1981–1995, 1991–2005, and 2001–2015.**

Country	Model	1981–1995		1991–2005		2001–2015	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
<i>Italy</i>	DNN	0.1672	<b>0.2017</b>	<b>0.1038</b>	<b>0.1418</b>	0.1210	<b>0.1591</b>
	linear-link	<b>0.1379</b>	0.2325	0.1561	0.2272	<b>0.1154</b>	0.1818
	Ševčíková et al. (2016)	0.2421	0.2995	0.2042	0.2598	0.2539	0.3278
<i>Japan</i>	DNN	<b>0.1730</b>	<b>0.2055</b>	<b>0.1182</b>	<b>0.1532</b>	<b>0.1137</b>	<b>0.1436</b>
	linear-link	0.4036	0.5473	0.2602	0.3242	0.1664	0.2095
	Ševčíková et al. (2016)	0.2444	0.2785	0.1906	0.2376	0.1656	0.2014
<i>United States</i>	DNN	<b>0.0572</b>	<b>0.0788</b>	<b>0.0873</b>	<b>0.1119</b>	<b>0.0944</b>	<b>0.1206</b>
	linear-link	0.0720	0.1196	0.0949	0.1541	0.1183	0.1683
	Ševčíková et al. (2016)	0.1291	0.1642	0.1081	0.1524	0.1227	0.1643
<i>Russia (2014)</i>	DNN	-	-	<b>0.1714</b>	0.2544	<b>0.1670</b>	<b>0.2393</b>
	linear-link	-	-	0.1942	0.2736	0.2266	0.3223
	Ševčíková et al. (2016)	-	-	0.1931	<b>0.2510</b>	0.1907	0.2558

## 6. Multi-population model

In the previous section, we compared the performance of the DNN model with the linear-link and Ševčíková et al. (2016)'s models, which are the closest comparable models available. However, the DNN model can be used in a more general way in contexts for which there is an estimate of life expectancy but no available past age-specific mortality. In such contexts, information from neighbouring countries or from countries with similar mortality dynamics could be used to estimate an age-specific mortality profile. From this perspective, the time dimension is lost and not needed. Therefore, the DNN model fills the gap left by predecessor models by relying only on past data and becomes more akin to indirect methods or model lifetables.

Here, we present a simple example of how to extend the DNN model for the multi-population (mp-DNN) case. Consider the case in which the full HMD is used to train a model and then to predict the age-specific mortality profile of a country's life expectancy. We still model the functional relationship between life expectancy at birth and death rates, as numerical inputs and outputs. To extend the framework to the mp-DNN model by adding other demographic features, such as country, year, and sex, we use what is known as embedding layers (Richman (2020); Bengio, Courville, and Vincent (2013)). Embedding is a tool that allows for the capture of relationships that are otherwise difficult to capture due to high dimensionality. This is the case when data present many categorical variables, the one-hot encoding schemes produce high dimensional sparse vectors, which

often causes calibration difficulties. Embedding allows a low-dimensional representation learning, mapping categorical variables into a vector space. In the present study, we deal with the following categorical variables:

$$c \in \mathcal{C} = \{\text{Italy}, \dots, \text{Russia}\}; \quad g \in \mathcal{G} = \{\text{Male}, \text{Female}\}; \\ a \in \mathcal{A} = \{0, \dots, 100\}; \quad t \in \mathcal{T} = \{1990, \dots, 2015\}.$$

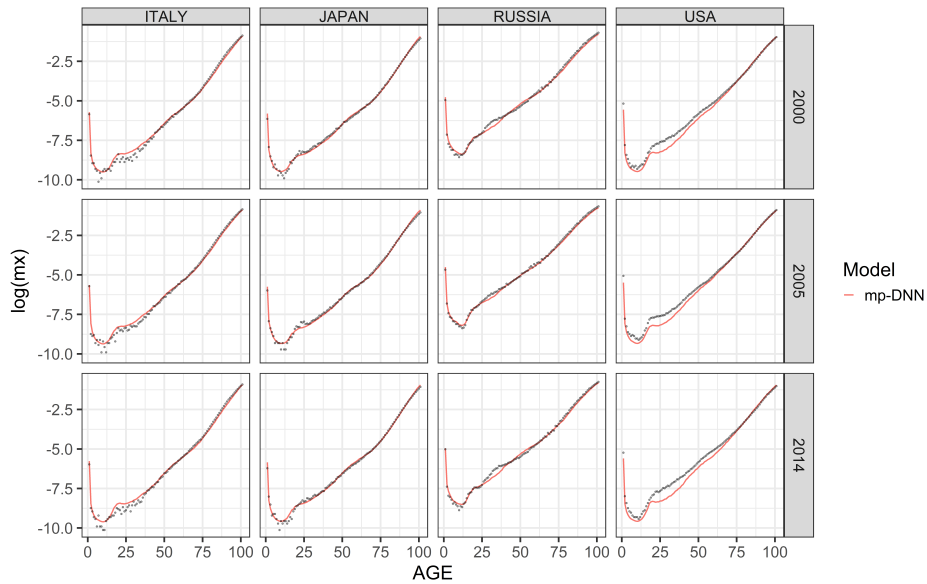
Embedding layers map these features into real-valued vectors, where for instance  $z_C(c)$  is the new representation of countries. Therefore,  $i = (c, g, a, t) \in \mathcal{I} = \mathcal{C} \times \mathcal{G} \times \mathcal{A} \times \mathcal{T}$  might be considered the categorical features space.

