European Actuarial Journal
https://doi.org/10.1007/s13385-024-00377-5

CASE STUDY

A neural network approach for the mortality analysis of multiple populations: a case study on data of the Italian population

Maximilian Euthum[1,2] · **Matthias Scherer**[1] · **Francesco Ungolo**[1,3,4]

Received: 27 February 2023 / Revised: 14 July 2023 / Accepted: 19 December 2023

Abstract
A Neural Network (NN) approach for the modelling of mortality rates in a multi-population framework is compared to three classical mortality models. The NN setup contains two instances of Recurrent NNs, including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) networks. The stochastic approaches comprise the Li and Lee model, the Common Age Effect model of Kleinow, and the model of Plat. All models are applied and compared in a large case study on decades of data of the Italian population as divided in counties. In this case study, a new index of multiple deprivation is introduced and used to classify all Italian counties based on socio-economic indicators, sourced from the local office of national statistics (ISTAT). The aforementioned models are then used to model and predict mortality rates of groups of different socio-economic characteristics, sex, and age.

Keywords Case Study on Mortality · Longevity Risk · Neural Network · Multi-population · Deprivation Index · Socio-economic characteristics · Italian data

✉ Matthias Scherer
scherer@tum.de

[1] Chair of Mathematical Finance, Technical University of Munich, Garching bei München, Germany

[2] Munich Reinsurance Company (Munich RE), Munich, Germany

[3] School of Risk and Actuarial Studies, University of New South Wales, Kensington, NSW, 2052, Australia

[4] ARC Centre of Excellence in Population Ageing Research, University of New South Wales, Kensington, NSW 2052, Australia

Published online: 06 March 2024

1 Introduction

1.1 Mortality modelling: motivation, background, and literature

Since the seminal work of Lee and Carter [17], several stochastic models for the estimation and projection of mortality rates were developed during the last decades, see, e.g., the contributions of Brouhns et al. [2], Renshaw and Haberman [31], Cairns et al. [3], and Plat [28]. While these pioneering approaches analyzed single populations, models for multiple populations gained considerable importance in subsequent years; after it was found that the mortality profiles of multiple populations tended to converge (cf. Wilson [40]). Indeed, the multi-population paradigm often has advantages over modelling mortality rates for each population separately. Most notably, multi-population mortality models can capture common features of the mortality profile of similar populations such as neighbouring countries, populations showing similar socio-economic-, environmental-, or biological characteristics, while simultaneously reflecting population-specific features. This motivates their use for producing coherent projections of mortality rates.

Recently, Neural Network (NN)-based approaches for mortality modelling were proposed as an appealing alternative to classical stochastic models. Among the first examples is the work of Richman and Wüthrich [33], who analyses the Swiss population, and Hainaut [11] who uses French, UK, and US mortality rates to compare a NN approach to the Lee–Carter model w/wo cohort effects. Another contribution is Nigri et al. [23], who use a deep learning algorithm based on a two-step Recurrent Neural Network (RNN) to enhance the forecasts obtainable under the Lee–Carter model. Indeed, there have been many recent developments in the use of NNs in the context of multi-population mortality modelling, such as Perla et al. [27], which considers the use of one-dimensional (period effect only) RNN with Long Short-Term Memory (LSTM) and of Convolutional Neural Networks (CNN) to provide direct forecasts of the mortality rates compared to the two-step approach of Nigri et al. [23]. Lindholm and Palmborg [19] consider similar models with a focus on the optimal use of data for projection. Schnürch and Korn [35] extend the RNN and CNN by proposing a two-dimensional approach involving age and period. In a slightly different fashion, Scognamiglio [36] proposes a NN architecture for the joint calibration of individual Lee–Carter models based on its classical log-normal representation as well as the Poisson Lee–Carter version of Brouhns et al. [2]. An approach that allows for coherent predictions within sub-groups of similar population is provided in Perla and Scognamiglio [26]. Finally, Wang et al. [38] develop a framework which 'augments' the mortality dataset to construct an image of neighborhood mortality data around the central death rate and use two CNN approaches for projecting mortality rates.

1.2 Applications of (multi-population) mortality models

A key application of multi-population mortality modelling approaches is the analysis of the mortality levels based on socio-economic characteristics. Among others, understanding mortality is relevant to policymakers in order to propose and plan sustainable state pension reforms and budgets, or to address disparities between socio-economic groups. Understanding mortality from a statistical point-of-view is also relevant for private sector players such as insurance companies and pension funds, offering mortality-linked products like annuities and pensions and designing effective solutions for longevity risk management and transfer.

Considering specific populations that have already been investigated in the literature, let us mention Wen et al. [39], who compare several stochastic mortality models to fit mortality rates of small geographic areas in the UK (Lower Layer Super Output Areas), grouped in deciles of their Index of Multiple Deprivation. Cairns et al. [6] develop a multi-population mortality modelling approach for the analysis of the Danish population on the basis of the deciles of a newly created affluence index, which accounts for individual information on income and wealth.

1.3 Contributions: methodology and investigated population

This paper investigates the use of NNs to jointly model the mortality rates of multiple populations and compares the empirical results to classical stochastic mortality models. The model for the dynamics of mortality rates draws on the work of Richman and Wüthrich [32], where they propose the use of Recurrent Neural Networks (RNN) such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). Although computationally expensive, we focus on RNNs to exploit their suitability to deal with the time-series structure of mortality rates. Indeed, due to their recurrent connections, RNNs allow to maintain memory of past observations, since these let the information pass from past time steps to the current one. In this way, it is possible to model the information throughout the entire time series. Furthermore, RNNs can handle time series of variable length, due to their sequential processing of the data, cf. Hsu [12]. The NN approaches are then compared to well-established stochastic mortality models for multiple populations such as the Lee–Carter extension of Li and Lee [18] and the Common Age Effect model of Kleinow [16], as well as to the single-population approach represented by the Plat [28] model.

