STUDIA DEMOGRAFICZNE

1–2(185–186) 2024

Original research paper

Łukasz Głąb
SGH Warsaw School of Economics
Institute of Statistics and Demography
Email: lglab@sgh.waw.pl
ORCID: 0009-0005-6150-571X

Wioletta Grzenda
SGH Warsaw School of Economics
Institute of Statistics and Demography
email: wgrzend@sgh.waw.pl
ORCID: 0000-0002-2226-4563

DOI: 10.33119/SD.2024.1-2.1

Mortality modelling and forecasting using generalised age-period-cohort models and neural networks

Abstract

Modelling and forecasting mortality risk are key tasks in demography as well as for social security institutions and insurance companies. Traditionally used stochastic mortality models such as the Lee-Carter model, require meeting assumptions which cannot always be met in real-life scenarios. These include, for example, the condition of time independence of age-specific improvement rates. An alternative approach to mortality modelling is based on deep neural networks. Previous works in the field primarily focus on recurrent neural networks, typically used in time series forecasting problems. This work aims to compare and analyse the effectiveness of both types of methods in mortality modelling and forecasting based on nine European populations. The study uses data from the Human Mortality Database.

Additionally, we propose a hyperparameter tuning framework for the feedforward neural network model used in the study.

Keywords: Human Mortality Database, mortality forecasting, neural networks, stochastic mortality models

Introduction

In recent years, the magnitude of shocks affecting mortality rates – both in Europe and globally – has been unprecedented. A shock like this was the COVID-19 pandemic, which caused mortality rates across various nations, cohorts and socio-economic groups to change. That affected not only multiple individuals but also social security institutions and insurance companies that traditionally use mortality models in their daily operations. Mortality modelling and forecasting became even more important. Accurate mortality modelling is fundamental for the financial stability of social security systems – over- or underestimation of mortality risk might lead to inadequate payouts of social benefits and affect public budgets and fiscal stability of countries. Inadequate estimation of mortality rates (and life expectancy) might lead to incentive distortions and miscalculations of financial sustainability (Ayuso et al., 2021). On average, across 31 OECD countries, public pension expenditure is expected to increase from 8.9% of GDP in 2020–2023 to 10.2% of GDP in 2050 (OECD, 2023). Thus, long-term mortality projections are crucial for responsible pension policy planning, as there is usually a lag between reform occurrence and its impact on social expenditures.

From the perspective of insurance and reinsurance companies, inaccurate mortality models, especially those related to long-term products (Gaille and Sherris, 2011), may have a direct impact on solvency, profitability and competitiveness. Underestimation of mortality risk could lead to insufficient technical provisions, which may lead to insolvency. On the other hand, overestimation of mortality risk may lead to inadequate product pricing, which can decrease the entity competitiveness.

A basic stochastic mortality model, utilising singular value decomposition (SVD) for decomposing the matrix of mortality rates with dimensions reflecting age and time, was introduced by Lee and Carter (1992). Since then, many extensions and alternatives of this model have been developed. There are stochastic mortality modelling methods related to single population models as well as multiple population models, both in discrete and continuous time.

As for the single population models, the most recognisable include works introducing a more formalised approach based on Poisson regression (Brouhns et al., 2002), incorporating the cohort effects into the model (Renshaw and Haberman, 2006) or assuming linearity of the one-year death probabilities logit for older ages within the two-factor Cairns-Blake-Dowd (CBD) model (Cairns et al., 2006). Furthermore, the characteristics of the classical Lee-Carter (LC) model and the CBD model were joined in the Plat model (Plat, 2009), allowing for capturing the cohort effect as well as estimation tailored to broader age ranges as compared to the CBD model. The developments by Hunt and Blake (2015) and Currie (2016) help to generalise the abovementioned models within an age-period-cohort structure and define such models as generalised linear models or non-linear models. The generalised age-period-cohort (GAPC) framework was also described by Villegas, Millossovich and Kaishev (2018).

As for the multiple population models, the basic Lee-Carter model was extended to a joint model of the mortality coefficients of multiple countries with a common mortality index (Li and Lee, 2005). Further extensions and modifications include the Bayesian (Antonio et al., 2015) or factor-based approach (Chen et al., 2015) as well as models taking into account a long-term convergence of the mortality rates of multiple populations (Li et al., 2017) or two-population discrete models with jumps (Zhou et al., 2013 and Özen and Şahin, 2021).

Advances in machine learning and deep learning have resulted in mortality models based on alternative approaches to stochastic mortality models. One of the initial works in the field explores the possibility of determining the causes of death in mortality modelling (Deprez et al., 2017) with the use of tree-based algorithms, which were then further extended by Levantesi and Pizzorusso (2018). A stacked regression ensemble approach is proposed by Kessy et al. (2022). As for the neural network-based approach, one of the first works in the field by Hainaut (2018) describes a two-step semiparametric model for estimation of log forces of mortality. The deep learning approach to mortality modelling was further explored by Nigri et al. (2019), who applied the long-short-term memory (LSTM) network to fit the time index trend being part of the traditional LC model. This approach turned out to outperform the classical model in forecasting capabilities. Richman and Wüthrich (2021) apply a neural network with embedding layers within a multi-population LC framework, which helps to incorporate categorical features into the model.

The most recent developments in the field include the applications of recurrent neural network (RNN) and convolutional neural network based model architectures to mortality forecasting, for example by Perla et al. (2021) and Wang et al. (2021) who combine neighbouring mortality framework with the CNN model to achieve better forecasting performance. Other recent developments include a locally coherent multi-population model by Perla and Scognamiglio (2022), applying neural networks to point and interval forecasts of mortality rates (Schnürch, Korn, 2022) or neural network-based calibrations of (Poisson) LC multi-population models (Scognamiglio, 2022). An alternative approach, based on transformer architecture which uses multi-head attention mechanism and positional encoding for key mortality related features extraction, was recently introduced by Wang et al. (2024), whereas multi-population mortality models handling long-term forecast divergence of mortality rates resulting from modelling multiple populations was simultaneously presented by Scognamiglio (2024).

The GAPC mortality models have a clear parametric structure which ensures better interpretability (i.e. specific model parameters related to age, period and cohort effects on mortality). Additionally, such models have well-defined statistical properties. They can be fitted using the maximum likelihood estimation method, which ensures consistency, equivariance and efficiency of the model parameter estimates (Fisher, 1922). GAPC models also allow for straightforward quantification of prediction uncertainty (e.g. using confidence intervals). Such models also utilise identifiability constraints to ensure unique parameter estimates and prevent overfitting. The estimation process itself is efficient and thus practically applicable in real-world scenarios (e.g. official mortality projections by social security institutions or estimations of longevity risk for Solvency II requirements). On the other hand, the majority of GAPC models assume (log) linear nature of mortality trends, which may not capture more complex dependencies between age, period and cohort effect, e.g. for smaller populations with non-trivial mortality patterns. They also usually treat each population independently, which does not allow for capturing common trends in mortality patterns across the analysed populations.