Once embedding vectors have been defined for each categorical variable, all variables, categorical and not, are concatenated into a single feature vector:  $x_{e_0, i} = (e_0, z_I(i))$ , which is used as input to the sub-neural network in order to predict the death rates in year  $t$  at age  $a$  for country  $c$  with gender  $g$  related to levels of observed life expectancy at birth. The network that accepts the input has a very similar structure to the single-population model. Specifically, a network consisting of two hidden layers with 150 neurons for each layer is used; also in this case, as in the single population model, we use regularisation techniques. Details regarding the choice of parameters are provided below in subsection 6.2.

## 6.1 Multi-population model results

Figures A-2 and 5 show age-specific death rates (in log scale) for females and males in Russia, Japan, Italy, and the United States, which were estimated by exploiting the mp-DNN framework trained on the whole HMD. At first glance, the mp-DNN provides smoothed estimation by nature, due to a wider training sample. The model is able to describe the general mortality shape and provides a good fit for Italy and Japan and remarkable accuracy for Russia, which represents a real challenge for the single-population models, even if it seems to underestimate mortality at old ages. The United States provides a particular example where mp-DNN constantly underestimates mortality at both young and older ages.

**Figure 5:** Estimated age-specific female log-mortality rates  $\log(m_{a,t})$  for the mp-DNN model by country for 2005, 2010 and 2014 based on the training period 1970–2000. Black dots are the observed log-mortality rates.

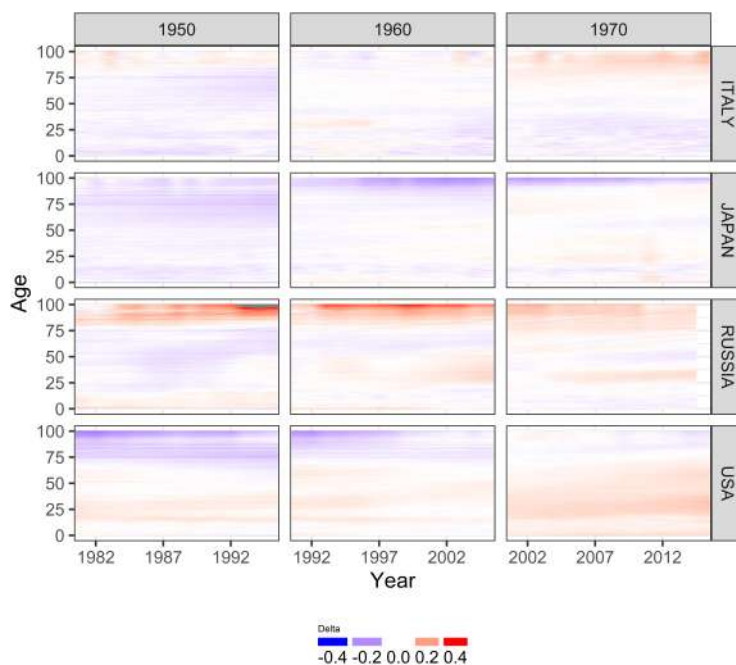


Figures 6 and A-3 show the accuracy of the mp-DNN model with the relative differences ( $\Delta_{a,t}$ ) between estimates and the observed death rates by age and time windows. Overall, mp-DNN shows small deviations with the observed mortality profile, across countries and periods, with the exception of the United States. In this case, the model provides inconsistent results, showing underestimations that increase over time. An alternative to treat atypical cases, such as the United States, could be used to train the model with data from countries with relatively high young mortality rates. For Italy, the model shows a high sensitivity over time of older ages, in particular for the female population. For Japan and Russia, as noted in the age-specific figures, the model provides reliable estimations, notably, deviations from the observed mortality decrease across all time windows.

Tables 2 and A-2 show the MAE and RMSE for the multi-population model estimation for the four countries over the three study periods for both sexes. We can confirm the inadequacy of this model to represent the United States mortality dynamics. However, we provide evidence of good accuracy for Italy and Japan and accurate results in terms

of errors referring to Russia. We underline that mp-DNN was able to provide reliable estimates for Russia over the first time window, where other models were not. Indeed, single population models, in order to estimate parameters, need training data that are not completely provided for Russia in the reference period 1950–1980. The mp-DNN leverages the multi-population estimated parameters by applying them to the Russia case of the out-of-sample window (1981–1995).