In our case study, we investigate Italian mortality data that is grouped according to socio-economic characteristics. More precisely, we first propose a new deprivation index on the basis of five variables. This index allows separating the 106 Italian counties into nine groups of different socio-economic level, which are then used as populations for the empirical analysis. To the best of our knowledge, this study is the first to explicitly address the use of NNs in the context of the analysis of multiple populations on the basis of socio-economic characteristics.

The structure of the remaining paper is as follows: Sect. 2 introduces and explains the Italian mortality data used for our analysis and the creation of the new Index of Multiple Deprivation. Section 3 describes how NNs are used in our study. Section 4 briefly introduces the models used for comparison. Then, Sect. 5 shows the results of the empirical analysis and, finally, Sect. 6 concludes.

2 Data

The collected mortality data comprises the 'number of deaths' and 'exposure-at-risk years' for males and females aged 50 to 95 living in Italy. It spans over the calendar years 1982–2018 and is granular to the level of the 106 Italian counties (called *provinces*). Our main source of data is the local office of national statistics (ISTAT).[1] The data were further processed to account for splits in some counties over the last thirty years (see Euthum [10] for details). Further elaborations were needed to calculate the central exposure to risk from the population data given for January 1st of each year.[2] We assume that the net migration effect does not bias our findings about mortality rates.[3] Furthermore, we collected a set of twelve indicators of socio-economic characteristics for each county, which we used to create an index of multiple deprivation. For further details regarding peculiarities and adjustment of data, we refer to the supplementary material[4].

2.1 Index of multiple deprivation

For each county, we originally collected a set of twelve indicators representing different aspects of the quality of life. Out of these indicators, a subset of five was ultimately selected, reflecting a proper mix of different types of socio-economic factors. These have been aggregated to a so-called 'Index of Multiple Deprivation' (IMD). In this way, the 106 provinces[5] from twenty regions could be classified to different groups based on their socio-economic indicators. The chosen variables are:

1. *Relative poverty (in percent):* The percentage of households with a consumption expenditure lower than the average *per-capita*, as estimated by ISTAT[6];

[1] The population data was collected from the website www.istat.it. The number of deaths from 1982 to 2002 were provided by the office of statistics.

[2] To approximate the central exposure for year i, populations from January 1st of year i and year $i+1$ were averaged. Further, some data cleaning was performed, for details we refer to Euthum [10].

[3] This assumption appears reasonable on the grounds of the relatively short length of the time series of available data, especially if we consider that censuses are conducted every ten years only, and also in consideration of the age-range we analyse, since migration (in general and so) between Italian provinces typically affects younger people. A detailed account of this observation is discussed in Cairns et al. [5] based on population data from England and Wales. Moreover, we are not aware of the existence of data that allows to track migration between Italian provinces.

[4] See the corresponding Github repository to this paper.

[5] 'Sardegna' has been kept to four provinces instead of five to keep things simpler for historic data.

[6] Accessible at https://www.istat.it/en/analysis-and-products/databases/statbase

2. *Primary care and residential and semi-residential facilities (measured in beds per 10,000 inhabitants).* This information is indicative of the expenditure in health care by the region where the county is located, which in its turn is affected by its wealth[7];
3. *Social services and benefits of municipalities, measured as costs in Euros per capita (2017 data).* Services include, e.g., day nursery, socio-educational services for early childhood, and so on. It is assumed that higher costs correspond to more investments in social services by municipalities and hence more benefits from a sociological point of view for the population;
4. *Unemployment rate (in percent for the population aged over 15)*;
5. *Number of felonies committed by persons convicted by final judgement (per 1,000 inhabitants).* From a statistical point of view, this variable is not significantly correlated with the other indicators used for building the index. It is assumed that a high relative number of felonies committed (by persons convicted by final judgement) suggests worse living conditions from a sociological point of view, but also comes along with a rather deficient economic situation.

A detailed description of these variables can be found on the 'Statbase portal' of Istat, which provides access to a large amount of data on the Italian population,[8] which is also the source of the underlying data (downloaded on 1st of March 2021 for year 2018).

The variables have been chosen based on a correlation analysis between them. The selection was then validated by a ranking measure, Kendall's tau, to check whether the rank of provinces changes significantly when omitting a variable from the selection. We found that Kendall's tau does not significantly change when omitting or adding a further variable to the five we selected.

Our index was constructed using z-scores to scale the five different variables to a comparable level and unit; the same approach is used in Osservatorio della salute [24]. This allowed to aggregate the single z-scores per province to a total value and finally permitted a ranking of the provinces. For each province, the z-score was calculated as

$$z^j = \sum_{i=1}^{5} z_i^j = \sum_{i=1}^{5} \frac{x_i^j - m_i}{s_i},$$

where i indicates the respective socio-economic classifier, x_i^j its value, j the county, m_i the mean of this classifier over all counties, and s_i the (unbiased) sample deviation over all counties.

The interpretation is intuitive: the higher the value z^j of a county, the worse is its socio-economic situation, or in other words, the deprivation in that area is higher. Implicitly, we assume that the standardized value of each variable has the same impact on the z-score (which could be easily generalized by introducing weights).

[7] Since our index is created with the interpretation of 'smaller values corresponding with better living conditions', we changed the sign of this covariate and also the next one.

[8] Accessible at https://www.istat.it/it/dati-analisi-e-prodotti/banche-dati/statbase

Then, the aim is to rank the counties on the basis of the values of their z-score. Hence, we created nine groups of counties, each with homogeneous populations, ranging between 6 and 7 million people. This split does not account for sex and age of the Italian population. This implicitly assumes that the counties have similar distributed population by age and sex. If, by using this criterion, we found that two counties had the same z-score value,[9] we allocate these in the same group. In this way, we aim at creating groups of comparable size, where the counties therein have similar socio-economic characteristics. The use of socio-economic indicators for the calendar year 2018 only, implies that the ranking and the corresponding groups do not change over time. This assumption may not reflect the socio-economic developments in Italy across the last 36 years. Nevertheless, we find this assumption reasonable in consideration of the results obtained in the preliminary analysis of the mortality and socio-economic data. Indeed, more deprived socio-economic counties are located in the south of Italy, compared to the historically wealthier areas in the north, as also demonstrated by the time series of the unemployment rate, and this evidence is further confirmed when analysing the evolution of the raw mortality rates.

