

# Disaggregating Death Rates of Age-Groups Using Deep Learning Algorithms

Andrea Nigri<sup>1</sup> , Susanna Levantesi<sup>2</sup>,  
and Salvatore Scognamiglio<sup>3</sup>

## Abstract

Reliable estimates of age-specific vital rates are crucial in demographic studies, while ages are, in most cases, commonly grouped in bins of five years. Indeed, public health and national systems require single age-specific data to achieve accurate social planning. This paper introduces a deep learning approach for splitting the abridged death rates, providing a more comprehensive perspective on the indirect age-specific vital rates estimation from grouped data. Additionally, we contribute to the existing literature by introducing a multi-population (countries and genders) approach, providing reliable estimates considering the heterogeneity of longevity dynamics over age, years, and across populations. We also contribute to the state of the art in indirect estimation by introducing, for the first time, a multi-population indirect estimation leveraging subnational data. Our model accurately captures mortality dynamics by age over time and among different populations. We prove the model's ability to estimate reliable predictions of age-specific mortality rates by also studying how the hyperparameters' choice affects the model reliability and analyzing the age-specific relative differences between the real and the estimated mortality rates.

## Keywords

mortality modeling, ungroup, deep learning

## 1. Introduction

Monitoring changes and inequality among populations is a prime aim in assessing population dynamics and social and public policies. Thus reliable predictions of age-specific

<sup>1</sup>Department of Economics, Management and Territory, University of Foggia, Foggia, Italy

<sup>2</sup>Department of Statistics, Sapienza University of Rome, Rome, Italy

<sup>3</sup>Department of Management and Quantitative Studies, University of Naples—Parthenope, Napoli, Italy

### Corresponding author:

Andrea Nigri, Department of Economics, Management and Territory, University of Foggia, Via Da Zara, 11 – 71121, Foggia 71122, Italy.

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



vital rates are crucial in demographic studies. Data deficiency usually refers to incomplete or misreported information, such as age exaggeration and age heaping in both death and population data.

Despite single age-specific data being desirable, ages are commonly grouped in bins of five years in most cases. It is the case of demographic data, which is followed by a broad or open-ended age class for older ages. Several methods have been proposed for the disaggregation problem of historical data and data from developing countries that lack functional systems of vital registration (Liu et al. 2011; McNeil et al. 1977; Schmertmann 2012; Smith et al. 2004). Using countries from the Human Fertility Database (HFD), Liu et al. (2011) derived age-specific fertility rates from abridged data comparing ten different methods. The authors concluded that the modified Beers method (Beers 1945) provided the best fit. Similarly, Schmertmann (2012) using schedules observed in the HFD and the US Census International Database (IDB) proposed a calibrated spline (CS) as a more accurate and flexible alternative to the Beers interpolation method that requires more computation.

Among mortality modeling, the main approaches aimed at ungrouping histograms or abridged life tables were based on parametric assumptions for the underlying distribution (Hsieh 1991; Kostaki 1991; Kostaki and Panousis 2001), also for fitting a non-parametric density to binned data are generally used histosplines (Boneva et al. 1971), kernel density estimators (Blower and Kelsall 2002) and local likelihood. One of the most prominent frameworks has been proposed by Rizzi et al. (2015) who developed a versatile method for ungrouping histograms based on the composite link model with a penalty added to ensure the smoothness of the target distribution. Estimates are obtained by maximizing a penalized likelihood. Further, Rizzi et al. (2016) identify and compare the performance of five non-parametric methods for ungrouping count data, two spline interpolation methods, two kernel density estimators, and the penalized composite link model introduced by Rizzi et al. (2015). They found that the latter model outperforms the other four when data are grouped in wide age classes, or classes are open-ended.

Nevertheless, Rizzi's model relies on a parametric framework, specifically the Poisson distribution, which assumes a certain regularity in demographic distributions. However, as observed in real-world scenarios, demographic patterns often deviate from parametric assumptions, exhibiting complexities such as asymmetries or bimodal trends. Additionally, working with subpopulations with a limited number of deaths, as often encountered in demographic studies, can further jeopardize the precision of estimates due to probabilistic assumptions.

It is worth pointing out that, Rizzi's method, akin to other ungrouping approaches, is centered around estimating individual populations. Notably, no prior study has proposed a multi-population approach encompassing both gender and country variations. Attempting such estimations independently for each population might not ensure accurate results. Indeed, estimations often require harmonization, comparability, and coherence among countries and gender. Working on a single-gender population, the models above do not guarantee the latter properties, jeopardizing the comparability across time and among countries. Thus relying on coarsely grouped data may hamper precise data analysis. In contrast to the present literature, our proposed methodology not only addresses these limitations but also introduces improvements. Therefore, in this paper,

we introduce a multi-population (countries and genders) approach for splitting the abridged demographic rates, providing a more comprehensive perspective on the age-specific vital rates estimation from grouped data. The model leverages deep learning algorithms based on deep neural networks (DNN) to uncover age-specific vital rates. It is worth pointing out that the ungrouping methods discussed above (for example, Beers 1945; Rizzi et al. 2015; Schmertmann 2012), besides working on a single population, obtain the ungrouped estimates as a latent realization of an underlying process, in most cases also using smoothing interpolations. Therefore, it does not require supervised learning, and thus, the Train versus Test splitting, as in a deep learning approach, is unnecessary. In light of that, due to the different methodological frameworks, the comparison with existing literature would be inadequate since any comparison against the proposed model would be unreliable.

Deep learning has shown promising results in many applications enhancing a general interest in this methodology to solve complex problems, make predictions, extract information from data, and provide reliable estimates.

The choice of employing deep learning, over traditional machine learning methods such as regression trees, random forest, or XGBoost algorithms is supported by various factors stemming from the abilities of deep learning architectures to capture intricate relationships and patterns in complex datasets, to handle non-linear relationships within the data. Furthermore, we can mention end-to-end learning, eliminating the need for manual feature engineering and preprocessing steps, capacity for representation learning (as the model can automatically discern relevant features), and finally, handling high-dimensional data. For a comprehensive treatment of these concepts, the reader can refer to Bengio et al. (2000).

While the advantages of deep learning are significant, it's essential to note that the choice between deep learning and traditional methods depends on factors such as the nature of the data, the complexity of the problem, and the availability of computational resources. Where intricate patterns in large and complex datasets need to be uncovered, deep learning often proves superior. Indeed, throughout the literature on non-traditional methods in mortality modeling, the literature on machine learning (ML) and deep learning (DL) is well-distinct.

Contributions of deep learning in longevity have been proposed in order to forecast demographic time series using recurrent neural networks (Levantesi et al. 2022; Nigri et al. 2021), and also in the field of actuarial science to predict death rates (see, e.g., Nigri et al. 2019; Perla et al. 2021; Richman and Wüthrich 2021; Scognamiglio 2022). The literature devoted to the longevity model for indirect estimation is based on the deep learning model. Indeed, more recently, Nigri et al. (2022d) formalized a deep neural networks approach to indirectly derive age-specific mortality from observed or predicted life expectancy by leveraging deep learning algorithms akin to demography's indirect estimation techniques.

We contribute to the discussion on the ability of deep learning to provide reliable predictions by using it to ungroup mortality data for multiple populations. More specifically, we build up a DNN model to estimate the multi-population age-specific death rates from grouped death rates in five-year age classes, except for the oldest age class (100–110), which is wider. We assess the model prediction performance on the out-of-sample

data through traditional error measures, Root Mean Square Error and Mean Absolute Error. These measures allow for detecting the aggregate prediction ability of the model.

We take a significant stride in advancing the indirect estimation literature by extending the proposed method to the sub-national level. In a context where all indirect methods have traditionally been designed for national-level applications, a multi-population (at the regional level) indirect model signifies noteworthy progress. This is particularly significant because subnational data may be subject to stochastic variation in vital event numbers owing to the smaller population size, thus putting the indirect model to the hard test. The proposed model shows remarkable results in terms of accuracy and ability to capture longevity dynamics even at the subnational level.