**Figure 6:** Relative differences ( $\Delta_{a,t}$ ) between estimates and the observed death rate by age for the mp-DNN model. Red hues indicate that the model underestimates mortality, while blue hues indicate overestimation. The female test period took place 2001–2015.



**Table 2: Out-of-sample test: MAE and RMSE for mp-DNN by country and sex. The estimation period for females took place 1981–1995, 1991–2005, and 2001–2015.**

Country	Model	1981–1995		1991–2005		2001–2015	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
<i>Italy</i>	mp-DNN	0.1380	0.1827	0.1024	0.1592	0.1395	0.1897
<i>Japan</i>	mp-DNN	0.1707	0.2026	0.09544	0.1215	0.09131	0.1314
<i>United States</i>	mp-DNN	0.1462	0.177	0.107	0.1336	0.2397	0.3115
<i>Russia (2014)</i>	mp-DNN	0.144	0.1728	0.1124	0.1382	0.1003	0.1265

## 6.2 Network evaluation

Despite DNN models having substantial advantages over traditional statistical methods that make them appealing for solving complex demographic tasks, the effect of hyper-parameters' choice on accuracy and stability has been often underrated. To facilitate the adoption of DNNs into longevity analysis tools, we attempt to explore this aspect for the indirect estimation of death rates. Here, we provide a complete perspective referring only to the multi-population model since it is being trained simultaneously on the whole HMD, therefore it is not affected by a single population's choice.

Taking the stochasticity in the input data as a realisation of an underlying stochastic process, which arises in all inferential methods, the main source of variability origins is due to the optimisation procedure. Despite the fact that the instability of the prediction towards different training setups is a common issue in the field of deep-learning, it might be problematic to explain in the field of mortality prediction.

Therefore, we study the neural network sensitivity towards parameter changes and how predictions vary each time the network is trained using different setup choices for different parts of the networks.

In a large space of combinations, we decided to move to a reasonable subspace, using the changes highlighted in the Tables 3 and 4. We carry out these experiments, comparing all possible combinations through the MSE and RMSE of the predictions averaged over the study countries. Tables 3 and 4 show that the adoption of dropout (set to 10%) gains better performances in the architecture composed of two layers, for all activation functions here considered (ReLU, Sigmoid, and Softmax). On the contrary, when we use one layer, the dropout technique seems to increase the error, especially using the Sigmoid activation function. This is due to the fact that the number of neurons drops dramatically, reducing the model's predictive ability. The most suitable architecture would seem to be the one with two intermediate layers, such as to better grasp the relationships between variables. From the tables, we can see how the ReLU activation function outperforms

the other tested functions, showing a better learning performance in networks with more complex architecture.

**Table 3: Males**

Layer	Error	ReLU		Sigmoid		Softmax	
		Drop-out	No Drop-out	Drop-out	No Drop-out	Drop-out	No Drop-out
1	MAE	0.14	0.14	0.31	0.23	0.28	0.28
	RMSE	0.18	0.18	0.40	0.31	0.37	0.35
2	MAE	0.13	0.16	0.15	0.23	0.20	0.44
	RMSE	0.17	0.24	0.21	0.32	0.29	0.63

**Table 4: Females**

Layer	Error	ReLU		Sigmoid		Softmax	
		Drop-out	No Drop-out	Drop-out	No Drop-out	Drop-out	No Drop-out
1	MAE	0.16	0.15	0.26	0.18	0.31	0.28
	RMSE	0.20	0.20	0.32	0.25	0.39	0.35
2	MAE	0.14	0.17	0.16	0.19	0.26	0.53
	RMSE	0.18	0.21	0.21	0.25	0.35	0.72

## 7. Conclusion

We presented a novel method to indirectly estimate a full mortality profile from a level of life expectancy at birth by leveraging deep neural networks using prior information on age-specific mortality. When tested with state-of-the-art methodologies, the DNN model performed the best in many cases with fewer assumptions than previous methods. The method outlined here is non-parametric and data-driven and does not rely on assumptions that may not be completely accurate. Nevertheless, the results show that the three models tested here perform in a satisfactory way, with the DNN model offering the best performance in most cases. As shown, the reconstruction of an accurate mortality surface from a given level of demographic summary measure, such as life expectancy, is challenging. However, we offer a new alternative based on machine-learning algorithms that complement the existing demographic toolbox. Moreover, the analysis of four countries with substantially different mortality and three sequential time windows of 30 years provided robust results. We confirm that the linear-link model, because of its dynamic constraint and consequential re-parametrisation, assures coherence with respect to input life expectancy level (providing a negligible error), while for the Ševčíková et al. (2016) and DNN models, deviations were larger.