We purposely opted for a simple and interpretable approach to create the index, that might be refined in further studies. Our main focus is to elaborate on mortality models as described later, and this simple and intuitive approach provides very plausible results. In any case, we remark how the creation of the groups was carried out on the basis of their order, rather than their index value.

To obtain a geographical impression of the index-based subdivision, the following map is reported, see Fig. 1. The color-scheme is as follows: Brighter colors indicate a lower index value, reflecting better socio-economic condition based on the IMD defined above. The original map was taken from the De Agostini website[10] and colored with standard graphic tools.

To obtain a first impression of mortality rates, the crude period death rates $\hat{m}(x,t,i) = \frac{d(x,t,i)}{E^c(x,t,i)}$ between 1982 and 2018 are plotted in Fig. 2 for the male and female population aged 68 and 83 years old, respectively. We observe:

- In general, mortality rates decrease over time and increase with age; which is consistent with the literature and the biological ageing process;
- Mortality rates for females are lower than for males, as widely observed in many other national population tables;
- The most deprived subpopulations (g1 in dark blue) appear to have higher mortality rates over the period analysed. The difference and ordering in subgroups is more pronounced for the female subpopulation, while for males, the difference in mortality rates is less evident for wealthier subgroups. This could be due to the chosen index or due to the underlying population. This effect may also be the consequence of a significant north–south division in terms of socio-economic well-being, while people still living longer in southern parts of the country, as

[9] As per two decimal digits.

[10] https://blog.geografia.deascuola.it/articoli/che-fine-faranno-le-province-italiane

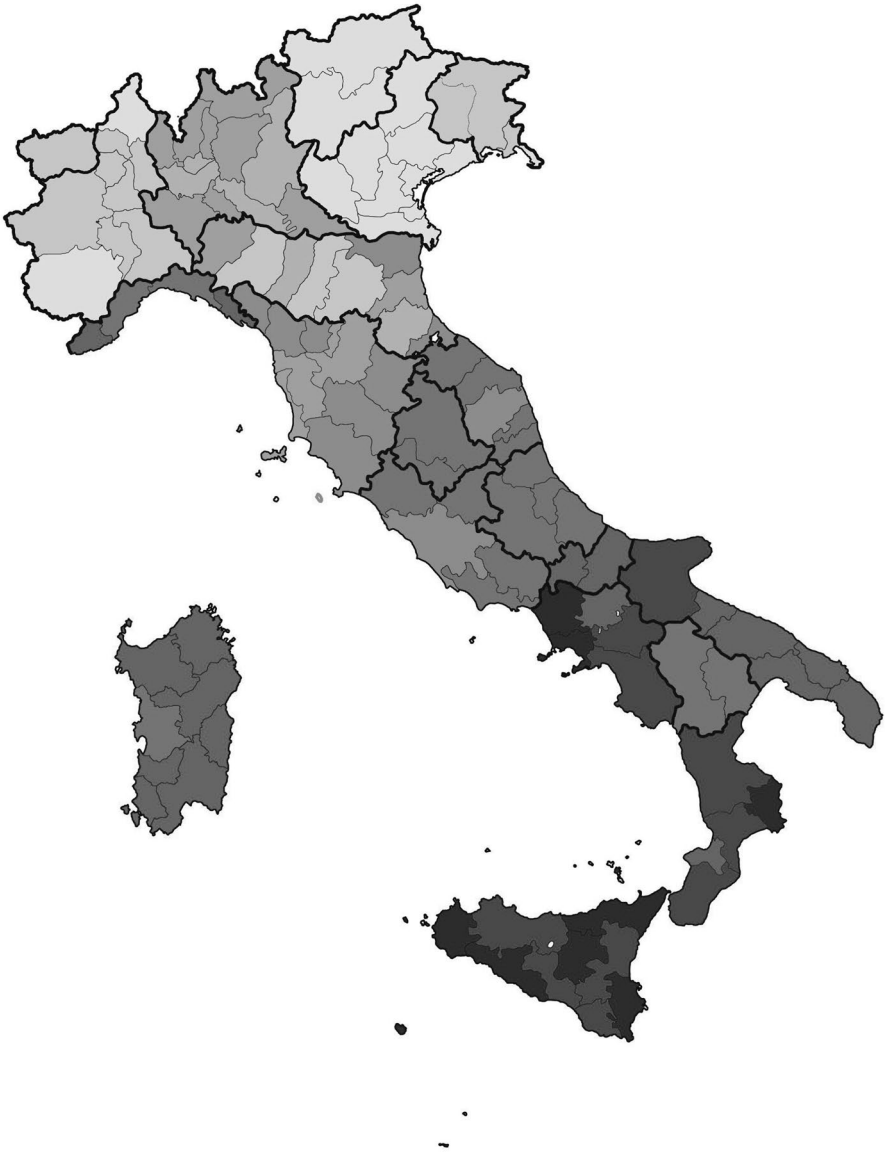

Fig. 1 Subdivision of the Italian provinces based on the new index. We observe the tendency of northern provinces having smaller index values, i.e. better living conditions

it is the case of regions like Sardinia and Calabria, which are famous for their exceptional longevity (see Poulain et al. [29]);

- In the earlier years of the study, deprivation trends and differences are harder to detect. This may be caused by the fact that the socio-economic analysis