The proposed model represents an advance in mortality modeling, offering the advantage of an indirect and complementary way to approximate age-specific death rates. It can be valuable in contexts where population-level mortality studies are hindered by financial or time constraints for national registries that do not support the open data system.

The remainder of the paper is structured as follows. Section 2 introduces the fundamentals of neural networks. Section 3 describes the specific framework of the neural network model for ungrouping mortality data. Section 4 presents the implementation of the neural network model and its structural reliability. Section 5 provides the results of the numerical experiment. Finally, Section 6 concludes the paper.

## 2. Neural Networks

Neural networks (NN) are high-dimensional and non-linear regression models that have achieved notable results in several fields such as computer vision and natural language processing. They consist of interconnected computational units, called neurons in the NN jargon, arranged on different layers that learn from data using training algorithms. The weights connect the units on the layers, and the connection configuration defines different kinds of NN. This section formally introduces the Fully-Connected Network (FCN) and the Embedding Network (EN) layers which have been used in this research.

Let  $x \in \mathbb{R}^{q_0}$  be the input vector; a FCN layer with  $q_1 \in \mathbb{N}$  units is a vectorial function that maps  $x$  to a  $q_1$ -dimensional real-valued space:

$$z^{(1)} : \mathbb{R}^{q_0} \rightarrow \mathbb{R}^{q_1}, \quad x \rightarrow z^{(1)}(x) = \left( z_1^{(1)}(x), z_2^{(1)}(x), \dots, z_{q_1}^{(1)}(x) \right)^T.$$

The output of each unit is a new feature  $z_j^{(1)}(x)$  which is a non-linear function of  $x$  :

$$z_j^{(1)}(x) = \phi \left( w_{j,0}^{(1)} + \sum_{i=1}^{q_0} w_{j,i}^{(1)} x_i \right) \quad j = 1, 2, \dots, q_1,$$

where  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  is the activation function,  $w_{j,0}^{(1)}$  indicates the bias/intercept (subscript 0 does not refer to a node in the previous layer but indicates that  $w_{j,0}^{(1)}$  does not depend on  $x$ ), and  $w_{j,i}^{(1)} \in \mathbb{R}$  represent the weights. In matrix form, the output  $z^{(1)}(x)$  of the FCN layer can be written as:

$$z^{(1)}(x) = \phi(w_0^{(1)} + W^{(1)}x).$$

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

$$y = \phi(w_0^{(o)} + \langle \omega^{(o)}, z^{(1)}(x) \rangle),$$

where  $w_0^{(o)} \in \mathbb{R}$ ,  $\omega^{(o)} \in \mathbb{R}^{q_1}$  and  $\langle \cdot, \cdot \rangle$  denotes the scalar product in  $\mathbb{R}^{q_1}$ . The upper index  $(o)$  of  $w_0^{(o)}$  and  $\omega^{(o)}$  emphasizes that these weights are related to the output layer. If the network is deep, the vector  $z^{(1)}(x)$  is used as input in the next layer for computing new features and so for the following layers. Let  $h \in \mathbb{N}$  be the number of hidden layers (depth of network), and  $q_k \in \mathbb{N}$ , for  $1 \leq k \leq h$ , be a sequence of integers that indicates the dimension of each FCN layer (widths of layers). A deep FCN can be described as follows:

$$x \rightarrow z^{(h:1)}(x) = (z^{(h)} \circ \dots \circ z^{(1)})(x) \in \mathbb{R}^{q_h},$$

where the vectorial functions  $z^{(k)} : \mathbb{R}^{q_{k-1}} \rightarrow \mathbb{R}^{q_k}$  have the same structure, and  $W^{(k)} = (w_j^{(k)})_{1 \leq j \leq q_k} \in \mathbb{R}^{q_k \times q_{k-1}}$ ,  $w_0^{(k)} \in \mathbb{R}^{q_k}$ , for  $1 \leq k \leq h$  are the network weights, and  $\circ$  is the composition operator. In the case of deep NN, the output layer uses the features extracted by the last hidden layer  $z^{(h:1)}(x)$  instead of those  $z^{(1)}(x)$ . The depth of the network  $h$  is a hyperparameter that should be suitably chosen. Indeed, a too deep NN would lead to *overfitting*, producing a model unable to generalize to new data points. One remedy is the application of regularization methods such as dropout. It is a *stochastic* technique that ignores some randomly chosen units during the network fitting. This is generally achieved by multiplying the output of the different layers by independent realizations of a Bernoulli random variable with parameter  $p \in [0, 1]$ . Mathematically, the introduction of the dropout in the  $k$ -th FCN layer induces the following structure:

$$\begin{aligned} r_j^{(k)} &\sim \text{Bernoulli}(p) \\ \dot{z}^{(k-1)}(x) &= \mathbf{r}^{(k)} * z^{(k-1)}(x) \\ z^{(k)}(x) &= \phi(w_0^{(k)} + W^{(k)}\dot{z}^{(k-1)}(x)), \end{aligned}$$

where  $*$  denotes the element-wise product and  $\mathbf{r}^{(k)}$  is a vector of independent Bernoulli random variables, each of which has a probability  $p$  of being 1. This mechanism leads to the value of some elements being reset to zero. FCN layers are useful tools for processing numerical data, however, sometimes, information is available as categorical variables. In the statistical literature, the standard procedures for dealing with categorical variables are the one-hot and dummy encoding. Nevertheless, these coding schemes produce high-dimensional sparse vectors, which often leads to computational issues when there are many categorical features, or one of them presents many levels. Embedding is an innovative technique to analyze categorical variables. They appear for the first time in the Natural Language Processing context (see Bengio et al. 2000), but recently, they are becoming very popular in the mortality literature (Perla et al. 2021; Richman and Wüthrich 2021; Scognamiglio 2022).

An Embedding Network (EN) Layer maps the levels of a categorical variable into a low-dimensional real-valued space. The dimensionality of the new space  $q_{\mathcal{L}} \in \mathbb{N}$  represents a hyperparameter chosen by the modelers. The levels of the categorical variable are mapped into a real-valued  $\mathbb{R}^{q_{\mathcal{L}}}$ -dimensional space, and the coordinates of the level are parameters to learn during the training process (Guo and Berkahn 2016). The distance of the levels in the new learned space reflects the similarity of levels concerning the target variable: similar levels will have a small euclidean distance, whereas very different categories will have a large one. Formally, let  $\mathcal{L} = \{l_1, l_2, \dots, l_{n_{\mathcal{L}}}\}$  be the set of categories of the qualitative variable and  $n_{\mathcal{L}}$  be its cardinality. An embedding layer is a mapping

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

The number of embedding parameters to learn during the network calibration is  $n_{\mathcal{L}}q_{\mathcal{L}}$ .

The elements of the matrices  $W^{(k)}$  and of the bias vectors  $w_0^{(k)}, k = 1, \dots, h$  of the FCN, and the coordinates of the levels in the new embedding space  $z_l(l), \forall l \in \mathcal{L}$  must be appropriately calibrated. Denoting by  $\theta$  the vector containing all the network parameters, one could argue the training process consists of an unconstrained optimization problem where chosen a suitable loss function  $L(\theta)$ ; its minimum is sought. The NN training is generally carried out using the Gradient Descent algorithm or one of its extensions, where the updating of the weights is based on the gradient of the loss function  $L(\theta)$ . The weights are iteratively adjusted to decrease the error of the network outputs with respect to some reference values. Let  $\theta^{(t)}$  be the vector of parameters at time  $t$ , the updating rule can be written as follows:

$$\theta^{(t+1)} = \theta^{(t)} + \eta \nabla L(\theta)$$

where  $\eta \in [0, 1]$  is the learning rate,  $\theta^{(0)}$  is the initial vector. We remark that the complexity of the training grows with the number of layers and units per layer in the network architecture.