The main advantages of the neural network-based approach to mortality modelling are, in particular, the ability to identify complex dependencies between age, period and cohort effects as well as the possibility to capture common trends between different populations, notwithstanding their characteristics. On the other hand, such models usually lack interpretability and explainability (although there are first developments in that field, e.g. as in the work of Perla, Richman, Scognamiglio and Wüthrich (2024)). As for the estimation process, it is usually more complex and time-consuming as compared to GAPC models, especially if training involves extensive hyperparameter tuning.

The main aim of this paper is to compare the predictive effectiveness of various single mortality models from the GAPC framework against a feedforward neural network architecture based on a multiple population approach similar to the one proposed by Richman and Wüthrich (2021). Additionally, in order to improve the performance of the neural network, we conduct hyperparameter tuning focusing on a broader set of hyperparameters as compared to the original paper.

This study examines the following research questions:

1. How does the predictive effectiveness of mortality models from the GAPC framework compare to a feedforward neural network architecture based on a multiple population approach, when applied to selected European populations?
2. Does hyperparameter tuning focusing on a broader set of parameters combined with ensembling techniques improve the predictive performance of the feedforward neural network model for mortality modelling?

The analysis was conducted based on the data from selected European countries available in the Human Mortality Database (HMD).[1]

Data

In our analysis, we consider mortality data for nine countries (Germany, Spain, France, Hungary, Italy, Lithuania, Latvia, Poland and Portugal) from HMD to account for the heterogeneity of mortality patterns among various populations for periods $T = \{1960, \ldots, 2018\}$ with mortality rates for ages $A = \{0, 1, \ldots, 99\}$, split by gender. We then split the data into two sets of calendar years, $T_{train} = \{1960, \ldots, 2007\}$ and $T_{test} = \{2008, \ldots, 2018\}$ to train and test the model performance respectively.

Additional data transformations we applied are related to unifying the datasets for East and West Germany (DEUTE and DEUTW respectively) as the combined DEUTNP dataset contained only the data from 1990 onwards. The calculation of age specific mortality rates for females and males from Germany was done by aggregating exposures and total number of deaths per group (male, female), age and year and calculating mortality rates based on such aggregated measures (for East and West Germany combined). In order to keep the stability of the neural network the maximum mortality rate is set to 1, whereas the minimum mortality rate was set to 0.000001 in cases where the number of deaths was equal to 0 to prevent errors in the logarithmic transformation of the mortality rates. For fitting models following the GAPC framework we utilise the *StMoMo* R package (Villegas, Millossovich and Kaishev, 2018), whereas for the feedforward neural network model we utilise the *TensorFlow* library (Abadi et al., 2016) through *Keras* API for R (Chollet, 2015).

[1] Human Mortality Database, https://mortality.org/

General notation for GAPC mortality models

Let us denote calendar year as t, random variable defining number of deaths of a given age x at time t as $D(x,t)$ and the observed number of deaths as $d(x,t)$ where the related central number of people of a given age exposed to risk is denoted as $E^c(x,t)$. Then, we can arrange these data into matrices respectively $D(d(x,t))$ and $E^c = E^c(x,t)$ having dimensions $n_a \times n_y$ (with n_a ages and n_y years and $n_b = n_a + n_y - 1$ cohorts. Then the force of mortality matrix can be defined as $\boldsymbol{m} = \left(\frac{d(x,t)}{E^c(x,t)}\right)$.

As per generalisation by Villegas, Millossovich and Kaishev (2018), we can then define the GAPC framework for single population mortality modelling as follows:

$$\ln m(x,t) = a(x) + \sum_{i=1}^{N} f^i(x)\kappa^i(t) + \gamma(c) \tag{1}$$

where $m(x,t)$ is the force of mortality for a given age $x \in [x_1, x_{n_a}]$ as of time period $t \in [t_1, t_{n_y}]$ and for cohort $c \in \{t_1 - x_{n_a}, \ldots, t_{n_y} - x_1]$, $a(x)$ is the age effect of the mortality pattern averaged out across the time periods, N is the number of age – period terms describing the temporal effect $\kappa^i(t)$, $f^i(x)$ is the age modulating function, whereas $\gamma(c)$ is a factor reflecting cohort effect (where $c = t - x$). Formula (1) denotes the systematic component of the model. Additionally the model contains a random component $D(x,t)$ having a Poisson distribution (Brouhns et al., 2002), i.e.

$$D(x,t) \sim Poiss(E^c(x,t) * m(x,t)). \tag{2}$$

The relation between systematic and random components is given by a link function g:

$$g\left(\mathbb{E}\left(\frac{D(x,t)}{E^c(x,t)}\right)\right) = \ln m(x,t) \tag{3}$$

which in case of the Poisson distribution is usually the log link function. Alternatively one could select a binomial distribution with logit link function.

The majority of stochastic mortality models are restricted with parameter constraints related to the abovementioned age, period and cohort effects (i.e. are identifiable to a certain transformation ensuring unique parameter estimates). These are applied using a constraint function mapping the initial vector of parameters into a vector of transformed parameters where the latter one satisfies the model constraints without effect on the log force of mortality $\ln\ m(x,t)$. The most popular single population mortality models based on the GAPC framework are listed below, together with associated model constraints. Lee Carter model (1992) is the simplest one and is defined as follows:

$$\ln m(x,t) = a(x) + \beta^{(1)}(x)\kappa^{(1)}(t), \tag{4}$$

with the following parameter constraints: $\sum_t \kappa^{(1)}(t) = 0, \sum_x \beta^{(1)}(x) = 1$. It is worth noting that this model is lacking the cohort effect, thus it cannot capture the non-linear tendencies in mortality rates well.

Such a component is incorporated in the Renshaw and Haberman (2006) (RH) model:

$$\ln m(x,t) = a(x) + \beta^{(1)}(x)\kappa^{(1)}(t) + \beta^{(0)}(x)\gamma(c), \tag{5}$$

with following parameter constraints:

$$\sum_t \kappa^{(1)}(t) = \sum_x \gamma^{(1)}(c) = 0, \sum_x \beta^{(1)}(x) = \sum_x \beta^{(0)}(x) = 1.$$

Setting $\beta^{(0)}$ to 1 simplifies the abovementioned structure to modified Renshaw Haberman (mRH) model (2011), which is more stable than the original one, thus it would be considered in the analysis part.

If the parameters $\beta^{(0)}$ and $\beta^{(1)}$of RH model are set to 1, then we obtain the so called age-period-cohort (APC) model (Cairns et al., 2009):

$$\ln m(x,t) = a(x) + \kappa^{(1)}(t) + \gamma(c), \tag{6}$$

with the following constraints $\sum_t \kappa^{(1)}(t) = \sum_x \gamma^{(1)}(c) = \sum_x c\gamma^{(1)}(c) = 0$.