A substantial advantage of the DNN model appears in the multiple-population framework. While the Ševčíková et al. (2016) and linear-link models were designed to derive an age-specific mortality profile for a future value of life expectancy based on past data of a single population, the DNN model could be used to estimate age-specific mortality for countries where there is no data available using information from countries with similar mortality dynamics. Therefore, the DNN model fills the gap of the predecessors' models by relying only on past data and becomes more akin to indirect methods or model lifetables. We take a step forward among demographic methods, offering a multi-population indirect estimation based on a data-driven approach, which can be fitted to many populations simultaneously, using DNN optimisation approaches. While we apply our methodology to country-specific scenarios, the model could be used to indirectly estimate mortality profiles for regions or subpopulations with similar mortality profiles. This characteristic makes our model appealing for countries where present information is lacking but past data are available from surrounding countries or populations, as we have shown in the Russian case.

We acknowledge that our model is subject to several limitations, including the choice of architecture (e.g., the number of hidden layers) and the parameters involved in the training phase. This remains a heuristic problem for neural network users, as indeed the choice often depends on the type of data and a preliminary round of fine-tuning, before the testing, which is highly desired, albeit somewhat time-consuming (Nigri et al. 2019). This issue is dimmed in the case of multi-populations. Indeed, this framework relies on a bigger data set, thus more examples during the training phases are required, which has two implications. On the one hand, the model provides more robustness towards structural changes; however, some country-specific dynamics may not be captured, as in the case of the United States. Therefore, we strongly recommend the careful selection of the countries' subgroup on which the DNN model will be trained. Finally, although some studies have attempted to provide a viable alternative (see, e.g., Marino, Levantesi, and Nigri (2022); Richman (2021)), how the uncertainty in the prediction could be derived when using deep-learning techniques is still considered a big challenge. Indeed, the resulting DNN estimate is a point estimation that does not provide any information on the uncertainty given by  $\hat{W}$ , since DNN suffers from different uncertainty sources that affect the learning process.

To conclude, here we propose a new approach that provides a valuable alternative tool to capture irregular mortality trajectories. While machine learning, deep or not, is not widely used in the field of demography, our method shows that it can be used to provide robust estimations of age-specific death rates. Due to its nature, our method can be leveraged in other demographic contexts, (e.g., to derive age-specific fertility profiles from observed or predicted mean age at childbearing). This may foster new research at the frontier of demographic studies using innovative, yet simple to implement, techniques such as the DNN model. This is even more important in the context of rapid population

ageing and fast mortality decline, but also in contexts where lacking mortality estimates can provide crucial information for policy planning.

## 8. Acknowledgements

**Material availability:** The replication scripts, written in the R statistical programming language (R. Core Team, 2013), are hosted on the Open Science Framework (OSF) at [https://osf.io/5dynw/?view\\_only=badbcfa0e67b4d98a78560acb50f0b8c](https://osf.io/5dynw/?view_only=badbcfa0e67b4d98a78560acb50f0b8c)

**Funding:** JMA acknowledges support from the British Academy Newton International Fellowship NIFBA19/190679, Leverhulme Trust Large Centre Grant and the Interdisciplinary Centre for Population Dynamics at University of Southern Denmark.

**Authors contributions:** AN conceived the idea. AN and JMA developed the methodology. AN wrote the code and performed the computations. AN and SL drafted the first version of the manuscript. JMA supervised the findings of this work. All authors discussed the results and contributed to writing and editing the final manuscript.

**Acknowledgements:** We thank Charles Rahal, Mark Verhagen, as well as two anonymous reviewers and the Associate Editor, for their useful discussions and comments on previous versions of this work.