### 3. The NN Model for Ungrouping Mortality Data

Let  $X = \{x_1, x_2, \dots, x_\omega\}$  be the set of individual ages,  $T = \{t_1, t_2, \dots, t_n\}$  the set of the calendar years, and  $I$  the set of populations considered. In particular, we consider a set of populations which differ among them for the region and the gender such that  $I = G \times R = \{\text{Male, Female}\} \times R$  (where  $R$  denotes the set of regions). The death rate at age  $x$ , at time  $t$  for the population  $i = (r, g)$  is defined as:

$$m_{x,t,r,g} = D_{x,t,r,g} / E_{x,t,r,g},$$

where  $D_{x,t,r,g}$  is the death count, and  $E_{x,t,r,g}$  is the number of exposure-to-risk. Sometimes the death rates are available for age-groups rather than for single ages. Let  $C = \{c_1, c_2, \dots, c_L\}$  be the set of contiguous age-groups with  $L \ll \omega$ . The death rate at the class  $c$  at time  $t$  in the population  $i = (r, g)$  is defined as:

$$m_{c,t,r,g} = \frac{D_{c,t,r,g}}{E_{c,t,r,g}} = \frac{D_{v(x),t,r,g}}{E_{v(x),t,r,g}} = \frac{\sum_{x \in \mathcal{X}} 1_{v(x)=c} D_{x,t,r,g}}{\sum_{x \in \mathcal{X}} 1_{v(x)=c} E_{x,t,r,g}}$$

$v : X \rightarrow C$  is a surjective function which assigns to each age  $x \in \mathcal{X}$  an age-group  $c$ ,  $v(x) \in \{c_1, c_2, \dots, c_L\}$  and  $v^{(c_j)} := v^{-1}(\{c_j\})$ ,  $j = 1, \dots, L$  denote the set of ages belonging of the  $j$ -th age-group. For a given population  $i = (r, g)$  and a given calendar year  $t$ , the problem consists of ungrouping the age-specific death rates  $\{m_{x,t,r,g}\}_{x \in \mathcal{X}}$  from the death rates of the age-groups  $\{m_{r,g}\}_{c \in C}$  in log scale.

We argue that the single-age death rate  $\{m_{x,t,r,g}\}$  in log scale can be modeled as a function that depends on some inputs: the age  $x$ , region  $r$ , the gender  $g$ , the time  $t$ , and the death rate in log scale of the age-group  $v(x)$  to which  $x$  belongs. In other words, we assume the existence of the function:

$$f : X \times R \times G \times T \times R \rightarrow R, \quad (x, r, g, t, \log(m_{v(x),t,r,g})) \rightarrow \log(m_{x,t,r,g}).$$

$f$  is unknown and potentially complex. In this framework, we employ deep NNs to find and approximation  $\hat{f}$  since they are known as universal function approximators (Hornik et al. 1989). The proposed NN model consists of some EN and FCN layers. First, the EN layers are used for processing the categorical variables and extracting from them the optimal numerical features with respect to the quantity of interest. In particular, we apply EN layers to the variable related to age, region, and gender. Second, the features extracted from the EN layers are concatenated with the numerical inputs (calendar year and related age-group death rate) and introduced through the FCNs to determine the single-age death rate. Let  $q_X, q_R, q_G \in \mathbb{N}$  be the hyper-parameter values defining the size of three embedding layers. They map  $r \in R$ ,  $g \in G$ , and  $x \in X$  into real-valued vectors:

$$\begin{aligned} z_X : X &\rightarrow \mathbb{R}^{q_X}, & x &\mapsto z_X(x) = \left( z_{X,1}(x), z_{X,2}(x), \dots, z_{X,q_X}(x) \right)^\top. \\ z_R : R &\rightarrow \mathbb{R}^{q_R}, & r &\mapsto z_R(r) = \left( z_{R,1}(r), z_{R,2}(r), \dots, z_{R,q_R}(r) \right)^\top, \\ z_G : G &\rightarrow \mathbb{R}^{q_G}, & g &\mapsto z_G(g) = \left( z_{G,1}(g), z_{G,2}(g), \dots, z_{G,q_G}(g) \right)^\top. \end{aligned}$$

$z_R(r)$  is the vector of the regional-specific features,  $z_G(g)$  is gender-specific, while  $z_X(x)$  is the vector of the age-specific features. We also define the complete vector of features concatenating the embedding weights and the numerical inputs:

$$z(x, t, r, g) = \left( z_X(x), t, z_R(r), z_G(g), \log(m_{v(x),t,r,g}) \right)^\top \in \mathbb{R}^{q_R + q_G + q_X + 2}$$

$z(x, t, r, g)$  is further processed with a three FCN layers of size  $(q_1, q_2, 1) \in \mathbb{N}^3$ . Since we are interested in modeling the single-age death rates, we design an architecture where the last layer of the network (output layer) has a size equal to 1 ( $q_3 = 1$ ). To avoid overfitting, we introduce dropout among the FCN layers. In notation, the mechanism of the network can be formalized as:

$$\begin{aligned} z^{(1)}(x, t, r, g) &= \phi \left( w_0^{(1)} + W^{(1)} z(x, t, r, g) \right) \\ r_j^{(1)} &\sim \text{Bernoulli}(p_1) \\ \dot{z}^{(1)}(x, t, r, g) &= r^{(1)} * z^{(1)}(x, t, r, g) \\ z^{(2)}(x, t, r, g) &= \phi \left( w_0^{(2)} + W^{(2)} \dot{z}^{(1)}(x, t, r, g) \right) \\ r_j^{(2)} &\sim \text{Bernoulli}(p_2) \\ \dot{z}^{(2)}(x, t, r, g) &= r^{(2)} * z^{(2)}(x, t, r, g) \\ \log(\hat{m}_{x,t,r,g}) &= \phi \left( w_0^{(3)} + w^{(3)} \dot{z}^{(2)}(x, t, r, g) \right), \end{aligned}$$

where  $p_1, p_2 \in [0, 1]$ , and  $w_0^{(1)}, w_0^{(2)}, w_0^{(3)}, W^{(1)}, W^{(2)}, w^{(3)}$  are the weights.

The weight matrices and bias vectors of the FCN layers and the parameters of the embedding layers need to be appropriately calibrated. These parameters are iteratively adjusted via the BP algorithm to minimize a specific loss function. Following the practice adopted in the mortality literature (Hainaut 2018; Perla et al. 2021; Richman and Wüthrich 2021), we fit our model using the Mean Squared Error (MSE) (Hainaut 2018; Perla et al. 2021; Richman and Wüthrich 2021). In such a case, the training of the network requires the minimization of the loss:



$$L(\theta) = \sum_{x \in X} \sum_{t \in T} \sum_{r \in R} \sum_{g \in G} \left( \log(m_{x,t,r,g}) - \log(\widehat{m}_{x,t,r,g}) \right)^2.$$

where  $\theta$ , denotes the NN parameters. The choice of the most appropriate loss function is also related to the modeling purpose. Indeed, if the object is to directly model death probabilities, an appropriate loss function could be the negative log-likelihood under the assumption of a binomial distribution of deaths.

### 4. Implementation

Let  $\left\{ \log(m_{v(x),t,r,g}) \right\}_{t=t_0}^{t_s}$ , for  $t_0 < t_s$ , be the country-specific observed grouped death rates in log scale. Then, each series is split into a train-validation set and a test set, where the first is used for fitting the model’s parameters, while the second is used to test the model’s prediction and calculate the error.

The best DNN setting during the training phase is used to obtain predictions in the test phase.

Hence, let  $t_\tau$ , with  $t_0 < t_\tau < t_s$ , be the calendar year corresponding to the last realization in the train-validation set. The values of  $\log(m_{v(x),t,r,g})$  over the period  $(t_0, t_\tau)$ ,  $\left\{ \log(m_{v(x),t,r,g}) \right\}_{t=t_0}^{t_\tau}$ , represent the input for train-validation, while the corresponding output is  $\left\{ \log(\widehat{m}_{x,t,r,g}) \right\}_{t=t_0}^{t_\tau}$ .