The model tends to be more efficient in the case of dataset structure alterations.

The next model by Cairns Blake and Dowd (2006) assumes that age is linearly interacting with log mortality and there is no cohort effect:

$$\ln\ m(x,t) = \kappa^{(1)}(t) + (x - \bar{x}) - \bar{x})\kappa^{(2)}(t). \tag{7}$$

Two extensions of the CBD model allow for more robust mortality patterns modelling. Cairns et al. (2009) introduce a more complex M7 model which allows for capturing additional age-related dynamics by introducing a quadratic component:

$$\ln m(x,t) = \kappa^{(1)}(t) + (x-\bar{x})\kappa^{(2)}(t) + \left((x-\bar{x})^2 - \hat{\sigma}_x^2\right)\kappa^{(3)}(t) + \gamma(c), \qquad (8)$$

given the following constraints:

$$\sum_t \kappa^{(1)}(t) = \sum\nolimits_t \kappa^{(2)}(t) = \sum\nolimits_t \kappa^{(3)}(t) = \sum_x \gamma^{(1)}(c) = \sum_x c^2 \gamma^{(1)}(c) = 0,$$
$$\sum_x \beta^{(1)}(x) = 1,$$

where $\hat{\sigma}_x^2$ is the average value of $(x-\bar{x})^2$.

Another extension of the CBD model is proposed by Plat (2009) with additional parameters related to age and period effects:

$$\ln m(x,t) = a(x) + \kappa^{(1)}(t) + (x-\bar{x})\kappa^{(2)}(t) + \left((x-\bar{x})^+\right)\kappa^{(3)}(t) + \gamma(c), \qquad (9)$$

with analogical constraints as the M7 model.

All of the abovementioned models can be fitted with the maximum likelihood function as specified by Villegas, Millossovich and Kaishev (2018):

$$\mathcal{L}\left(d(x,t), \hat{d}(x,t)\right) = \sum_x \sum_t \omega(x,t)\left(d(x,t)\ln \hat{d}(x,t) - \hat{d}(x,t) - \ln d(x,t)!\right) \qquad (10)$$

where the weights $\omega(x,t)$ equal to 1 if a given data point (x,t) is included in the model and 0 otherwise. Given the Poisson distribution of number of deaths, the expected number of deaths is given by the following formula:

$$\hat{d}(x,t) = E^c(x,t)e^{\left(a(x) + \sum_{i=1}^{N} f^i(x)\kappa^i(t) + \gamma(c)\right)}. \qquad (12)$$

Multiple population modelling using neural networks

While designing the neural network model for mortality modelling purposes, we follow a setup defined by Richman and Wüthrich (2021). The model is fit to mortality rates for the calendar years 1960–2007, and then the prediction of mortality rates is performed on the years 2008–2018. The feature space consists of age, calendar year, region and gender, where the calendar year (or year of death) is treated as numerical input to the neural network to allow for forecasting future mortality rates beyond the calendar years on which the model was fitted. 10% of the data from the training set is used as a validation set (obtained through stratified sampling with regard to age and calendar year) in order to select the best combination of hyperparameters minimising the mean squared error (MSE). We treat region, gender and age as categorical features transformed through embedding layers (Bengio et al., 2003), which transform values of categorical features into a lower-dimensional vector whose parameters are learned during model training. We can define the embedding layer $\imath_{\rho}$ for $\rho = \{\rho_1, \dots, \rho_n\}$ being number of unique categories of cardinality n. Then such a layer transforms $\rho_j \in \rho$ into a q_{ρ} – dimensional vector as defined below:

$$\varepsilon_{\rho}(\rho_j) = \left(\varepsilon_{\rho,1}(\rho_j), \dots, \varepsilon_{\rho,q_{\rho}}(\rho_j)\right)'. \tag{13}$$

For the purposes of multiple population modelling based on neural networks, we can define a single feature vector being a concatenation of previously defined embedding vectors as

$$\boldsymbol{f}(t, x, i, j) = \left(t, \boldsymbol{a}(x)', \boldsymbol{r}(i)', \boldsymbol{s}(j)'\right)', \tag{14}$$

which serves as an input layer for the neural network model used for predicting (log) mortality as of calendar year t for age x, region i and gender j. Then we can define the structure of the neural network as comprising of intermediate layers $\boldsymbol{Z}^{(1)}, \dots, \boldsymbol{Z}^{(h)}$ with θ_h neurons and the final output layer $\hat{y}$ (being prediction of log mortality rates) as follows:

$$\boldsymbol{Z}^{1} \equiv \boldsymbol{F}(\boldsymbol{f}(t,x,i,j)) = \varphi_0\left(\boldsymbol{\omega}_0^{(0)} + \boldsymbol{\omega}^{(0)}\boldsymbol{f}(t,x,i,j)\right), \tag{15}$$

$$\boldsymbol{Z}^{(h)} \equiv \boldsymbol{F}(\boldsymbol{Z}^{(h-1)}) = \varphi_{h-1}\left(\boldsymbol{\omega}_0^{(h-1)} + \boldsymbol{\omega}^{(h-1)}\boldsymbol{Z}^{(h-1)}\right),\ h = 2,\ \dots,\ H-1 \tag{16}$$

$$\hat{\boldsymbol{y}} \equiv \boldsymbol{F}(\boldsymbol{Z}^{(H-1)}) = \varphi_H\left(\boldsymbol{\omega}_0^{(H)} + \boldsymbol{\omega}^{(H)}\boldsymbol{Z}^{(H-1)}\right), \tag{17}$$

where $\boldsymbol{\omega}_0^{(j)}$ is a θ_h-element intercept vector, $\boldsymbol{\omega}^{(j)}$ is a ($\theta_j \times \theta_{j-1}$)-dimensional weight matrix and φ_{h-1} is the activation function which calculates the impact of each neuron (these may vary between the layers).

Following the approach from Richman and Wüthrich (2021), we set the dimension of all embedding layers within the input feature vector to 5. The number of intermediate layers in the network is subject to hyperparameter tuning. For each of the intermediate layers, we also add dropout layers for regularisation purposes (Hinton et al., 2012), (where dropout probability for each neuron is subject to hyperparameter tuning) as well as batch normalisation layers (Ioffe, Szegedy, 2015) which speed up and stabilise the training process through recentring and rescaling of inputs to network layers. Similarly to Richman and Wüthrich (2021), we use the back-propagation algorithm to fit the model. The activation function for intermediate layers is set to Rectified Linear Unit (ReLU) (Nair, Hinton, 2010), whereas the last layer has the sigmoid activation function. Additionally, following the original work by Richman and Wüthrich (2021), we introduce skip connections between the feature layer and the last hidden layer as well as the feature layer and the first hidden layer to diminish the risk of the vanishing gradient problem.