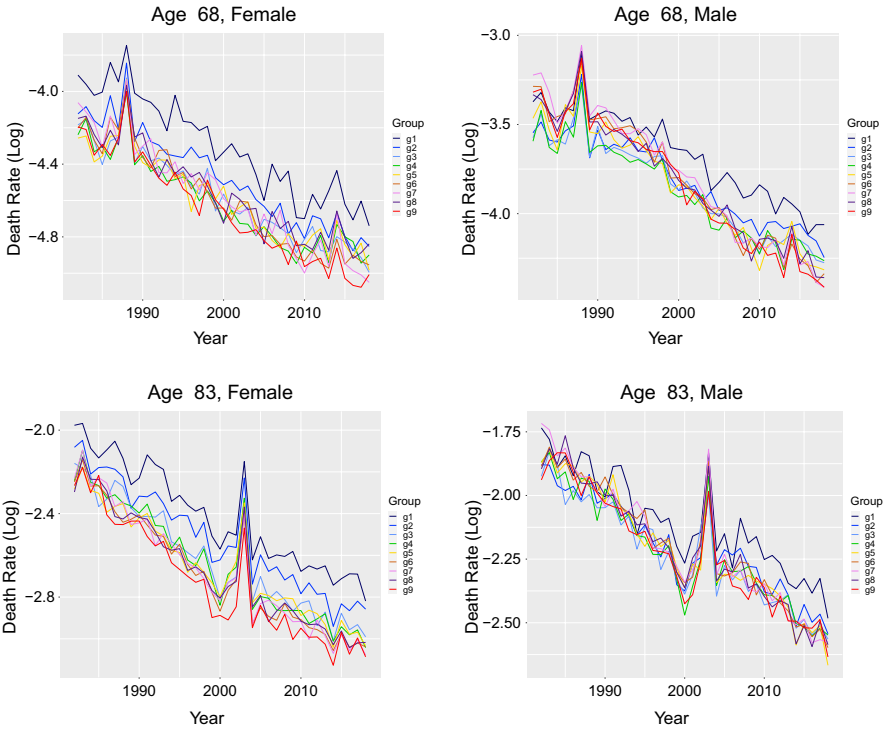

Fig. 2 Crude death rates in log-scale for ages 68 resp. 83, female and male population. Group 1 (g1) is the worst group socio-economically speaking

was based on indicators of the year 2018, while different provinces evolved differently over decades;

- The spike in the mortality rates for males and females aged 83 in 2003 is likely to be due to the massive heatwave of the summer in 2003 (Johnson et al. [15]). Further discussion is included in Sect. 5.

Throughout the analysis, we use the mortality rates of the first 33 calendar years 1982–2014 when training the models (training set), while those of 2015–2018 are used for predictions and forecasts (test set). We opted for this split in about 90%–10% due to the short length of the time-series of available data. We believe this split is a reasonable compromise between the need of sufficient data to fit the models (especially on a restricted age-range of a 50–95 years old population), whilst using the latest years to assess their predictive ability.

3 Multi-population NNs

In this section we introduce the NN approach adopted in this work, which draws on the work of Richman and Wüthrich [32], to analyse the Italian population. The key idea is to use age, sex, calendar year, and deprivation group (described in Sect. 2) as input, in order to obtain as output an estimate for mortality rates. The time series nature of the data requires the use of RNNs with Long Short-Term Memory structure or Gated Recurrent Unit structure, as pioneered in Richman and Wüthrich [33]. The general structure of a NN for regression includes:

- An *input layer* formed by several features or covariates;
- One or more *hidden layers* where inputs are processed, that is weighted and mapped inputs are passed on among different layers;
- An *output layer*, which returns a fitted value of the dependent variable.

Suppose there are $K \geq 1$ hidden layers in the network. Each layer includes $q_k \in \mathbb{N}$ neurons (by convenience q_0 is the dimension of the feature space which provides the input layer). The layers $z^{(k)}$ represent a mapping:

$$z^{(k)} : \mathbb{R}^{q_{k-1}} \to \mathbb{R}^{q_k}, \quad \mathbf{z} \mapsto z^{(k)}(\mathbf{z}) = \left(z_1^{(k)}(\mathbf{z}), \ldots, z_{q_k}^{(k)}(\mathbf{z})\right)', \tag{3.1}$$

$k = 1, \ldots, K$. $z_l^{(k-1)}$ is the l-th neuron from layer $k-1$:

$$z_j^{(k)}(\mathbf{z}) = \phi_j^{(k)}\left(w_{j,0}^{(k)} + \sum_{l=1}^{q_{k-1}} w_{j,l}^{(k)} z_l^{(k-1)}\right) =: \phi_j^{(k)}\left\langle \mathbf{w}_j^{(k)}, \mathbf{z}^{(k-1)} \right\rangle, \quad \text{for } j = 1, \ldots, q_k, \tag{3.2}$$

where $\mathbf{w}_j^{(k)} = (w_{j,0}^{(k)}, \ldots, w_{j,q_{k-1}}^{(k)})' \in \mathbb{R}^{1+q_{k-1}}$ are the weights and parameters to be trained in the model. $\phi_j^{(k)}$ denotes the activation function. This function determines the output of a layer, or neuron, in other words it decides whether a neuron is activated or not. The aim of the activation function is to introduce non-linearity into the output of a neuron.

3.1 Recurrent neural networks (RNN)

For RNNs, one has given input variables $(\tilde{x}_1, \ldots, \tilde{x}_T)$ in time series structure with components $\tilde{\mathbf{x}}_t \in \mathbb{R}^{q_0}$ for $t = 1, \ldots, T$. For the k-th layer, we define the mappings as follows:

$$z^{(k)} : \mathbb{R}^{q_{k-1}+q_k} \to \mathbb{R}^{q_k}, \quad (\mathbf{z}_t^{(k-1)}, \mathbf{z}_{t-1}^{(k)}) \mapsto z_t^{(k)} = z^{(k)}(\mathbf{z}_t^{(k-1)}, \mathbf{z}_{t-1}^{(k)}), \tag{3.3}$$

where

$$
\begin{aligned}
z_t^{(k)} &= z^{(k)}(\mathbf{z}_t^{(k-1)}, \mathbf{z}_{t-1}^{(k)}) \\
&= \left(\phi\Big(\langle \mathbf{w}_1^{(k)}, \mathbf{z}_t^{(k-1)}\rangle + \langle \mathbf{u}_1^{(k)}, \mathbf{z}_{t-1}^{(k)}\rangle\Big), \ldots, \phi\Big(\langle \mathbf{w}_{q_k}^{(k)}, \mathbf{z}_t^{(k-1)}\rangle + \langle \mathbf{u}_{q_k}^{(k)}, \mathbf{z}_{t-1}^{(k)}\rangle\Big)\right)^T \\
&=: \phi\Big(\Big\langle W^{(k)}, \mathbf{z}_t^{(k-1)}\Big\rangle + \Big\langle U^{(k)}, \mathbf{z}_{t-1}^{(k)}\Big\rangle\Big).
\end{aligned}
$$