The values of grouped death rates over a subsequent period,  $\left\{ \log(m_{v(x),t,r,g}) \right\}_{t=t_\tau+1}^{t_s}$ , represent the input for test, while the corresponding output is  $\left\{ \log(\widehat{m}_{x,t,r,g}) \right\}_{t=t_\tau+1}^{t_s}$ . Thereby, denoting  $\widehat{f}_{nn}$  as a composition of functions defined based on the NN architecture, the model can be described by:

$$\left\{ \log(\widehat{m}_{x,t,r,g}) \right\}_{t=t_\tau+1}^{t_s} = \widehat{f}_{nn} \left( \left\{ \log(m_{v(x),t,r,g}) \right\}_{t=t_\tau+1}^{t_s} \mid \widehat{\theta} \right), \tag{1}$$

where  $\left\{ \log(\widehat{m}_{x,t,r,g}) \right\}_{t=t_\tau+1}^{t_s}$  are the death rates in log scale in single year age-group, in the test set obtained by  $\widehat{f}_{nn}$ , that involves the NN weights  $\widehat{\theta}$  estimated during the network training.

#### 4.1. Parameters and Structural Reliability Analysis of DNN

Deep learning modeling often outperforms statistical methods, especially for prediction tasks, approximating even the most complex functional structure. To achieve this result, it is necessary to identify the optimal NN setting (i.e., the number of hidden layers,

neurons, and parameters); generally, it is obtained by performing a fine-tuning phase. Nevertheless, NN is a model subject to different uncertainty sources that might affect the learning phase (Richman 2021), and the effect of hyperparameters' choice on model reliability is still an open question, often underrated.

The fine-tuning aims to choose a reliable DNN architecture, considering, for example, the number of hidden layers, and activation functions. The most suitable structure depends on the data type and is generally selected according to the validation error minimization. The outcome prediction sensitivity for different training setups is a common issue in deep learning. Despite the variability in the input data being considered in all inferential methods, when it comes to deep learning, one should also view the source of variability that originates from the optimization procedure.

We seek to justify our structural choice by examining the reliability of the Deep Neural Network (DNN) in response to changes in training conditions. Our focus is on understanding how predictions may vary each time the network is trained using different configurations and network architectures.

In the following sub-sections, we analyze a large space of combinations, moving to a proper subspace, using the changes highlighted in the table. We test and explore different options as we motivate the final setting choice, whose outcomes will be discussed in the results section. To summarize the performance of the methods and to evaluate their accuracy, we report the MAE and RMSE on the test sets given by:

$$\text{MAE} : \frac{\sum_{x \in \mathcal{X}} \sum_{t \in \mathcal{T}} |\log(m_{x,t,r,g}) - \log(\hat{m}_{x,t,r,g})|}{|\mathcal{X}| \cdot |\mathcal{T}|},$$

$$\text{RMSE} : \sqrt{\frac{\sum_{x \in \mathcal{X}} \sum_{t \in \mathcal{T}} [\log(m_{x,t,r,g}) - \log(\hat{m}_{x,t,r,g})]^2}{|\mathcal{X}| \cdot |\mathcal{T}|}}.$$

We carry out these experiments, comparing all possible combinations through the MAE and RMSE of the predictions averaged over the study countries. The main network architecture applies the dropout set to 10%. Here, we test whether the dropout tool within the intermediate network might provide better results. Tables 1 and 2 show that the adoption of dropout gains better performances in the architecture composed of one layer and Relu function, but when we use two and three layers, the dropout technique seems to drop the number of neurons drastically, increasing the error. The tables show that the best DNN framework is one with two layers using the Sigmoid activation function, which outperforms the other tested with better learning performance.

## 5. Numerical Experiments

We consider historical mortality data collected by the Human Mortality Database (HMD 2018) for all available countries and both genders. Aiming to assess the multi-population model robustness and consistency toward the historical data, we carry out an out-of-sample test. The splitting choice is influenced not merely by a rigid 70 to 30% split but

**Table 1.** Reliability Analysis of the DNN on Male Populations.

Layer	Error	Relu		Sigmoid		Softmax	
		Drop-out	No Drop-out	Drop-out	No Drop-out	Drop-out	No Drop-out
1	MAE	0.09	0.10	0.11	0.10	0.16	0.11
	RMSE	0.15	0.16	0.17	0.16	0.23	0.18
2	MAE	0.10	0.09	0.09	<b>0.08</b>	0.15	0.11
	RMSE	0.16	0.15	0.15	<b>0.14</b>	0.21	0.18
3	MAE	0.14	0.10	0.10	0.09	2.32	2.33
	RMSE	0.19	0.16	0.16	0.16	2.69	2.69

The values in bold refer to the best performances.

**Table 2.** Reliability Analysis of the DNN on Female Populations.

Layer	Error	Relu		Sigmoid		Softmax	
		Drop-out	No Drop-out	Drop-out	No Drop-out	Drop-out	No Drop-out
1	MAE	0.11	0.12	0.12	0.12	0.14	0.13
	RMSE	0.18	0.19	0.19	0.19	0.21	0.21
2	MAE	0.13	0.11	0.10	<b>0.10</b>	0.16	0.13
	RMSE	0.20	0.18	0.17	<b>0.17</b>	0.23	0.20
3	MAE	0.15	0.11	0.11	0.11	2.54	2.55
	RMSE	0.21	0.18	0.18	0.18	2.90	2.91

The values in bold refer to the best performances.

rather by the temporal progression of mortality trends, consequently, opting for a training-validation phase spanning thirty years (1970–2000), allocating the remaining years (2001–2015) for the model test. Specifically, the time frame 1970 to 2000 is used as a train-validation set, and the years 2001 to 2015 are used for model test. The train-validation set is, in turn, split into training and validation sets according to the 80 to 20% splitting rule, that is, 1970 to 1994 and 1995 to 2000.

We validate the DNN model performance using illustrative applications. These examples are dedicated to investigating whether the approaches can capture (a) regular and irregular trends over time and (b) dynamics of age-specific mortality improvements. The analysis includes numerical and graphical representations of the goodness of fit.

To assess the models' accuracy, we calculate the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) on the out-of-sample period, which in the present analysis corresponds to a 2001 to 2015 time window.

Figure 1 depicts the MAE and RMSE for countries over the time window we studied for females and males, respectively, considering the best DNN architecture. For the sake of completeness, we have included several plots (Figures 5, 6, 7) in the Supplemental Material (SM), depicting the RMSE related to the different populations for the tested architectures.