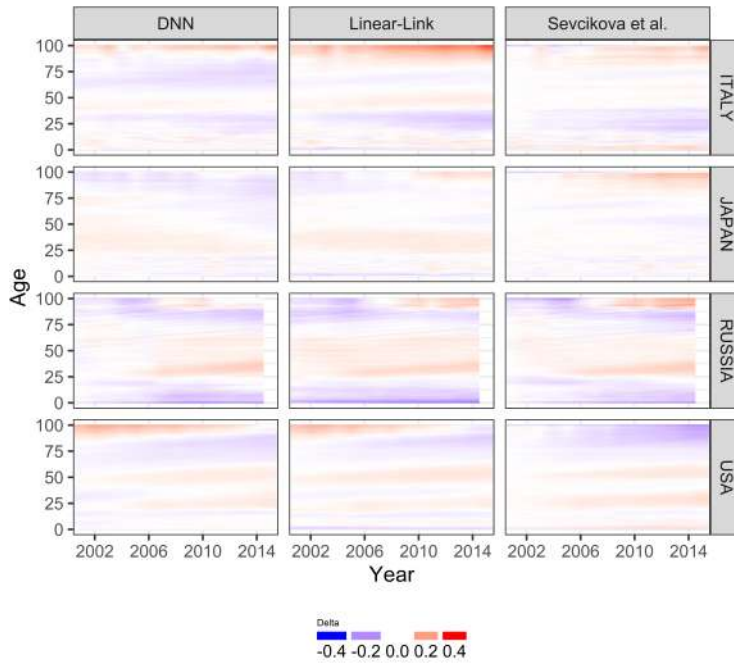
## References

- Aburto, J.M., Kashyap, R., Schöley, J., Angus, C., Ermisch, J., Mills, M., and Dowd, J.B. (2021). Estimating the burden of the covid-19 pandemic on mortality, life expectancy and lifespan inequality in England and Wales: A population-level analysis. *Journal of Epidemiology & Community Health* doi:10.1136/jech-2020-215505.
- Aburto, J.M., Beltrán-Sánchez, H., García-Guerrero, V.M., and Canudas-Romo, V. (2016). Homicides in Mexico reversed life expectancy gains for men and slowed them for women, 2000-10. *Health Affairs* 35(1): 88–95. doi:10.1377/hlthaff.2015.0068.
- Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8): 1798–1828. doi:10.1109/TPAMI.2013.50.
- García, J. and Aburto, J.M. (2019). The impact of violence on Venezuelan life expectancy and lifespan inequality. *International Journal of Epidemiology* 48(5): 1593–1601. doi:10.1093/ije/dyz072.
- Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In: Gordon, G., Dunson, D., and Dudík, M. (eds.). *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Fort Lauderdale, FL, USA: JMLR Workshop and Conference Proceedings, vol. 15 of *Proceedings of Machine Learning Research*, 315–323.
- Hainaut, D. (2018). A neural-network analyser for mortality forecast. *ASTIN Bulletin* 48(2): 481–508.
- He, X. and Ng, P. (1999). Cobs: Qualitatively constrained smoothing via linear programming. *Computational Statistics* 14: 315–337. doi:10.1007/s001800050019.
- Hinton, G., Srivastava, N., and Swersky, K. (2013). Neural networks for machine learning. lecture 6a: Overview of mini-batch gradient descent. *Department of Computer Science University of Toronto*.
- HMD (2021). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). <https://www.mortality.org>. Data downloaded on 01/01/2019.
- Ho, J. and Hendi, A. (2018). Recent trends in life expectancy across high income countries: Retrospective observational study. *BMJ* 362(k2562).
- Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521(7553): 436–444. doi:10.1038/nature14539.
- Lee, R. and Miller, T. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography* 38(6): 537–549.
- Lee, R. (2006). Mortality forecasts and linear life expectancy trends. Perspectives on mortality forecasting. The linear rise in life expectancy: History and prospects. *Social Insurance Studies* 3: 19–39.
- Lee, R. and Carter, L. (1992). Modeling and forecasting us mortality. *Journal of the American Statistical Association* 87: 659–671.
- Levantesi, S., Nigri, A., and Piscopo, G. (2022). Clustering-based simultaneous forecasting of life expectancy time series through long-short term memory neural networks. *International Journal of Approximate Reasoning* 140: 282–297. doi:10.1016/j.ijar.2021.10.008.
- Li, N., Lee, R., and Gerland, P. (2013). Extending the Lee-Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography* 50(6): 2037–2051.
- Luy, M. (2003). Causes of male excess mortality: Insights from cloistered populations. *Population and Development Review* 29(4): 647–676.