The individual neurons $1 \leq j \leq q_k$ are modeled as

$$
z_{j,t}^{(k)} = \phi\Big(\langle \mathbf{w}_j^{(k)}, \mathbf{z}_t^{(k-1)}\rangle + \langle \mathbf{u}_j^{(k)}, \mathbf{z}_{t-1}^{(k)}\rangle\Big) = \phi\left(w_{j,0}^{(k)} + \sum_{l=1}^{q_{k-1}} w_{j,l}^{(k)} z_{l,t}^{(k-1)} + \sum_{l=1}^{q_k} u_{j,l}^{(k)} z_{l,t-1}^{(k)}\right), \tag{3.4}
$$

where ϕ is the (non-linear) activation function, which will be the same for all neurons. The weights are $W^{(k)} = (\mathbf{w}_1^{(k)}, \ldots, \mathbf{w}_{q_k}^{(k)})^T \in \mathbb{R}^{q_k \times (1+q_{k-1})}$ (including an intercept, see above) and $U^{(k)} = (\mathbf{u}_1^{(k)}, \ldots, \mathbf{u}_{q_k}^{(k)})^T \in \mathbb{R}^{q_k \times q_k}$ (excluding an intercept) which are identical for all time points t, as these weight matrices are homogeneous over time.

Note, $\tilde{\mathbf{x}}_t \in \mathbb{R}^{q_0}$ is an input in layer $k = 1$.

3.1.1 Long short-term memory structure (LSTM)

The LSTM type RNN has cycles in information transmission which are provided by a so-called cell state process. This process stores the available memory and allows for information from previous points in time to be included in further time steps when modelling the network. Shortly, this leads to the main equations for this specific NN:

$$
\mathbf{z}_t^{(k)} := z^{(k)}\Big(\mathbf{z}_t^{(k-1)}, \mathbf{z}_{t-1}^{(k)}, \mathbf{c}_{t-1}^{(k)}\Big) = o_t^{(k)} \circ \phi\Big(\mathbf{c}_t^{(k)}\Big) \in \mathbb{R}^{q_k}, \tag{3.5}
$$

where the cell state $\mathbf{c}_t^{(k)}$ is given by

$$
\begin{aligned}
\mathbf{c}_t^{(k)} :=& c^{(k)}\Big(\mathbf{z}_t^{(k-1)}, \mathbf{z}_{t-1}^{(k)}, \mathbf{c}_{t-1}^{(k)}\Big) \\
=& f_t^{(k)}\Big(\mathbf{c}_{t-1}^{(k)}\Big) + i_t^{(k)} \circ \phi_{\tanh}\Big(\langle W_c^{(k)}, \mathbf{z}_t^{(k-1)}\rangle + \langle U_c^{(k)}, \mathbf{z}_{t-1}^{(k)}\rangle\Big) \in \mathbb{R}^{q_k}.
\end{aligned} \tag{3.6}
$$

Here, $f_t^{(k)}$, $i_t^{(k)}$, and $o_t^{(k)}$ are the *Forget gate*, the *Input gate*, and the *Output gate*, respectively. These are trained to decide which information is handed in and which one is excluded at time t. $\phi_{\tanh}$ denotes the selected activation function.[11] The mentioned gates have their own weight matrices and are defined as follows:

- *Forget gate* (loss of memory gate):

$$
f_t^{(k)} := f^{(k)}\Big(\mathbf{z}_t^{(k-1)}, \mathbf{z}_{t-1}^{(k)}\Big) = \phi_\sigma\Big(\langle W_f^{(k)}, \mathbf{z}_t^{(k-1)}\rangle + \langle U_f^{(k)}, \mathbf{z}_{t-1}^{(k)}\rangle\Big) \in (0,1)^{q_k}, \tag{3.7}
$$

[11] Here, the activation function is defined as $x \mapsto \phi(x)$, $\phi(\cdot) = \tanh(\cdot)$

- *Input gate* (memory update gate):

$$i_t^{(k)} := i^{(k)}\left(\mathbf{z}_t^{(k-1)}, \mathbf{z}_{t-1}^{(k)}\right) = \phi_\sigma\left(\langle W_i^{(k)}, \mathbf{z}_t^{(k-1)}\rangle + \langle U_i^{(k)}, \mathbf{z}_{t-1}^{(k)}\rangle\right) \in (0,1)^{q_k}, \tag{3.8}$$

- *Output gate* (release of memory information rate):

$$o_t^{(k)} := o^{(k)}\left(\mathbf{z}_t^{(k-1)}, \mathbf{z}_{t-1}^{(k)}\right) = \phi_\sigma\left(\langle W_o^{(k)}, \mathbf{z}_t^{(k-1)}\rangle + \langle U_o^{(k)}, \mathbf{z}_{t-1}^{(k)}\rangle\right) \in (0,1)^{q_k}. \tag{3.9}$$

3.1.2 Gated recurrent unit (GRU)

The second RNN architecture used in this work is the GRU, first introduced by Cho et al. [7]. A problem when using LSTMs is that they are rather complex and computationally expensive. GRU networks are slightly simpler and can still provide good results as we observe in Sect. 5. The output structure is the same as for the LSTM, while the corresponding gates show some differences. In GRU architectures, only two different gates are used, the so-called *Reset gate* and the *Update gate*, denoted respectively by $r_t^{(k)}$ and $u_t^{(k)}$.

The gate activations are given by

$$\begin{aligned}\mathbf{z}_t^{(k)} :=& z^{(k)}\left(\mathbf{z}_t^{(k-1)}, \mathbf{z}_{t-1}^{(k)}\right)\\ =& r_t^{(k)}\left(\mathbf{z}_{t-1}^{(k)}\right) + (1 - r_t^{(k)}) \circ \phi\left(\langle W^{(k)}, \mathbf{z}_t^{(k-1)}\rangle + u_t^{(k)} \circ \langle U^{(k)}, \mathbf{z}_{t-1}^{(k)}\rangle\right) \in \mathbb{R}^{q_k},\end{aligned} \tag{3.10}$$

for general weight matrices of dimensions as those above. Here, no cell process c_t comes into play.