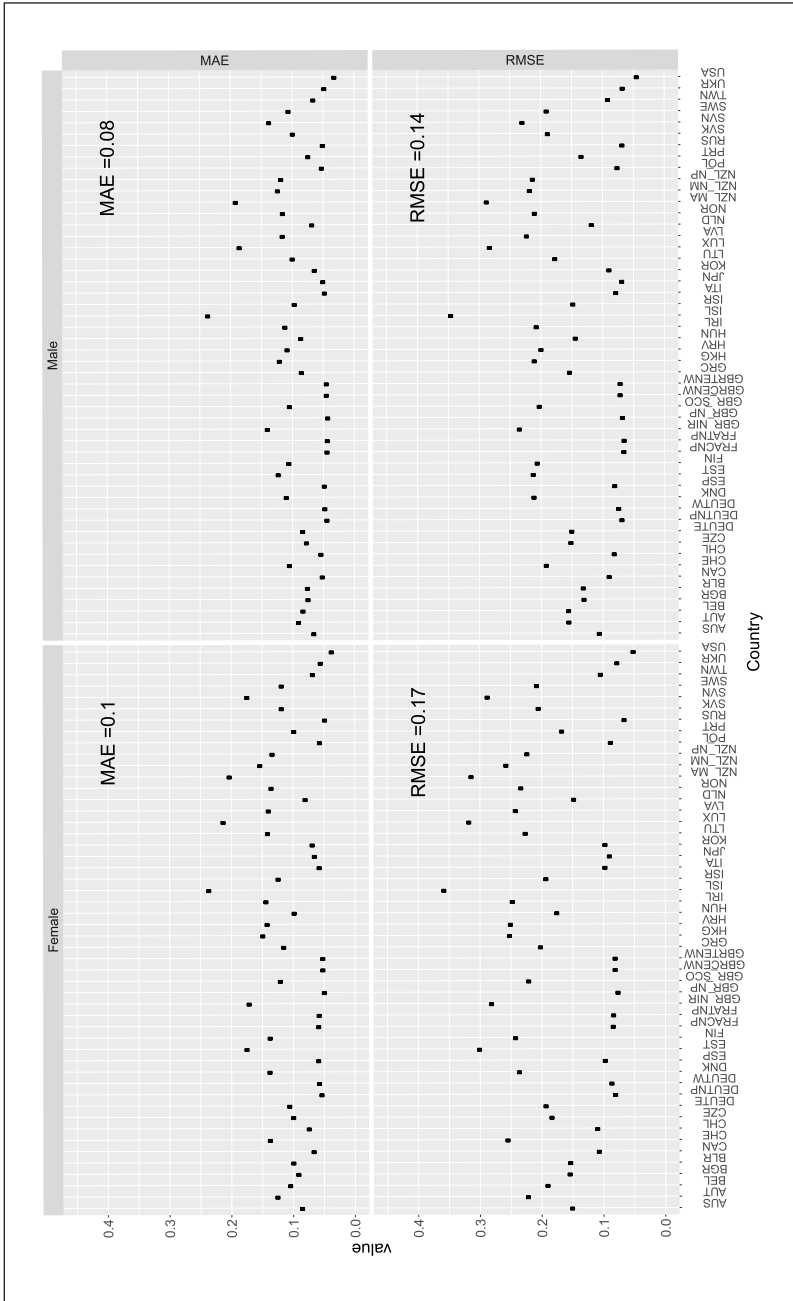


Figure 1. Out-of-sample MAE and RMSE obtained in the different countries with the two-layered DNN model with sigmoid activation and no dropout.

Overall, our DNN model provides remarkable accuracy. The USA is the country that presents the lowest MAE and RMSE values for both genders. Considering males, Germany's total population (DEUTNP), France's (both total [FRATNP] and civilian population [FRACNP]), and the UK's total population (GBR\_NP) follow. For females, Russia follows. Finally, Iceland shows the highest MAE and RMSE for both genders. On average, both the error measures are higher for female populations (MAE = 0.10 and RMSE = 0.17) compared to male populations (MAE = 0.08 and RMSE = 0.14).

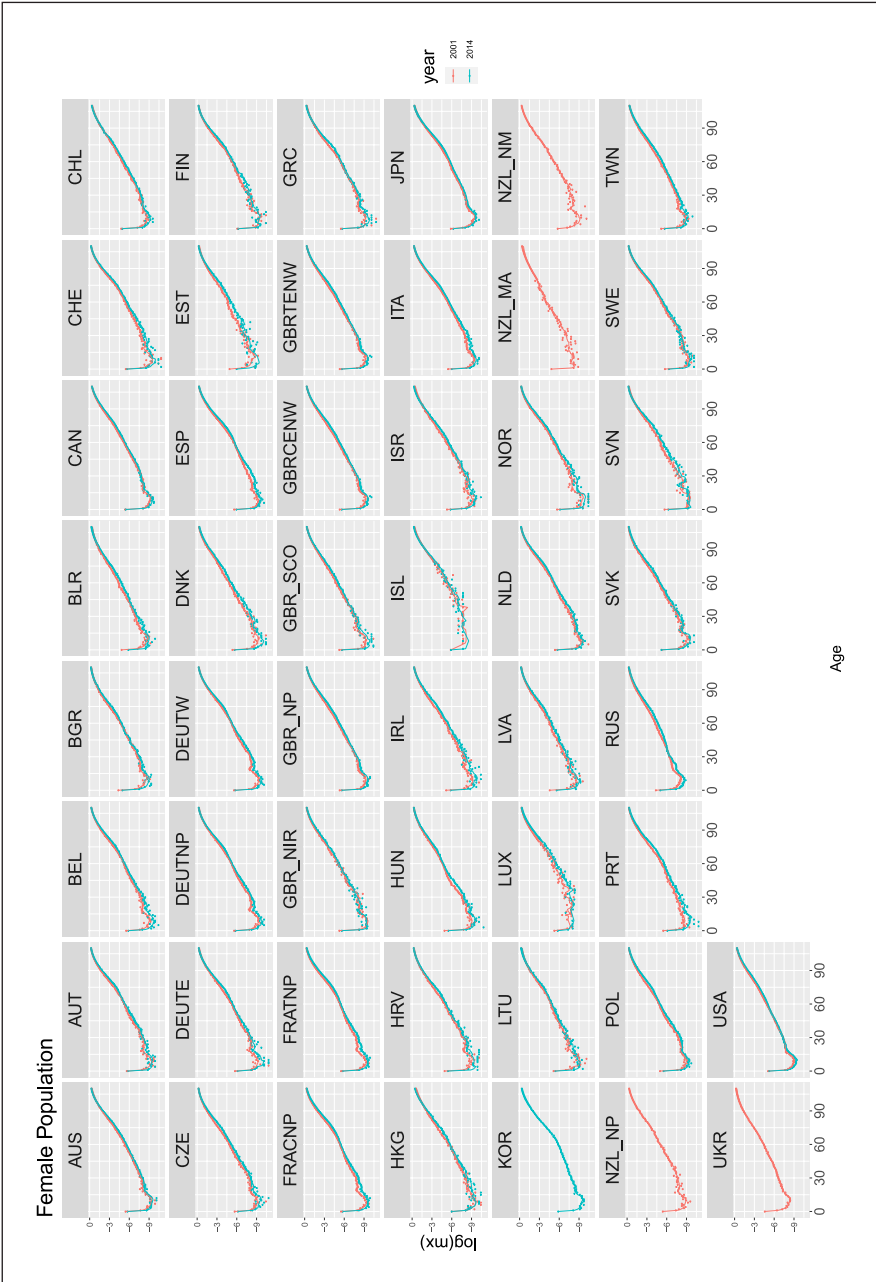
Figures 2 and 3, show age-specific death rates (in log scale) for females and males in the HMD in the years 2001 and 2014. The observed mortality profile is shown with dots and estimated ungrouped values from the models using the training period from 1970 to 2000, corresponding to the red line. Some countries, such as "UKR," do not display data for the year 2014 due to its unavailability in the temporal series provided by HMD.

Overall, the proposed model captures the general pattern of mortality, with a decreasing trend from birth to around age fifteen and increasing linearly from around age thirty. Indeed, the DNN model adequately captures the mortality patterns, even if a few countries fail to accurately capture the sharp decrease from infancy.