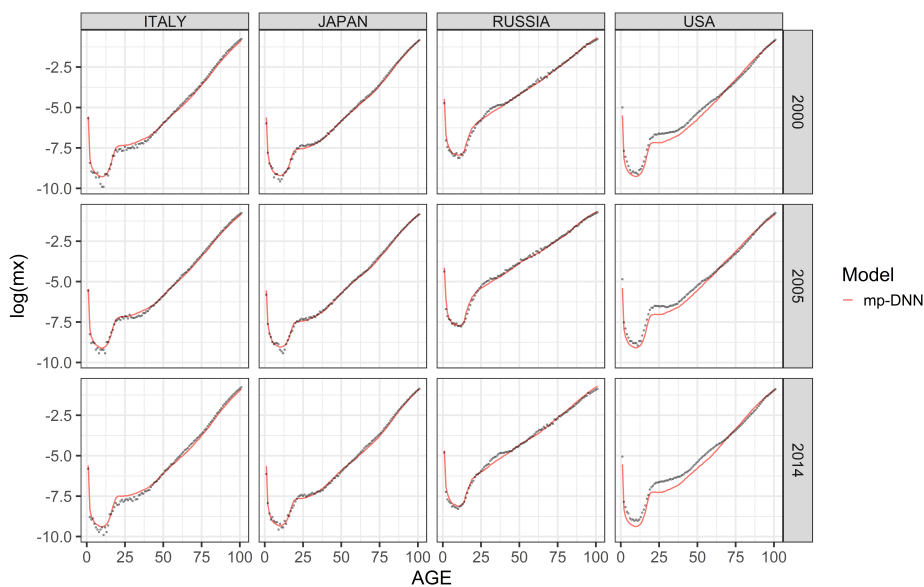
- Marino, M., Levantesi, S., and Nigri, A. (2022). A neural approach to improve the Lee-Carter mortality density forecasts. *North American Actuarial Journal* doi:10.1080/10920277.2022.2050260.
- Mehta, N.K., Abrams, L.R., and Myrskylä, M. (2020). US life expectancy stalls due to cardiovascular disease, not drug deaths. *Proceedings of the National Academy of Sciences* 117(13): 6998–7000.
- Montavon, G., Samek, W., and Müller, K. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73: 1–15.
- Nigri, A., Barbi, E., and Levantesi, S. (2021). The relationship between longevity and lifespan variation. *Statistical Methods and Applications* doi:10.1007/s10260-021-00584-4.
- Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., and Perla, F. (2019). A deep learning integrated Lee-Carter model. *Risks* 7(1). doi:10.3390/risks7010033.
- Nigri, A., Levantesi, S., and Marino, M. (2021). Life expectancy and lifespan disparity forecasting: a long short-term memory approach. *Scandinavian Actuarial Journal* 2021(2): 110–133. doi:10.1080/03461238.2020.1814855.
- Oeppen, J. and Vaupel, J.W. (2002). Broken limits to life expectancy. *Science* 296(5570): 1029–1031. doi:10.1126/science.1069675.
- Pascariu, M.D., Basellini, U., Aburto, J.M., and Canudas-Romo, V. (2020). The linear link: Deriving age-specific death rates from life expectancy. *Risks* 8(4).
- Pascariu, M.D., Canudas-Romo, V., and Vaupel, J.W. (2018). The double-gap life expectancy forecasting model. *Insurance: Mathematics and Economics* 78: 339–350. doi:10.1016/j.insmatheco.2017.09.011.
- Perla, F., Richman, R., Scognamiglio, S., and Wüthrich, M. (2021). Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal* 7: 572–598.
- Raftery, A.E., Chunn, J.L., Gerland, P., and Ševčíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography* 50(3): 777–801. doi:10.1007/s13524-012-0193-x.
- Richman, R. (2020). AI in actuarial science – a review of recent advances - part 1. *Annals of Actuarial Science* 1–23. doi:10.1017/S1748499520000238.
- Richman, R. (2021). Mind the gap - safely incorporating deep learning models into the actuarial toolkit. *Available at SSRN* 7. doi:10.2139/ssrn.3857693.
- Richman, R. and Wüthrich, M. (2021). A neural network extension of the Lee-Carter model to multiple populations. *Annals of Actuarial Science* 15(2): 346–366. doi:10.1017/S1748499519000071.
- Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning representations by back-propagating errors. *Nature* 323. doi:https://doi.org/10.1038/323533a0.
- Scognamiglio, S. (2022). Calibrating the Lee-Carter and Poisson Lee-Carter via neural networks. *ASTIN Bulletin* 52(2): 519–561. doi:10.1017/asb.2022.5.
- Torri, T. and Vaupel, J.W. (2012). Forecasting life expectancy in an international context. *International Journal of Forecasting* 28(2): 519–531. doi:10.1016/j.ijforecast.2011.01.009.
- United Nations (2019). *World Population Prospects 2019: Methodology of the United Nations population estimates and projections*. Department of Social Affairs.
- Ševčíková, H., Li, N., Kantorová, V., Gerland, P., and Raftery, A.E. (2016). Age-specific mortality and fertility rates for probabilistic population projections. In: *Dynamic Demographic Analysis*. Springer Series on Demographic Methods and Population Analysis: 285–310, vol. 39. doi:10.1007/978-3-319-26603-915.

## Appendix

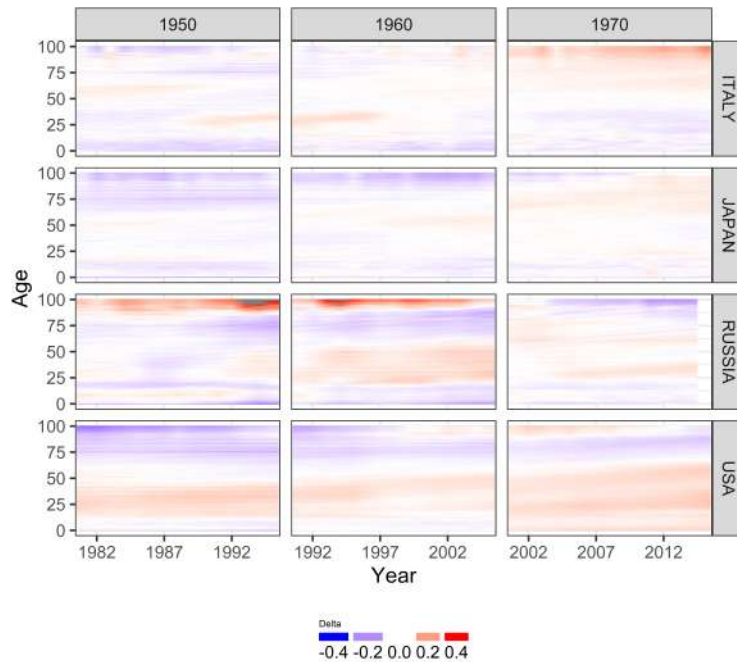
**Figure A-1: Relative differences ( $\Delta_{a,t}$ ) between estimates and the observed death rate by age for each model. Red hues indicate that the model underestimates mortality, while blue hues indicate overestimation. The male test period took place between 2001–2015.**