Note, if r_t approaches 1 in a component, then there is no reset for this component (neuron) in the sense that only the old activation from the previous time step is taken also in time step t. If r_t approaches 0, the old value is reset and replaced by a new value. This new value undergoes some update for information from the previous layer or not, depending on the values of u_t. Consequently, the number of parameters decreases. The gates are defined as follows:

- *Reset gate:*

$$r_t^{(k)} := r^{(k)}\left(\mathbf{z}_t^{(k-1)}, \mathbf{z}_{t-1}^{(k)}\right) = \phi_\sigma\left(\langle W_r^{(k)}, \mathbf{z}_t^{(k-1)}\rangle + \langle U_r^{(k)}, \mathbf{z}_{t-1}^{(k)}\rangle\right) \in (0,1)^{q_k}, \tag{3.11}$$

- *Update gate:*

$$u_t^{(k)} := u^{(k)}\left(\mathbf{z}_t^{(k-1)}, \mathbf{z}_{t-1}^{(k)}\right) = \phi_\sigma\left(\langle W_u^{(k)}, \mathbf{z}_t^{(k-1)}\rangle + \langle U_u^{(k)}, \mathbf{z}_{t-1}^{(k)}\rangle\right) \in (0,1)^{q_k}. \tag{3.12}$$

Layer (type)	Output Shape	Param #	Connected to
Input (InputLayer)	[(None, 10, 5)]	0	
LSTM1 (LSTM)	(None, 10, 20)	2080	Input[0][0]
LSTM2 (LSTM)	(None, 10, 15)	2160	LSTM1[0][0]
LSTM3 (LSTM)	(None, 10)	1040	LSTM2[0][0]
Groups (InputLayer)	[(None, 1)]	0	
Concat (Concatenate)	(None, 11)	0	LSTM3[0][0] Groups[0][0]
Output (Dense)	(None, 1)	12	Concat[0][0]

Total params: 5,292
Trainable params: 5,292
Non-trainable params: 0

Fig. 3 Parameters and concatenations, example of three layered LSTM network model, female population, R-output. This table is a standard R-output which shows the underlying construction of the model

3.2 Implementation of the NN approach

RNNs were implemented in R (R Core Team [30]). Training and forecasting was performed using the package `keras` Chollet et al. [8]. The choice of the parameters is motivated by the work of Richman and Wüthrich [33], as number of layers, number of neurons, type of activation functions, and so on. When experimenting with different hyper-parameters we did not notice any substantial differences or improvements in the results. The R code for data pre-processing, similar in spirit to Richman and Wüthrich [33], can be found in the Github repository https://github.com/maxeuthum/Multipopulation-Mortality-Models[12]. This repository also includes the code for fitting the models, with a detailed description of the performed operations line-by-line.

Input values were smoothed over 5 neighbouring ages to predict mortality at central age x. Therefore, for group i we obtain:

$$\mathbf{y}_{t,x} = \left(\log(m)_{(x-2)\vee 50,t}, \log(m)_{(x-1)\vee 50,t}, \log(m)_{x\vee 50,t}, \log(m)_{(x+1)\wedge 95,t}, \log(m)_{(x+2)\wedge 95,t}\right)^T \in \mathbb{R}^5,$$

where, for general variables $x, y \in \mathbb{R}$, $x \vee y = \max\{x, y\}$, and $x \wedge y = \min\{x, y\}$.

Furthermore, a look-back period of $T = 10$ years was determined. This means that the previous 10 years were taken as an input to predict mortality rates for year t. This is the same look-back period as in Richman and Wüthrich [33].

The training data set $\mathcal{T}$ was defined as

$$\mathcal{T} = \{\left(\mathbf{y}_{t-T,x}, \ldots, \mathbf{y}_{t-1,x}, \mathbf{y}_{t,x}\right); \quad 50 \le x \le 95, \quad 1982 + T \le t \le 2014\},$$

where $\mathbf{y}_{t,x}$ denote the log-mortality rates. Hence, we have log-mortality rates for 46 1-year age groups (50 to 95), denoted by variable x, over 23 calendar

[12] For legal reasons the full dataset cannot be available. The file `Finale Daten_Github.xlsx` in the repository contains an example of the available data to provide an idea of how these have been formatted for use in the analysis.

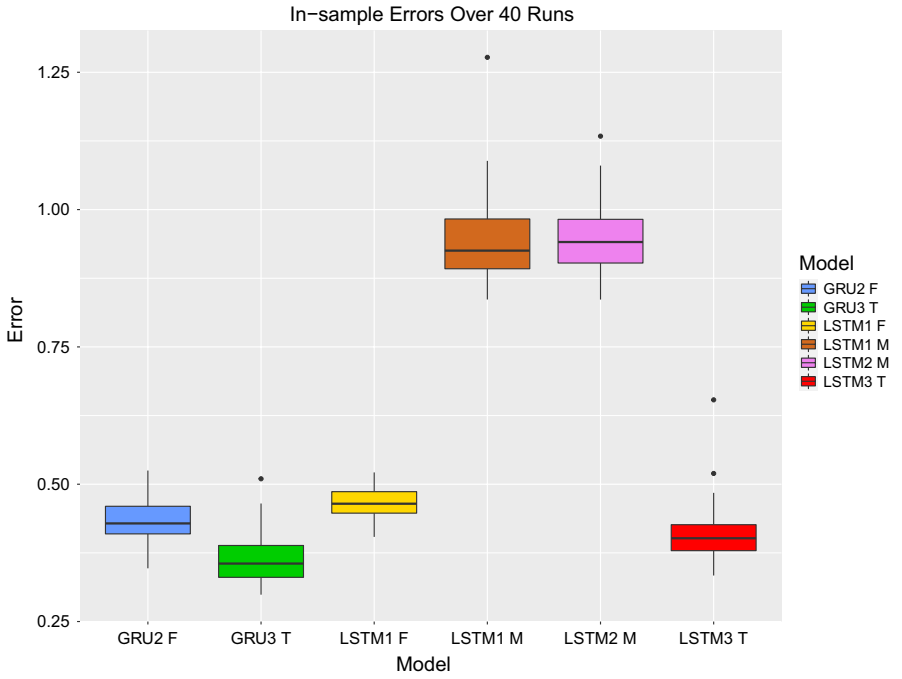

Fig. 4 In-sample loss for different models and populations, errors on 10^{-4}-scale. F, M, and T abbreviate females, males, and total (for combined), respectively