As for the key hyperparameters utilised in the training process, we focus on both the ones related to the general network architecture as well as the ones responsible for training process optimization. The hyperparameters related to first area are the number of layers, number of neurons in each layer and batch size. As for the training process, we utilise dropout probability, learning rate and patience related to early stopping or learning rate reduction (i.e. a hyperparameter indicating the number of epochs with no improvement in MSE value on the validation set after which training will be stopped or learning rate will be reduced in a given iteration). We select a 5% subsample of all combinations within the hyperparameter space to speed up the training process.

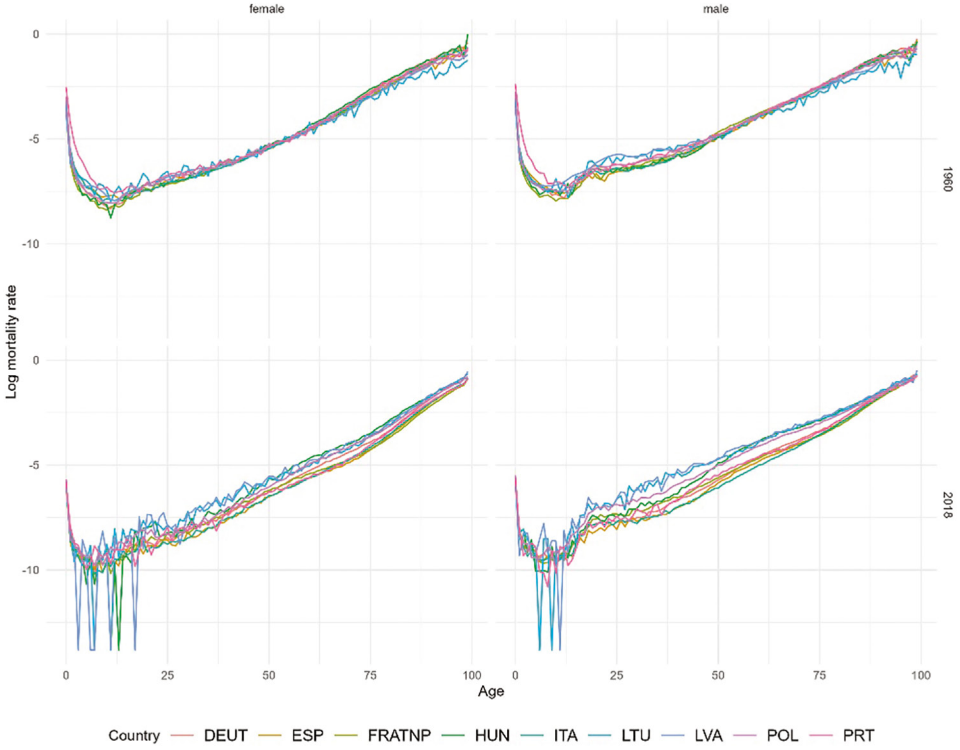
To increase the robustness of the model, we average out the predictions of the 4 best models obtained during the hyperparameter tuning process. Such an approach can help obtain more stable results on the test set and ensure better generalisation on the data not used during the training process.

Results

Exemplary mortality patterns for the ages and countries mentioned, compared as of the earliest (1960) year and latest (2018) year in the dataset, are presented in Figure 1.

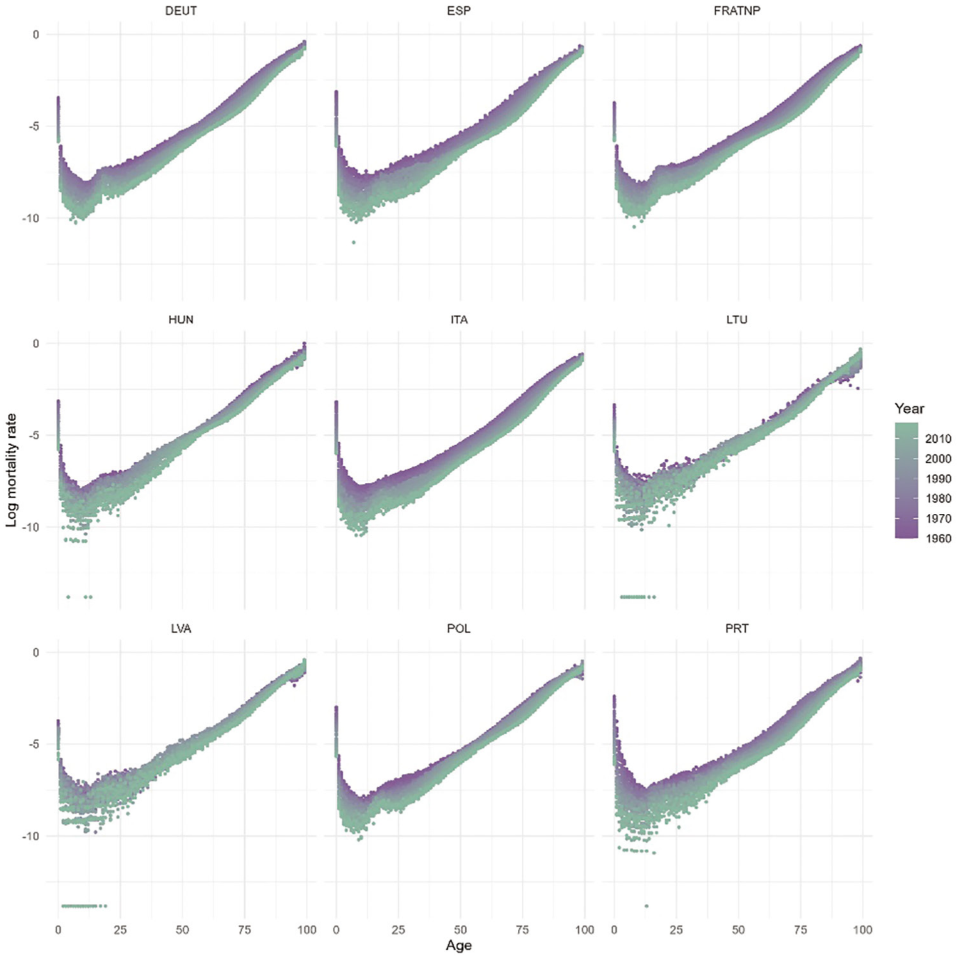
Looking at the high-level visualisations, it is easy to infer that mortality rates were lower in 2018 compared to 1960, they are generally lower for females than for males, and each country has a specific mortality pattern. Even though mortality decreased in European countries as compared to past century, one can still observe noticeable differences in the levels of mortality between Eastern and Western European countries, resulting not only from differences in healthcare or social security systems but also on supra-national developments between these two clusters of countries (Carracedo et al., 2018). As for the year 2018, interestingly, in some Baltic states (i.e. Lithuania and Latvia) as well as in Hungary, significantly lower mortality rates can be observed than in other countries analysed in specific younger age groups, indicating very low or even zero mortality rates for these. Common methods for overcoming such issues include introducing age-standardised rates (ASR) or smoothing techniques like those indicated by Perez-Panades et al. (2020). Those cases are more frequent for lower age groups in the latest years, which can be observed respectively for females and males in Figure 2 and Figure 3 (mortality patterns for the latest years are marked green).