**Figure A-2: Estimated age-specific male log-mortality rates  $\log(m_{a,t})$  for the mp-DNN model by country for 2005, 2010 and 2014 based on the training period from 1970–2000. Black dots are the observed log-mortality rates.**



**Figure A-3: Relative differences ( $\Delta_{a,t}$ ) between estimates and the observed death rate by age for the mp-DNN model. Red hues indicate that the model underestimates mortality, while blue hues indicate overestimation. The male test period took place between 2001–2015.**



**Table A-1: Out-of-sample test: MAE and RMSE for DNN, linear-link, and Ševčíková et al. (2016) by country and sex. The estimation period for males took place between 1981–1995, 1991–2005, and 2001–2015.**

Country	Model	1981–1995		1991–2005		2001–2015	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
<i>Italy</i>	DNN	<b>0.1419</b>	<b>0.2203</b>	<b>0.1107</b>	<b>0.1493</b>	<b>0.1194</b>	<b>0.1566</b>
	linear-link	0.1840	0.2846	0.1993	0.2780	0.1541	0.2158
	Ševčíková et al. (2016)	0.1477	0.2596	0.1112	0.1701	0.1274	0.1992
<i>Japan</i>	DNN	0.1049	<b>0.1241</b>	0.0943	0.1264	0.0986	0.1254
	linear-link	0.1585	0.2088	0.1014	0.1392	0.1011	0.1480
	Ševčíková et al. (2016)	<b>0.1042</b>	0.1348	<b>0.0874</b>	<b>0.1205</b>	<b>0.0684</b>	<b>0.0958</b>
<i>United States</i>	DNN	<b>0.0746</b>	<b>0.1088</b>	<b>0.0730</b>	<b>0.0907</b>	<b>0.0955</b>	<b>0.1127</b>
	linear-link	0.1020	0.1561	0.1029	0.1437	0.1029	0.1367
	Ševčíková et al. (2016)	0.07974	0.1218	0.0907	0.1310	0.1085	0.1437
<i>Russia (2014)</i>	DNN	-	-	0.1989	<b>0.2366</b>	<b>0.1497</b>	0.2412
	linear-link	-	-	<b>0.1739</b>	0.2544	0.1951	0.3067
	Ševčíková et al. (2016)	-	-	0.2287	0.2940	0.1515	<b>0.2246</b>

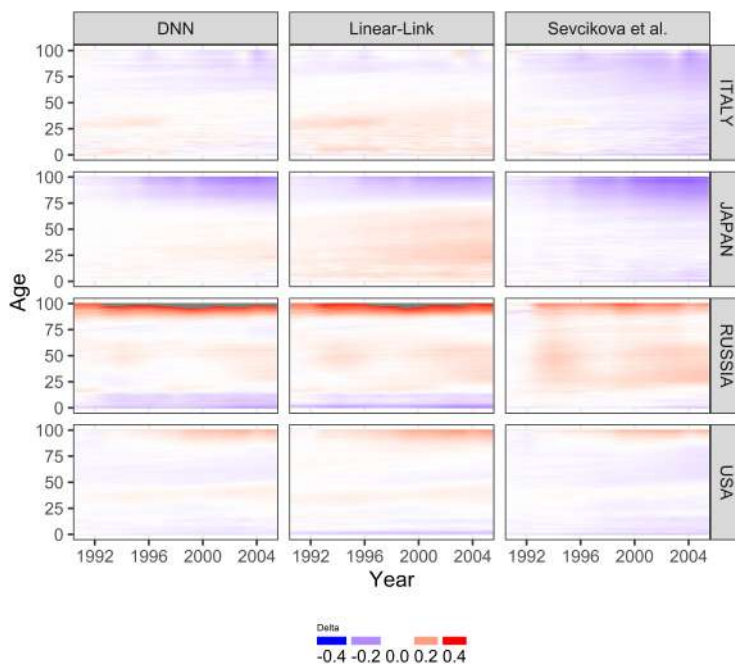
**Table A-2: Out-of-sample test: MAE and RMSE for mp-DNN by country and sex. The estimation period for males took place between 1981–1995, 1991–2005, and 2001–2015.**

Country	Model	1981–1995		1991–2005		2001–2015	
		MAE	RMSE	MAE	RMSE	MAE	RMSE
<i>Italy</i>	mp-DNN	0.1276	0.1971	0.0972	0.1659	0.1257	0.1603
<i>Japan</i>	mp-DNN	0.117	0.1683	0.0875	0.1232	0.08049	0.1095
<i>United States</i>	mp-DNN	0.2093	0.2614	0.1451	0.1805	0.2535	0.3038
<i>Russia (2014)</i>	mp-DNN	0.1356	0.1854	0.1849	0.2267	0.08918	0.1216