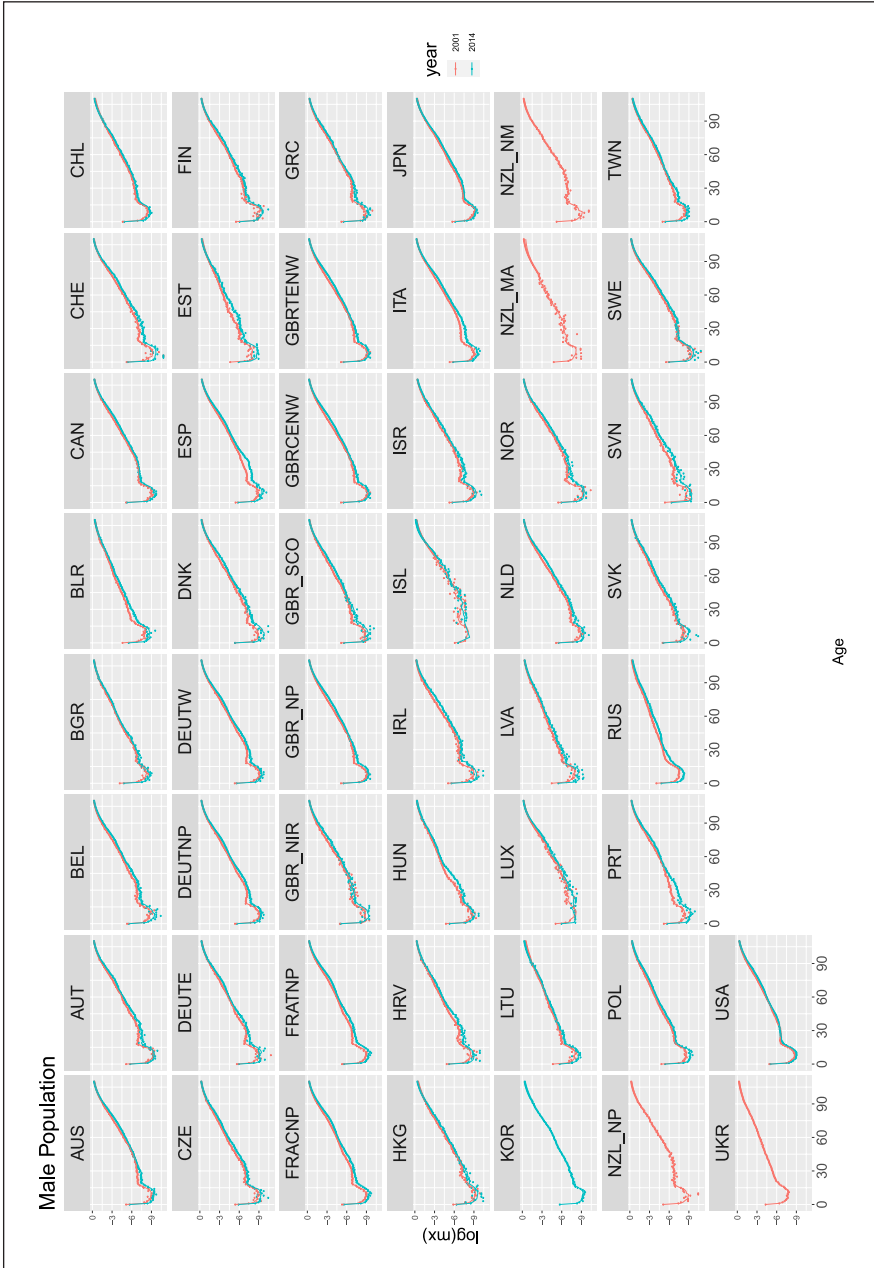
Estimating death rates is not an easy task, even more using indirect methods. Indeed capturing longevity dynamics can be challenging due to irregularity and peculiar behaviors that populations may play.

In the last decades, developed countries have seen an important mortality decline at all ages, involving a remarkable improvement in life expectancy, without evidence of deceleration. Therefore, it is important to investigate the age-specific differences between the real and the estimated mortality data according to specific phases of longevity evolution namely regularities, improvements, and stagnation, that can be represented by Japan, Italy, and the USA respectively. Specifically, Italian life expectancy exhibits a long-run transition with considerably upward shifts, converging to longevity records (Nigri et al. 2021a, 2022b, 2022c). This result has been achieved by a mix of a relatively healthy lifestyle and an efficient health system. In recent years, among longevous populations, Japan exhibits the most relevant regularities in leading the maximum human life expectancy at birth after a long period of low longevity. These peculiar dynamics make estimating Japan's mortality not straightforward. Even estimating the US mortality is a challenging task. After the first decade of the new millennium, the rise in US life expectancy stalled. Scholars bring evidence of a stagnating decline in cardiovascular disease mortality (Mehta et al. 2020).

In doing so, as a further tool for model assessment, we provide plots (Figures 1 and 2 in the SM), that show estimated versus observed values  $(\log m_{x,t,r,g}, \log \hat{m}_{x,t,r,g})$  for all countries, for male and female, to visualize the similarity between the actual and reconstructed death rates. Looking at our results, we underline that our findings can be very insightful in the demographic scenarios, being capable of adequately capturing the aforementioned dynamics over ages and years, in the context of longevity improvements (the case of Italy), longevity regularities (the case of Japan), and longevity stagnation (the case of USA).



**Figure 2.** Estimated ungrouped death rates by countries for years 2001 and 2014 based on the period 1970 to 2000 (training and validation). Dots refer to the observed rates. Solid lines are the DNN reconstructed ungrouped death rates. Females.



**Figure 3.** Estimated ungrouped death rates by countries for years 2001 and 2014 based on the period 1970 to 2000 (training and validation). Dots refer to the observed rates. Solid lines are the DNN reconstructed ungrouped death rates. Males.

### 5.1. Subnational Modeling

The development of indirect methods has been implemented to work at the national level. However, there is a considerable need and demand for subnational estimation, for example, the official statistics at the subnational level are essential for planning by local governments and the private sector and for health and social science research on subnational variation and inequality.

In this section, we extend the proposed method to the sub-national level. This task represents a crucial step to test and validate our model further since the subnational context is different from the national one. Indeed, some subnational areas can have populations so small that stochastic variation in numbers of vital events, usually ignored in population modeling, can have a significant impact on that area's population estimation, thus potentially affecting data quality.

The data for evaluating our proposed method comes from ISTAT, the Italian National Institute of Statistics, which serves as the authoritative source for official statistical information in Italy, encompassing diverse subjects, including mortality. The data, accessible through <http://dati.istat.it/>, comprises mortality statistics categorized by year, age, and region. Leveraging this dataset, we obtain a dataset spanning eighteen regions from 1974 to 2016, thereby facilitating comprehensive analyses. The model training procedure adhered to the methodology elucidated in Section 4, systematically exploring various network configurations and discerning optimal performance. The dataset's temporal and age-related granularity enabled the creation of a training-validation time frame (1974–2003), with the subsequent years (2004–2016) for model testing. The train-validation set is, in turn, split into training and validation sets according to the 80 to 20% splitting rule. Tables 3 and 4 reveal that the best DNN configuration refers to the one with three layers, devoid of dropout, and employs a Rectified Linear Unit (ReLU) activation function. It is conceivable that, when operating at the subnational level, an additional layer may enhance precision in estimation, particularly when dealing with datasets characterized by heightened stochasticity.

Figures 4 and 5 illustrate age-specific death rates (in logarithmic scale) for females and males across Italian regions in the years 2004 and 2016. The depicted mortality profile utilizes points for observed values and lines for estimated ungrouped values derived from models trained and validated during the 1974 to 2003 period.

It is imperative to emphasize that, although the model employs data from eighteen regions for training, the results exhibit sixteen regions due to the unavailability of data for the omitted regions during the years considered in the test set.

The deep neural network model adeptly captures mortality patterns, even at a subregional level, where data exhibit heightened stochasticity and diminished quality. Through this analysis, we substantiate the model's capacity not only for delivering high accuracy but also for demonstrating a remarkable degree of flexibility, accommodating both national and subnational data, as well as data of varying quality levels.

We can speculate that by working on subpopulations, the network requires an additional layer to achieve greater accuracy in estimation, particularly when dealing with data characterized by higher stochasticity. This aspect was not necessary for the model estimated on the HMD, where the additional layer, and therefore a more complicated



**Table 3.** Reliability Analysis of the DNN on Male Subnational Populations.

Layer	Error	Relu		Sigmoid		Softmax	
		Drop-out	No Drop-out	Drop-out	No Drop-out	Drop-out	No Drop-out
1	MAE	0.08	0.11	0.21	0.08	0.20	0.219
	RMSE	0.11	0.16	0.33	0.14	0.29	0.304
2	MAE	0.09	0.09	0.14	0.08	0.16	0.548
	RMSE	0.13	0.14	0.19	0.13	0.21	0.79
3	MAE	0.13	<b>0.07</b>	0.10	0.08	2.18	2.177
	RMSE	0.16	<b>0.11</b>	0.14	0.13	2.51	2.51

The values in bold refer to the best performances.