Figure 1. Actual mortality rates (log scale) by gender for ages 0–99 from selected countries – comparison for years 1960 and 2018.



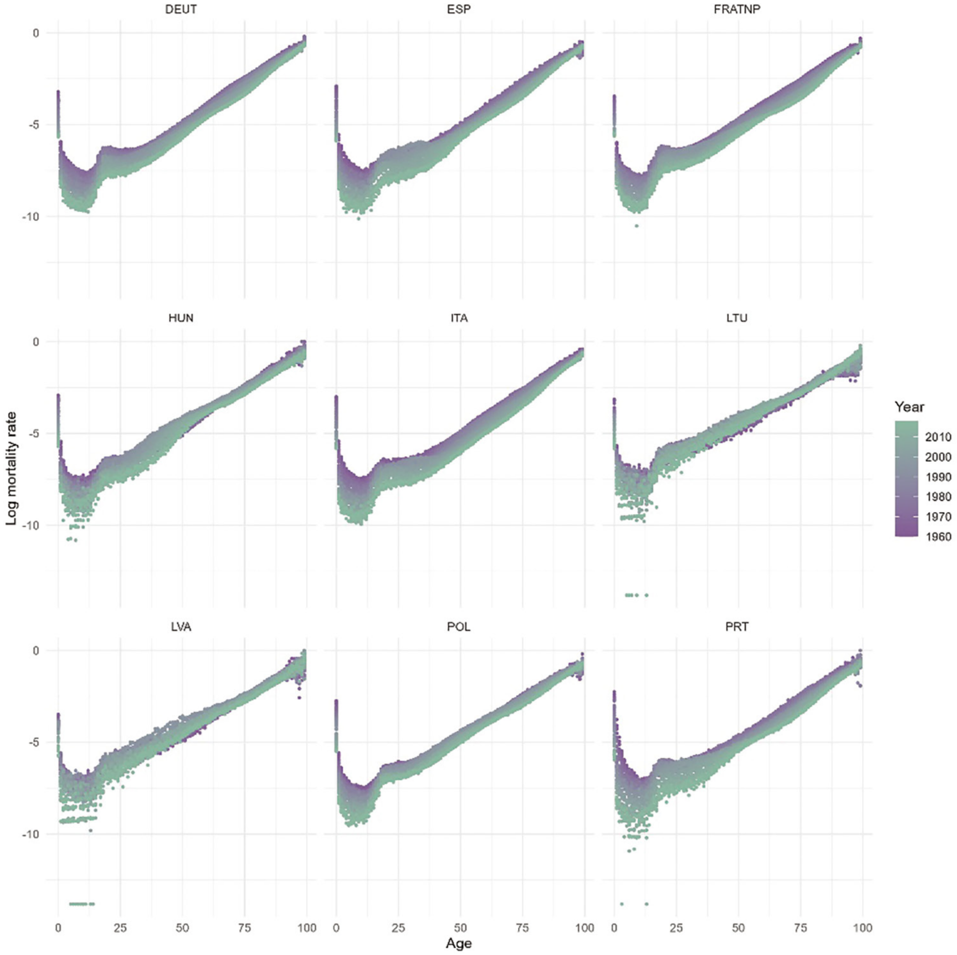
Source: Authors' own compilation based on per gender HMD mortality tables for ages 0–99 from selected countries for years 1960–2018.

Figure 2. Mortality rates (log scale) for females for ages 0–99 from selected countries for years 1960–2018



Source: Authors' own compilation based on per gender HMD mortality tables for ages 0–99 from selected countries for years 1960–2018.

Figure 3. Mortality rates (log scale) for males for ages 0–99 from selected countries for years 1960–2018



Source: Authors' own compilation based on per gender HMD mortality tables for ages 0–99 from selected countries for years 1960–2018.

The baseline neural network-based model hyperparameters and results on the training set are presented in Table 1. The four best models obtained in the hyperparameter tuning process are shown in Table 2. The hyperparameter tuning process results in a decrease in MSE values. The ensemble of the 4 best models will be used for prediction on the test set.

Table 1. Results of the baseline neural network model with key hyperparameter values on the training set (years 1960–2007)

	MSE val	MSE train	Learning rate (LR)	Hidden layers	Dropout	Neurons	Starting LR	Patience (early stopping)	Patience (LR reduction)	Batch size	Epochs	Epochs ended
0	0.0208	0.069	0.0000001	6	0.05	128	0.1	45	20	400	250	155

Source: Authors' own compilation based on per gender HMD mortality tables for ages 0–99 from selected countries for years 1960–2007.

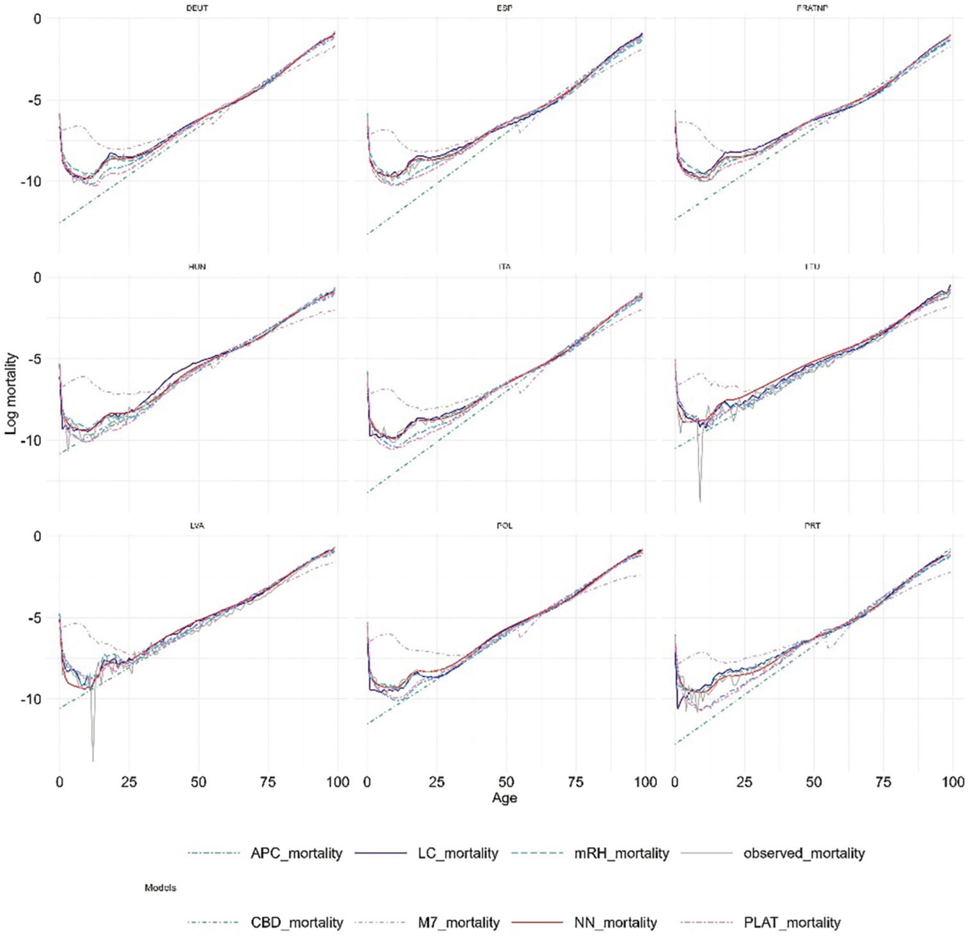
Table 2. Results of the 4 best neural network models with key hyperparameter values on the training set (years 1960–2007)