**Table A-3: Hyperparameter selection for single population model by country and sex. The estimation period took place between 1981–1995, 1991–2005, and 2001–2015.**

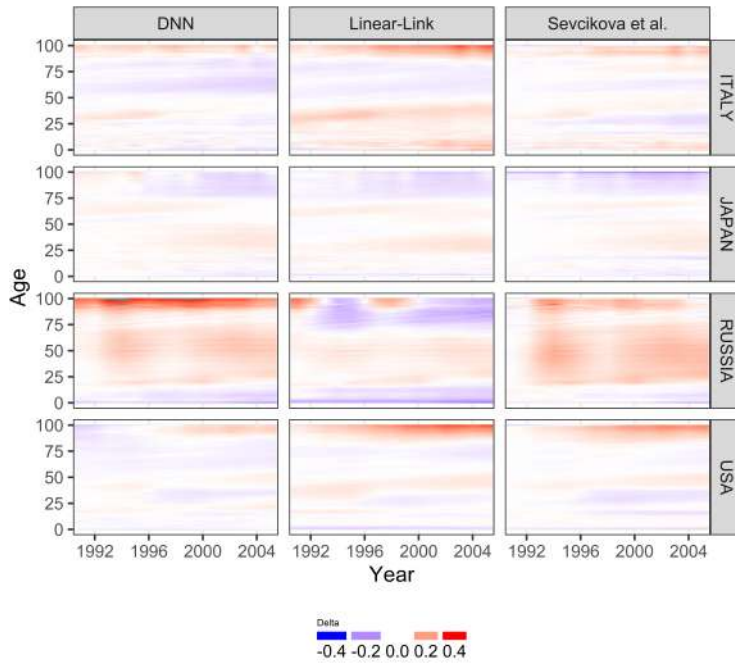
Country	1981-1995		1992-2005		2001-2015	
	NEURONS	EPOCHS	NEURONS	EPOCHS	NEURONS	EPOCHS
<i>Italy Male</i>	150	800	150	800	200	800
<i>Japan Male</i>	350	800	150	800	600	350
<i>United States Male</i>	150	800	250	800	105	300
<i>Russia Male</i>	-	-	210	800	210	800
<i>Italy Female</i>	250	800	150	400	600	350
<i>Japan Female</i>	200	800	100	450	200	800
<i>United States Female</i>	150	800	80	800	700	600
<i>Russia Female</i>	-	-	190	800	200	800

**Figure A-4:** Relative differences ( $\Delta_{a,t}$ ) between estimates and the observed death rate by age for each model. Red hues indicate that the model underestimates mortality, while blue hues indicate overestimation. The female test period took place between 1991–2005.

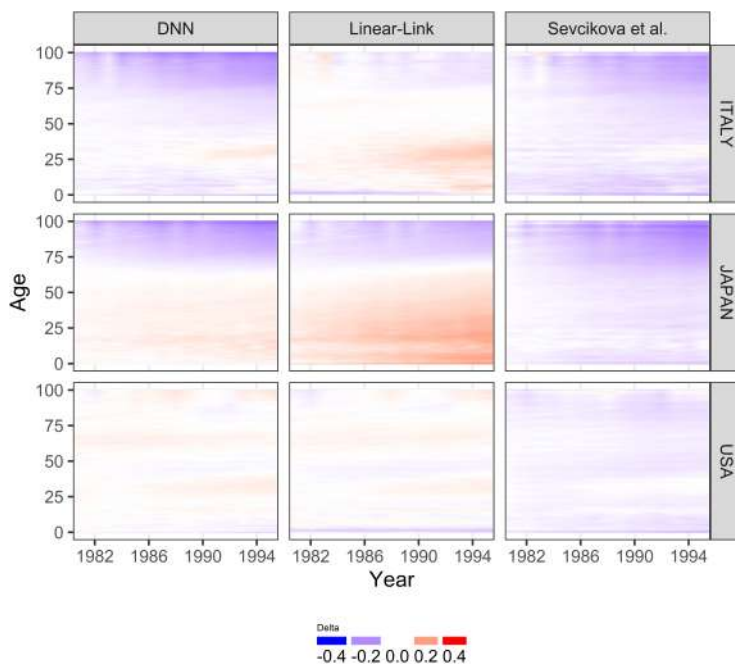




**Figure A-5: Relative differences ( $\Delta_{a,t}$ ) between estimates and the observed death rate by age for each model. Red hues indicate that the model underestimates mortality, while blue hues indicate overestimation. The male test period took place between 1991–2005.**



**Figure A-6:** Relative differences ( $\Delta_{a,t}$ ) between estimates and the observed death rate by age for each model. Red hues indicate that the model underestimates mortality, while blue hues indicate overestimation. The female test period took place between 1981–1995.



**Figure A-7: Relative differences ( $\Delta_{a,t}$ ) between estimates and the observed death rate by age for each model. Red hues indicate that the model underestimates mortality, while blue hues indicate overestimation. The male test period took place between 1981–1995.**