**Table 4.** Reliability Analysis of the DNN on Female Subnational Populations.

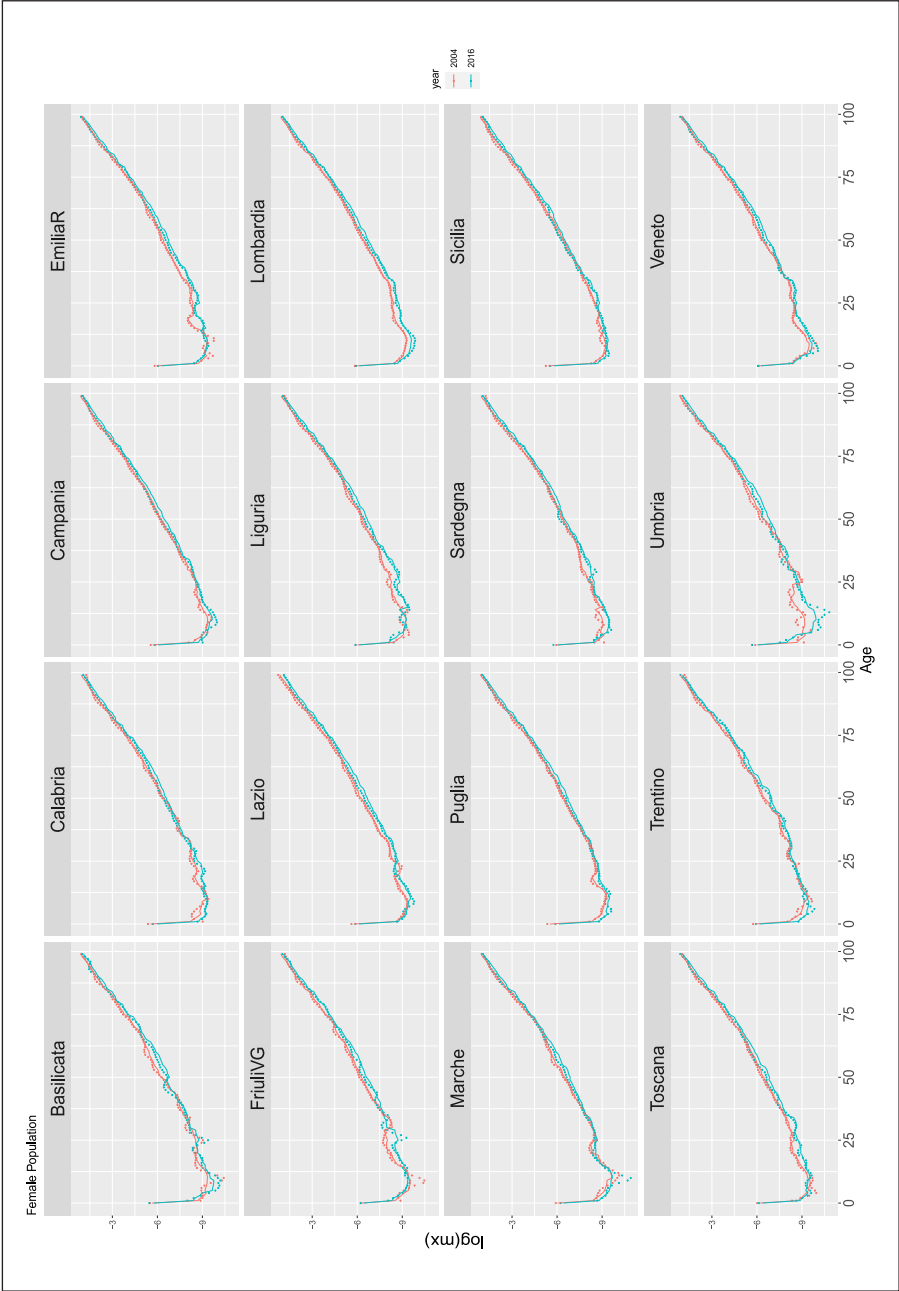
Layer	Error	Relu		Sigmoid		Softmax	
		Drop-out	No Drop-out	Drop-out	No Drop-out	Drop-out	No Drop-out
1	MAE	0.09	0.13	0.21	0.10	0.18	0.205
	RMSE	0.13	0.19	0.31	0.16	0.26	0.30
2	MAE	0.11	0.10	0.13	0.10	0.145	0.771
	RMSE	0.15	0.15	0.19	0.16	0.213	0.99
3	MAE	0.13	<b>0.08</b>	0.11	0.10	2.37	2.366
	RMSE	0.17	<b>0.12</b>	0.16	0.16	2.72	2.72

The values in bold refer to the best performances.

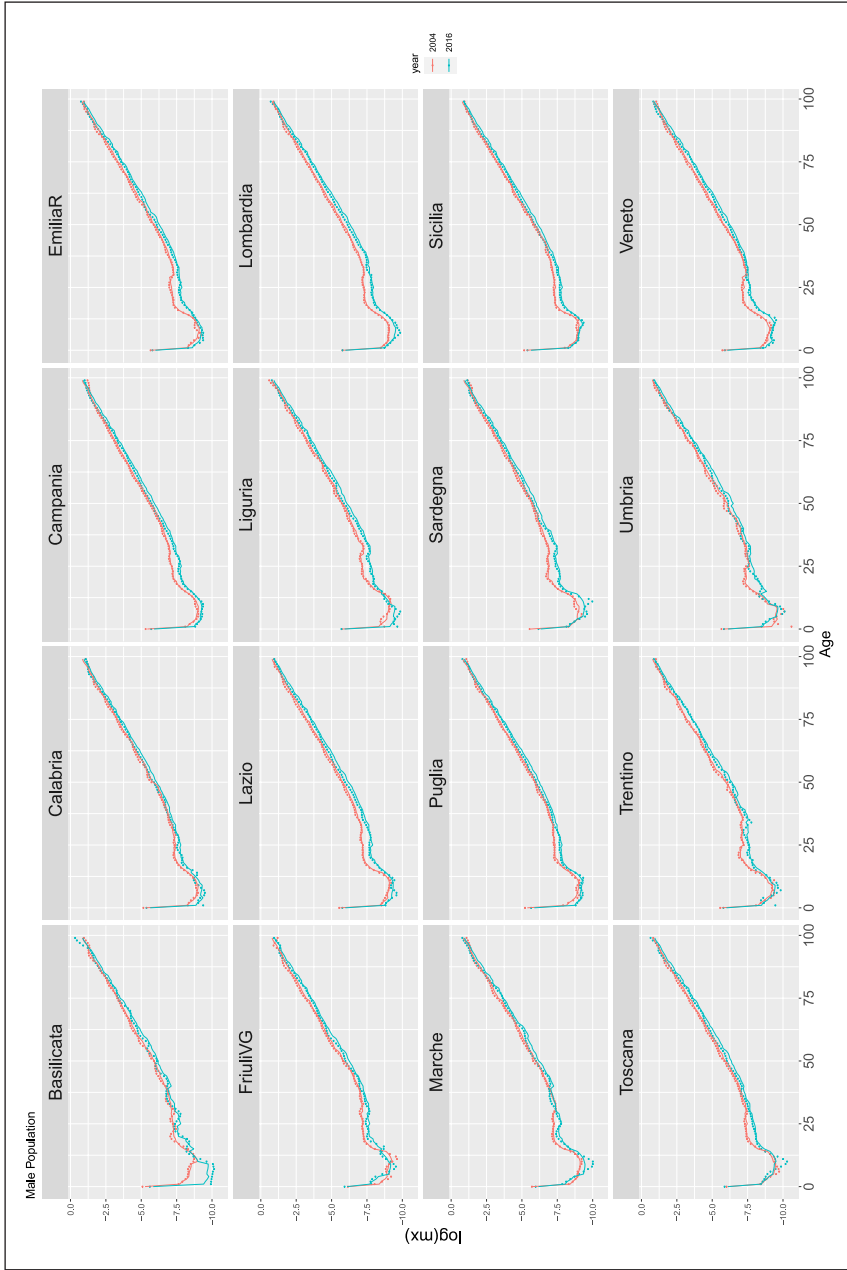
model, did not bring any benefit to the estimation. Also in this case, we provide in the Supplemental Material scatter plots  $(\log m_{x,t,r,g}, \log \hat{m}_{x,t,r,g})$  for all regions (Figures 3, and 4 in SM) to visualize the similarity between the actual and reconstructed death rates.