years $(2014 - (1982 + 10) + 1 = 23)$, and nine deprivation groups. This gives $46 \times 23 \times 9 = 9,522$ training samples of dimension 10×5 as input x (10 for the look-back period, and 5 for the neighbouring ages). In R, this was stored in 5 arrays of dimension $9,522 \times 10$. Last, $\mathbf{y}_{t,x}$ was stored in an array of dimension 9, 522 (for this array of course no look-back period or smoothing takes place).

As a final step for training data pre-processing, scaling was applied on the prediction data using the MinMax-scaler. In the end, the group indicator has been mapped to the data to obtain the input data consisting of these two parts returning the total input of the NN. We remark that several choices for the input dimension are reasonable, allowing for further features to be included in the model.

In this work, different types of RNN models have been modelled, namely LSTM and GRU networks for all groups simultaneously, for all three sex groups (males, females, and combined) separately. Respectively one, two, and three layers have been specified (Figs. 3, 4).

We opted for running 500 epochs[13] of the data with a batch size of 100 (again based on the work of Richman and Wüthrich [33]). To prevent overfitting, the training data set has been split into a training and validation set with a proportion of 80%

[13] Epochs indicate the number of times the learning algorithm runs through the entire (training) dataset. The batch size defines the number of data points that are run before an update of the internal model parameters happens.

and 20%, respectively. Furthermore, we implemented callbacks in order to select the calibration with the minimum loss on validation data among the 500 epochs.

As in other cases of NN-modelling, gradient descent algorithms were used to solve the optimization problem of the NN. Usually, these algorithms run until some stopping criterion is met such as the case where the error which we seek to minimize lies within a certain range. As stated in Richman and Wüthrich [33], an issue using early stopped solutions of gradient descent methods is that the resulting calibrations depend on the chosen seed of the algorithm. As a consequence, the results from different runs of the NN may be substantially deviating. For example, a certain run may lead to really good predictions, the next run to quite bad predictions. Hence, in our case study each model has been run 40 times, each time with different starting values. Then, we average the outcomes to assure not to have an outlying model by sheer coincidence. Hence, for each model the output (the log-mortality rates) has been transformed by the exponential function to obtain the mortality rates, which are then finally averaged.

3.2.1 Forecasting

The prediction of the log-mortality rates for the years 2015–2018 is carried out by means of an iterative approach, which makes no explicit use of time-series techniques. For example, the 10-year window 2005–2014 is needed to predict mortality for the year 2015. These new predictions are combined with the observations from the years 2006–2014, which yields a further 10-year window range to predict mortality for 2016, and so on. This recursive method allows for predictions of the 4 out-of-sample years. However, for longer prediction intervals this method may cause some implausible predictions.

4 Competing stochastic mortality models

We compare the results obtained from the NN approach of this work with three well established stochastic mortality (multi-)population models (thereafter referred as SMM), namely the Li and Lee [18] model, the Common Age Effect (CAE) model of Kleinow [16], and the Plat [28] model. All three models assume that the number of deaths $D(x, t, i)$ at age x, calendar year t, for group i is Poisson distributed with rate $E^c(x,t,i) \cdot m(x,t,i)$, where $m(x, t, i)$ denotes the underlying death rate and $E^c(x,t,i)$ is the central exposed at risk:

$$D(x,t,i) \sim \text{Poi}\big(E^c(x,t,i) \cdot m(x,t,i)\big).$$

The parameters underlying the specification are estimated by using maximum likelihood, as proposed by Enchev et al. [9]. We briefly introduce these models below.[14]

[14] The R code for implementing these model is available in the Github repository https://github.com/maxeuthum/Multipopulation-Mortality-Models

4.1 Li and Lee (LL) model

This model extends the single-population Lee and Carter [17] model to the analysis of multiple populations. The Li and Lee [18] model introduces a set of parameters which are common to the set of analysed groups, as well as group-specific parameters capturing the unexplained variance.

Model 4.1 (Li and Lee model) For age (in years) x, time period t, and group i, the Li & Lee model describes the logarithm of the 'central mortality rate' $m(x, t, i)$ as

$$\log\left(m(x,t,i)\right) = \alpha(x,i) + B(x)\,K(t) + \beta(x,i)\,\kappa(t,i), \quad t_{min} \le t \le t_{min} + T - 1. \tag{4.1}$$

Here, $m(x, t, i)$ can be approximated as the ratio between the death counts, denoted as $D(x, t, i)$, and the central exposure at risk $E^c(x, t, i)$.

$\alpha(x, i)$, $\beta(x, i)$, and $\kappa(t, i)$ are group-specific parameters. $\alpha(x, i)$ indicates the average over time of the log mortality rate. The common function $K(t)$ explains the evolution of the mortality over time for all groups, and $B(x)$ is a global age modulating parameter, indicating how rates change by age for changes in the time factor $K(t)$. $\beta(x, i)$ and $\kappa(t, i)$ have the same role as $B(x)$ and $K(t)$, but act on a group-specific level.

4.2 Common age effect (CAE) model

The CAE model of Kleinow [16] assumes that age effects are common to all populations, following the assumption that age effects may be very similar in countries sharing a similar socio-economic structure.

Model 4.2 (Kleinow model) For age x, time period t, and group i, the Kleinow model of order p assumes the following model for the logarithm of the central death rate

$$\log\left(m(x,t,i)\right) = \alpha(x,i) + \beta^{(1)}(x)\,\kappa^{(1)}(t,i) + \ldots + \beta^{(p)}(x)\,\kappa^{(p)}(t,i). \tag{4.2}$$

The order p follows from the allowance of further age–time interaction parameters. In our analysis we set $p = 2$.