	MSE val	MSE train	Learning rate (LR)	Hidden layers	Dropout	Neurons	Starting LR	Patience (early stopping)	Patience (LR reduction)	Batch size	Epochs	Epochs ended
1	0.0150	0.0311	0.00000001	9	0.01	192	0.1	45	20	1200	250	250
2	0.0151	0.0379	0.00000001	6	0.01	224	0.05	35	30	800	250	250
3	0.0153	0.0444	0.00000001	9	0.03	160	0.1	35	30	400	250	219
4	0.0154	0.0391	0.00000000001	9	0.01	256	0.05	45	20	400	250	240

Source: Authors' own compilation based on per gender HMD mortality tables for ages 0–99 from selected countries for years 1960–2007.

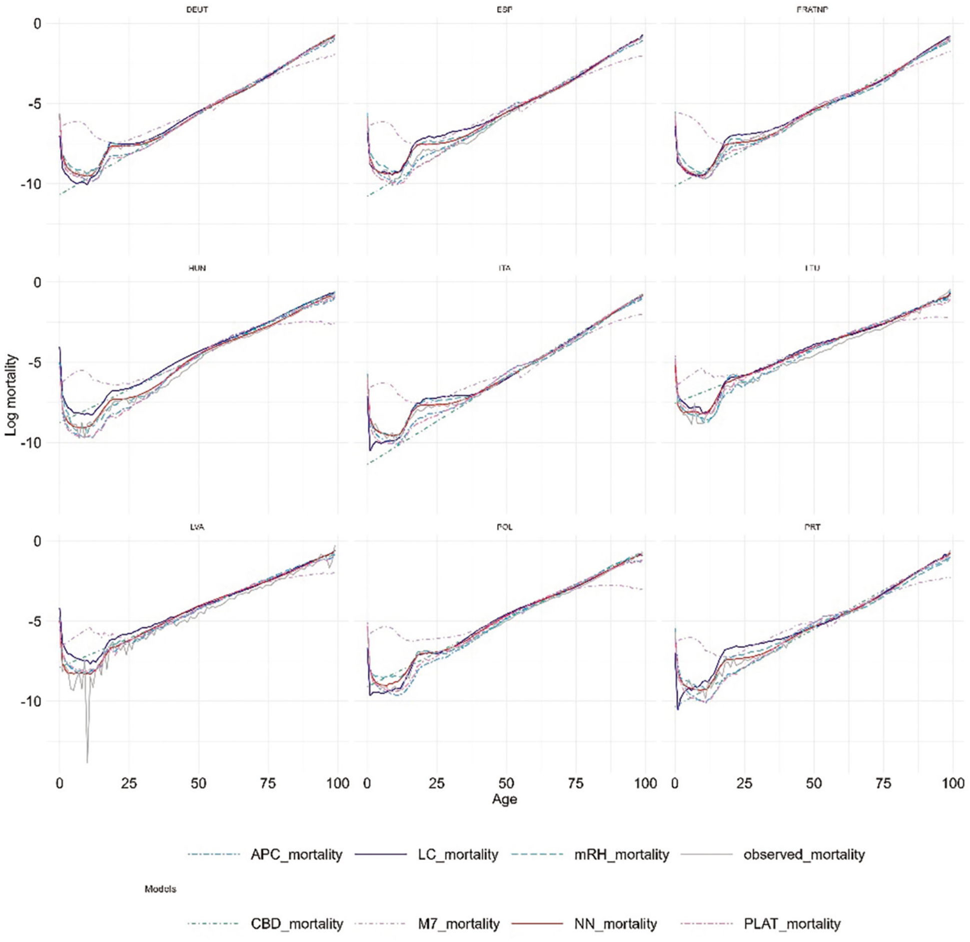
The results of GAPC and neural network-based ensemble models are illustrated graphically for the exemplary year (2014) against the observed mortality rates for females and males in Figures 4 and 5, respectively.

Figure 4. Comparison of predictions for ages 0–99 against observed mortality rates (log scale) per country for females as of the year 2014 from the test set



Source: Authors' own compilation based on per gender HMD mortality tables for ages 0–99 from selected countries for years 1960–2018.

Figure 5. Comparison of predictions for ages 0–99 against observed mortality rates (log scale) per country for males as of the year 2014 from the test set



Source: Autor’s own compilation based on per gender HMD mortality tables for ages 0–99 from selected countries for years 1960–2018.

Mean squared errors for each model per country and gender group on the test set are presented in Tables 3 and 4.

Table 3. Total MSE of mortality rates for the test set (years 2008–2018)

NN base	NNtuned	LC	CBD	mRH	PLAT	APC	M7
0.000261	0.000175	0.000221	0.000331	0.000405	0.000501	0.000621	0.004260

Source: Authors’ own compilation based on per gender HMD mortality tables for ages 0–99 from selected countries for years 1960–2018.