## 6. Conclusions

This paper contributes to the current literature on the demographic methods for ungrouping vital rates leveraging deep learning, which has provided reliable estimates in many fields of application. We propose a DNN model in a multi-population framework to ungroup death rates from rates gathered in five-year age-groups. This method represents an advance in mortality modeling as it may be used to estimate vital rates for single ages in regions or populations where present information is lacking. We investigate the ability of our model to provide reliable predictions of age-specific death rates by also studying how the hyperparameters' choice affects the model's reliability. We measure the accuracy of our method by analyzing the age-specific relative differences between the real and the estimated death rates. The results of the numerical experiments show the high accuracy of the proposed model, which captures the dynamics of mortality by age over time and between different populations. Indeed, mortality modeling is challenging due to



**Figure 4.** Estimated ungrouped death rates by Italian regions for years 2004 and 2016 based on the period 1974 to 2003 (training and validation). Dots refer to the observed rates. Solid red lines are the DNN reconstructed ungrouped death rates. Females. Color version of the figure is available online.



**Figure 5.** Estimated ungrouped death rates by Italian regions for years 2004 and 2016 based on training period 1974 to 2003 (training and validation). Dots refer to the observed rates. Solid red lines are the DNN reconstructed ungrouped death rates. Males. Color version of the figure is available online.

its dynamic; thus one of the main tasks is to grasp the heterogeneity of mortality in different regions.

We also contribute to the state of the art in indirect estimation by introducing a multi-population indirect estimation leveraging subnational data. The proposed model yields impressive results, even when using a lower-quality data source.

Through our model, we can offer a comprehensive picture of specific mortality levels, providing reliable results that might be exploited by public health and national systems, aiming to obtain granular information on mortality profiles for more accurate social planning.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: A.N. was supported by the MUR-PRIN 2022 project CARONTE (Prot. 2022KBTEBN), funded by the European Union - Next Generation EU.

## ORCID iD

Andrea Nigri  <https://orcid.org/0000-0002-2707-3678>

## Supplemental Material

Supplemental material for this article is available online.

## References

- Beers, H. S. 1945. "Modified-Interpolation Formulas That Minimize Fourth Differences." *Record of the American Institute of Actuaries* 34 (69): 14–20.
- Bengio, Y., R. Ducharme, and P. Vincent. 2000. "A Neural Probabilistic Language Model." *The Journal of Machine Learning Research* 3: 1137–1155.
- Blower, G., and J. E. Kelsall. 2002. "Nonlinear Kernel Density Estimation for Binned Data: Convergence in Entropy." *Bernoulli* 8 (4): 423–49. DOI: <https://www.jstor.org/stable/3318847>.
- Boneva, L. I., D. G. Kendall, and I. Stefanov. 1971. "Spline Transformations: Three New Diagnostic Aids for the Statistical Data-Analyst." *Journal of the Royal Statistical Society. Series B (Methodological)* 33 (1): 1–71. DOI: <https://www.jstor.org/stable/2986005>.
- Guo, C., and F. Berkhahn. 2016. "Entity Embeddings of Categorical Variables." *arXiv preprint arXiv:1604.06737*.
- Hainaut, D. 2018. "A Neural-Network Analyzer for Mortality Forecast." *Astin Bulletin* 48: 481–508. DOI: <https://doi.org/10.1017/asb.2017.45>.
- Hornik, K., M. Stinchcombe, and H. White. 1989. "Multilayer Feedforward Networks are Universal Approximators." *Neural Networks* 2 (5): 359–66.
- Hsieh, J. J. 1991. "Construction of Expanded Continuous Life Tables a Generalization of Abridged and Complete Life Tables." *Mathematical Biosciences* 103 (2): 287–302.
- Human Mortality Database. 2018. "University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)." Available at: <https://www.humanmortality.org>.
- Kostaki, A. 1991. "The Heligman-Pollard Formula as a Tool for Expanding an Abridged Life Table." *Journal of Official Statistics* 7 (3): 311–23.
- Kostaki, A., and V. Panousis. 2001. "Expanding an Abridged Life Table." *Demographic Research* 5 (1): 1–22.

- Levantesi, S., A. Nigri, and G. Piscopo. 2022. "Clustering-Based Simultaneous Forecasting of Life Expectancy Time Series Through Long-Short Term Memory Neural Networks." *International Journal of Approximate Reasoning* 140: 282–97.
- Liu, Y., P. Gerland, T. Spoorenberg, V. Kantorova, and K. Andreev. 2011. "Graduation Methods to Derive Age-Specific Fertility Rates From Abridged Data: A Comparison of 10 Methods Using HFD Data." Presented at the First HFD Symposium, MPIDR, Rostock, Germany.
- McNeil, D. R., T. J. Trussell, and J. C. Turner. 1977. "Spline Interpolation of Demographic Data." *Demography* 14 (2): 245–52.
- Mehta, N. K., L. R. Abrams, and M. Myrskylä. 2020. "US Life Expectancy Stalls Due to Cardiovascular Disease, Not Drug Deaths." *Proceedings of the National Academy of Sciences of the United States of America* 117 (13): 6998–7000.
- Nigri, A., J. M. Aburto, U. Basellini, and M. Bonetti. 2022a. "Evaluation of Age-Specific Causes of Death in the Context of the Italian Longevity Transition." *Scientific Reports* 12 (1): 22624.
- Nigri, A., E. Barbi, and S. Levantesi. 2022b. "The Relationship Between Longevity and Lifespan Variation." *Statistical Methods and Applications* 31 (3): 481–93. DOI: <https://doi.org/10.1007/s10260-021-00584-4>.
- Nigri, A., E. Barbi, and S. Levantesi. 2022c. "The Relay for Human Longevity: Country-Specific Contributions to the Increase of the Best-Practice Life Expectancy." *Quality & Quantity* 56 (6): 4061–73. DOI: <https://doi.org/10.1007/s11135-021-01298-1>.
- Nigri, A., S. Levantesi, and J. M. Aburto. 2022d. "Leveraging Deep Neural Networks to Estimate Age Specific Mortality From Life Expectancy at Birth." *Demographic Research* 47: 199–232. DOI: <https://doi.org/10.4054/DemRes.2022.47.8>.
- Nigri, A., S. Levantesi, and M. Marino. 2021. "Life Expectancy and Lifespan Disparity Forecasting: A Long Short-Term Memory Approach." *Scandinavian Actuarial Journal* 2021 (2): 110–33. DOI: <https://doi.org/10.1080/03461238.2020.1814855>.
- Nigri, A., S. Levantesi, M. Marino, S. Scognamiglio, and F. Perla. 2019. "A Deep Learning Integrated Lee-Carter Model." *Risks* 7 (1): 33. DOI: <https://doi.org/10.3390/risks7010033>.
- Perla, F., R. Richman, S. Scognamiglio, and M. V. Wüthrich. 2021. "Time-Series Forecasting of Mortality Rates Using Deep Learning." *Scandinavian Actuarial Journal* 7: 572–98.
- Richman, R. 2021. "Mind the Gap - Safely Incorporating Deep Learning Models Into the Actuarial Toolkit." Available at: <https://ssrn.com/abstract=3857693>.
- Richman, R., and M. V. Wüthrich. 2021. "A Neural Network Extension of the Lee-Carter Model to Multiple Populations." *Annals of Actuarial Science* 15 (2): 346–66.
- Rizzi, S., J. Gampe, and P. H. C. Eilers. 2015. "Efficient Estimation of Smooth Distributions From Coarsely Grouped Data." *American Journal of Epidemiology* 182 (2): 138–47.
- Rizzi, S., M. Thinggaard, G. Engholm, N. Christensen, T. B. Johannesen, J. W. Vaupel, and R. Lindahl-Jacobsen. 2016. "Comparison of Non-Parametric Methods for Ungrouping Coarsely Aggregated Data." *BMC Medical Research Methodology* 16: 59. DOI: <https://doi.org/10.1186/s12874-016-0157-8>.
- Schmertmann, C. 2012. "Calibrated Spline Estimation of Detailed Fertility Schedules From Abridged Data." MPIDR Working Paper WP-2012-022, Rostock, Germany.
- Scognamiglio, S. 2022. "Calibrating the Lee-Carter and the Poisson Lee-Carter Models via Neural Networks." *ASTIN Bulletin: The Journal of the IAA* 52 (2): 519–61.
- Smith, L., R. Hyndman, and S. Wood. 2004. "Spline Interpolation for Demographic Variables: The Monotonicity Problem." *Journal of Population Research* 21 (1): 95–7.

Received: May 4, 2023

Accepted: February 14, 2024