Table 4. Mean squared errors of mortality rates for test set per country gender groups (years 2008–2018)

Country	Gender	NN base	NN tuned	LC	CBD	mRH	APC	M7	PLAT	Min MSE
DEUT	female	0.000283	0.0001558	**0.0000793**	0.0001128	0.0001520	0.0005235	0.0025421	0.0002030	LC
DEUT	male	0.000233	0.0000748	**0.0000591**	0.0001991	0.0001261	0.0009113	0.0050821	0.0004344	LC
ESP	female	0.000142	0.0000387	**0.0000193**	0.0000372	0.0005183	0.0002321	0.0018829	0.0000577	LC
ESP	male	0.000162	**0.0000328**	0.0000355	0.0000605	0.0000601	0.0004364	0.0036437	0.0000776	NN
FRATNP	female	0.000138	0.0000375	**0.0000147**	0.0002913	0.0004528	0.0003332	0.0012783	0.0001915	LC
FRATNP	male	0.000137	0.0000465	**0.0000345**	0.0006689	0.0003109	0.0006058	0.0029403	0.0002782	LC
HUN	female	0.0000859	**0.0000734**	0.0000754	0.0000961	0.0001295	0.0003548	0.0040376	0.0001146	NN
HUN	male	**0.000135**	0.0001429	0.0006844	0.0003265	0.0006025	0.0008207	0.0077306	0.0005451	NN
ITA	female	0.000189	0.0001161	**0.0000323**	0.0000324	0.0005925	0.0003307	0.0022790	0.0000938	LC
ITA	male	0.000281	0.0000827	0.0000453	**0.0000416**	0.0002508	0.0004890	0.0041765	0.0001217	CBD
LTU	female	0.000865	**0.0004290**	0.0005442	0.0008427	0.0004316	0.0014187	0.0047923	0.0012673	NN
LTU	male	0.000663	**0.0006212**	0.0007650	0.0010937	0.0007129	0.0013373	0.0078459	0.0024125	NN
LVA	female	0.000241	**0.0002343**	0.0003238	0.0002514	0.0004120	0.0004736	0.0028595	0.0003190	NN
LVA	male	**0.000836**	0.0008975	0.0009866	0.0010115	0.0012216	0.0013821	0.0068007	0,0016099	NN
POL	female	0.000055	**0.0000300**	0.0000636	0.0001892	0.0000451	0.0002420	0.0039039	0.0001790	NN
POL	male	**0.000085**	0.0000874	0.0001040	0.0003299	0.0002810	0.0006934	0.0075708	0.0009420	NN
PRT	female	0.000021	**0.0000176**	0.0000247	0.0002112	0.0003095	0.0002222	0.0026182	0.0000377	NN
PRT	male	0.000149	**0.0000383**	0.0000951	0.0001649	0.0006894	0.0003662	0.0046890	0.0001277	NN

Source: Authors' own compilation based on per gender HMD mortality tables for ages 0–99 from selected countries for years 1960–2018.

The analysis of the results by country and gender groups indicates that in the majority of cases, the neural network models perform better than the other models (11 country gender groups out of 18). The traditional Poisson-based Lee-Carter model is the second most frequent model with minimal MSE (6 cases out of 18). As for the other GAPC-based models, the CBD model has performed the best in a single country gender group (Males from Italy). The tuned neural network-based model has also outperformed the base neural network model in 8 out of 11 country gender groups, as well as regarding total MSE on the test set.

Discussion and conclusions

This article aims to compare the effectiveness of traditional models and neural networks in mortality modelling and forecasting. Traditionally used stochastic mortality models have certain limitations, the main one being difficulties in fitting these models to the data (Richman and Wüthrich, 2021). The development of artificial intelligence methods has made machine learning models an effective tool in mortality modelling as well (Deprez et al., 2017; Hainaut, 2018; Levantesi and Pizzorusso, 2018; Kessy et al., 2022).

The results presented in this paper, based on the data from nine European countries, show that neural network-based models might have better performance than the traditional GAPC models.

Referring to the first research question, our results show that the neural network-based models achieve better results in the majority of analysed country gender groups (11 out of 18). The Poisson-based Lee-Carter model is the second-best-performing model (6 cases out of 18).

GAPC models assume (log) linear nature of mortality trends, which may not capture more complex dependencies between age, period and cohort effect, e.g. for smaller populations with non-trivial mortality patterns. They also usually treat each population independently, which does not allow for capturing common trends in mortality patterns across the analysed populations. On the other hand, such complex dependencies between age, period and cohort effects as well as the possibility to capture common trends between different populations, notwithstanding their characteristics, is possible using the neural network approach. The architecture of such a model is elastic and can be tailored to analysed populations during the hyperparameter tuning process. However, it is worth emphasising that the graphs of logarithms of mortality rates by gender that we obtained, showing that the decline in mortality became clearly observable after 1990, are consistent with the probability of dying between the ages of 15 and 60 based on the same database, i.e. HMD (World Bank Group, 2024).

We can also conclude that the neural network based models achieve better results for countries that might be perceived as emerging or developing markets (e.g. Baltic states, Hungary) whereas the Lee-Carter model performs better mostly for the countries perceived as developed economies (in particular Germany, France and Italy), e.g. as per classification by Morgan Stanley Capital International (MSCI) (MSCI, 2024).[2] As for the total mean squared errors measured on the test set (i.e. not differentiating between country gender groups), the neural network and Lee Carter models are also the best performing models. As per analysis by Shen et al. (2024), some of the abovementioned countries fall into the same clusters based on mortality pattern similarities.

As for the second research question, we have found that extended hyperparameter tuning and ensembling of the best neural network architectures result in better predictions on the test set than the baseline model without applying these techniques. It is also worth noting that the baseline neural network model still performed better as compared to models based on the GAPC framework for selected countries with smaller populations (e.g. Hungary, Lithuania or Latvia).

The model always would need to be selected on a case–by–case basis, and, in the case of more complex ones (neural network-based), a thorough hyperparameter selection could contribute to the robustness of the model and increased prediction stability.

Another area worth investigating while analysing GAPC and neural network-based mortality models is the impact of mortality shocks (e.g. COVID-19 pandemic) on the model behaviour. As for GAPC models, pandemic shocks could negatively affect all of the mortality pattern-related effects, i.e. the temporal effect (shift in gradual decrease in mortality over time), age group effect (especially for the advanced age groups), as well as the cohort effect. Simple extrapolation of mortality patterns observed for given age intervals using the traditional ARIMA approach would not necessarily reflect the impact of pandemic shocks on future mortality trends accurately (Ashofteh, Bravo, 2021). Even though such models explicitly incorporate the age effect, the future trends may be misestimated if pandemic shock disproportionately affects specific age groups. Such unusual mortality pattern shifts could also increase the forecast uncertainty (i.e. result in wider confidence intervals of the post-pandemic forecast). The neural network-based models, on the other hand, are expected to be

[2] Morgan Stanley Capital International, https://www.msci.com/our-solutions/indexes/developed-markets, https://www.msci.com/our-solutions/indexes/emerging-markets.

more flexible in handling mortality shocks but could also overfit to the training data without proper hyperparameter tuning.

As for future developments taking into account mortality shocks, one might consider comparative analysis with the use of combined neural networks of various types (recurrent, convolutional, transformer-architecture based, generative adversarial networks, etc.) as well as model ensembling techniques between traditional and neural network based architectures. Ensembling allows for increased forecasting accuracy by integrating several predictions from underlying models, making the model more robust and less prone to overfitting (Breiman, 1996). One could also consider the incorporation of the Bayesian approach to account for uncertainties related to mortality shocks and external factors related to irregularities in mortality patterns, such as emigration and immigration (Wiśniowski et al., 2015). Such an approach could increase the model robustness and allow for a better generalisation about future trends, also accounting for post pandemic factors affecting mortality patterns in various populations.

References

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y. Zhang, X. (2016). TensorFlow: a system for large-scale machine learning, *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, 265–283.

[2] Antonio, K., Bardoutsos, A., Ouburg, W. (2015). Bayesian Poisson log-bilinear models for mortality projections with multiple populations, *European Actuarial Journal*, 5(2), 245–281.

[3] Ashofteh A., Bravo, J.M. (2021). Life Table Forecasting in COVID-19 Times: An Ensemble Learning Approach, *16th Iberian Conference on Information Systems and Technologies (CISTI)*, Chaves, Portugal, 1–6.

[4] Ayuso, M., Bravo, J.M., Holzmann, R. (2021). Getting life expectancy estimates right for pension policy: period versus cohort approach. *Journal of Pension Economics and Finance*, 20(2), 212–231.

[5] Bengio, Y., Ducharme, R., Vincent, P. Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3(2), 1137–1155.

[6] Breiman, L. (1996). Stacked regressions. *Machine learning*, 24, 49–64.

[7] Brouhns, N., Denuit M., Vermunt, J. (2002). A Poisson Log-Bilinear regression approach to the construction of projected life tables, *Insurance: Mathematics and Economics*, 31(3), 373–393.

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

[9] Cairns, A.J., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A., Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from England and Wales and the United States, *North American Actuarial Journal*, 3(1), 1–35.

[10] Carracedo, P., Debón, A., Iftimi, A., Montes, F. (2018). Detecting spatio-temporal mortality clusters of European countries by sex and age. *International journal for equity in health*, 17, 1–19.

[11] Chen, H., MacMinn, R., Sun, T., (2015). Multi-population mortality models: a factor copula approach, *Insurance: Mathematics and Economics,* 63, 135–146.

[12] Chollet, F. (2015). Keras: the Python deep learning library.

[13] Deprez, P., Shevchenko, P., Wüthrich, M. (2017). Machine learning techniques for mortality modeling, *European Actuarial Journal*, 7(2), 337–352.

[14] Fisher, R.A. (1922). On the Mathematical Foundations of Theoretical Statistics", *Philosophical Transactions of the Royal Society of London. Series A (222)*, 309–368.

[15] Gaille, S., Sherris, M. (2011). Modelling mortality with common stochastic long-run trends. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 36, 595–621.

[16] Hainaut, D. (2018). A neural-network analyzer for mortality forecast, *Astin Bulletin*, 48(2), 481–508.

[17] Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I. Salakhutdinov, R. (2012). *Improving neural networks by preventing co-adaptation of feature detectors*. arXiv, arXiv:1207.0580.

[18] Ioffe, S. Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning*, PMLR 37, 448–456.

[19] Kessy, S.R., Sherris, M., Villegas, A.M., Ziveyi, J. (2022). Mortality forecasting using stacked regression ensembles, *Scandinavian Actuarial Journal*, 2022(7), 591–626.

[20] Lee, R.D., Carter L.R. (1992). Modeling and Forecasting U.S. Mortality, *Journal of the American Statistical Association*, 87(419), 659–671.

[21] Levantesi, S., Pizzorusso, V. (2019). Application of Machine Learning to Mortality Modeling and Forecasting, *Risks*, 7(1), 26.

[22] Li, J.S.-H., Chan, W.-S., Zhou, R. (2017). Semicoherent multipopulation mortality modeling: the impact on longevity risk securitization, *Journal of Risk and Insurance*, 84(3), 1025–1065.

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

[24] Nair, V. Hinton, G. (2010). Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning*, 807–814.

[25] Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., Perla, F. (2019). A Deep Learning Integrated Lee–Carter Model, *Risks*, 7(33), 1–16.

[26] OECD (2023), *Pensions at a Glance 2023: OECD and G20 Indicators*, OECD Publishing, 214.

[27] Özen, S., Şahin, Ş. (2021). A two-population mortality model to assess longevity basis risk, *Risks*, 9(2), 44.

[28] Perez-Panades, J., Botella-Rocamora, P., Martinez-Beneito, M.A. (2020). Beyond standardized mortality ratios; some uses of smoothed age-specific mortality rates on small areas studies. *International Journal of Health Geographics*, 19, 1–14.

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

[30] Perla, F., Scognamiglio, S. (2022), Locally-coherent multi-population mortality modelling via neural networks, *Decisions in Economics and Finance*, 46(1), 157–176.

[31] Perla, F., Richman, R., Scognamiglio, S., Wüthrich, M.V. (2024). Accurate and explainable mortality forecasting with the LocalGLMnet. *Scandinavian Actuarial Journal*, 2024(7), 739–761.

[32] Plat, R. (2009). On Stochastic Mortality Modeling, *Insurance: Mathematics and Economics*, 45(3), 393–404.

[33] Renshaw A.E., Haberman S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors, *Insurance: Mathematics and Economics*, 38(3), 556–570.

[34] Haberman S, Renshaw A. (2011). A Comparative Study of Parametric Mortality Projection Models, *Insurance: Mathematics and Economics*, 48(1), 35–55.

[35] Richman, R., Wüthrich, M. (2021). A neural network extension of the Lee–Carter model to multiple populations. *Annals of Actuarial Science*, 15(2), 346–366.

[36] Shen, Y., Yang, X., Liu, H., Li, Z. (2024). Advancing mortality rate prediction in European population clusters: integrating deep learning and multiscale analysis. *Scientific Reports*, 14(1), 6255.

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

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

[39] Scognamiglio, S. (2024). Multi-population mortality modelling and forecasting with divergence bounds. *Annals of Operations Research*, 1–19.

[40] Villegas, A.M., Kaishev, V.K., Millossovich, P. (2018). StMoMo: An R Package for Stochastic Mortality Modeling, *Journal of Statistical Software*, 84(3), 1–38.

[41] Wang, J., Wen, L., Xiao, L., Wang, C. (2024). Time-series forecasting of mortality rates using transformer, *Scandinavian Actuarial Journal*, 2024(2), 109–123.

[42] Wiśniowski A, Smith P.W., Bijak J, Raymer J, Forster JJ. (2015). Bayesian Population Forecasting: Extending the Lee-Carter Method. *Demography*. 52(3), 1035–1059.

[43] World Bank Group (2024). *Mortality rate, adult, male (per 1,000 male adults) - European Union*, https://data.worldbank.org/indicator/SP.DYN.AMRT.MA?locations=EU

[44] Zhou, R., Li, J.S.-H., Tan, K.S. (2013). Pricing standardized mortality securitizations: a two-population model with transitory jump effects, *The Journal of Risk and Insurance*, 80(3), 733–774.