4.3 Plat model

The third model we use for comparison was initially proposed by Plat [28]. However, in this work we use a simplified version, which includes two period-specific factors, without accounting for the cohort effect.

Model 4.3 (Plat model) For age x, time period t, and group i, the Plat model (without cohort effects) models the logarithm of the central death rates as

$$\log\left(m(x,t,i)\right) = \alpha(x,i) + \kappa^{(1)}(t,i) + \kappa^{(2)}(t,i)(x-\bar{x}), \tag{4.3}$$

where $\bar{x}$ denotes the average age of the observed age range. The first stochastic component $\kappa^{(1)}$ represents changes in the level of mortality for all ages, while $\kappa^{(2)}$ allows for changes in mortality to vary between ages.

4.3.1 Forecasting

For the models of Li & Lee, Kleinow, and Plat, we performed forecasts via a classical time series approach. In fact, time dependent κ-processes are modelled as a stochastic time series to predict mortality rates through ARIMA (Auto Regressive Integrated Moving Average) models. These can be readily implemented in R, e.g. by using the function `auto.arima` from the package `forecast` (Hyndman and Khandakar [14]).

5 Empirical results

All models are fitted based on data spanning from 1982 to 2014. However, fitted values are just compared for the years 1992 to 2014, since the NN-based approach does not deliver values for the first ten years, see Richman and Wüthrich [33]. In what follows, then $i \in \{1,\dots,9\}$, $x \in \{50,\dots,95\}$, and $t \in \{1992,\dots,2014\}$. Furthermore, we graphically inspect the models by using the standardized residuals, as defined in Wen et al. [39], see also Table 4.

5.1 In-sample fit

We first compare the three competing stochastic mortality models, that we briefly introduced in Sect. 4, in terms of their explanation ratio. Then we compare these to the rates obtained by using the NN models of this paper.

Table 1 shows the explanation ratios for the models of Li & Lee, Kleinow, and Plat, defined for group i and model M as

$$R_i^M = 1 - \frac{\sum_{x,t}\left(\log\frac{d(x,t,i)}{E^c(x,t,i)} - \log\left(\hat{m}(x,t,i)\right)\right)^2}{\sum_{x,t}\left(\log\frac{d(x,t,i)}{E^c(x,t,i)} - \alpha^c(x,i)\right)^2}, \tag{5.1}$$

where $\alpha^c(x,i) := \frac{1}{T}\sum_t \log\frac{d(x,t,i)}{E^c(x,t,i)}$ is the average log crude death rate over time. The explanation ratio is useful for analysing how much information about the crude death rates $\frac{d(x,t,i)}{E^c(x,t,i)}$ is explained by the respective model.

We observe how the Li & Lee model performs best in terms of explanation ratios for eight out of nine females subgroups, and the same is noted for the

Table 1 Explanation ratios for different models

Deprivation	Female			Male			Combined		
Group	R_i^{LL}	R_i^{CAE}	R_i^{Plat}	R_i^{LL}	R_i^{CAE}	R_i^{Plat}	R_i^{LL}	R_i^{CAE}	R_i^{Plat}
1	**0.862**	0.853	0.848	0.782	**0.823**	0.811	**0.881**	0.867	0.874
2	**0.868**	0.856	0.838	**0.822**	0.810	0.804	**0.877**	0.854	0.870
3	**0.862**	0.857	0.822	0.842	**0.846**	0.826	**0.893**	0.879	0.875
4	**0.859**	0.848	0.802	0.855	**0.856**	0.836	**0.892**	0.881	0.872
5	**0.831**	**0.831**	0.774	0.868	**0.872**	0.842	**0.886**	0.879	0.852
6	**0.876**	0.866	0.826	0.896	**0.903**	0.882	**0.917**	0.910	0.901
7	0.876	**0.882**	0.842	0.888	**0.912**	0.894	**0.929**	0.924	0.917
8	**0.866**	0.857	0.826	0.880	**0.890**	0.864	**0.910**	0.902	0.895
9	**0.866**	0.861	0.831	0.885	**0.898**	0.866	**0.914**	0.906	0.909
Average	**0.863**	0.857	0.823	0.857	**0.868**	0.847	**0.900**	0.889	0.885

The highest ratios for a specific sex are emphasized in bold

Table 2 AIC values for different models

Model	Female	Male	Both sexes
Li and Lee	**-66,185,533**	-67,260,680	**-134,742,870**
Kleinow	-66,188,446	**-67,256,443**	-134,750,703
Plat	-66,194,573	-67,261,519	-134,749,793

Lowest values for specific sex in bold

Kleinow model when looking at the male population. The Plat model, analysed at the level of a single population, still yield comparable explanation ratios, despite being lower compared to the Li & Lee and the CAE model of Kleinow. Furthermore, the Plat model has a lower number of parameters, which may in part explain these results. Our results are further confirmed by the analysis of their Akaike Information Criterion (AIC) (Akaike [1]), shown in Table 2, which indicates the relative quality of a statistical model with a penalty term for the number of parameters. The AIC is calculated as

$$AIC = -2 \cdot \ell(\hat{\theta}) + 2 \cdot p, \tag{5.2}$$

where $\ell(\hat{\theta})$ indicates the log-likelihood value at (optimal) parameter $\hat{\theta}$ and p denotes the number of parameters of the respective model (see Table 9).

Tables 3 shows the mean squared error for model M and subpopulation i, calculated as follows:

$$\text{MSE}_i^M = \frac{1}{n \cdot T} \sum_{x,t} \left(m(x,t,i) - \hat{m}(x,t,i)^M \right)^2, \tag{5.3}$$

where $\hat{m}(x,t,i)^M$ denotes the fitted mortality rate derived from model M. Table 4 shows the MSE for the RNN models analysed in this paper.

Table 3 Mean squared errors for different models and groups

Deprivation	Female			Male			Combined		
Group	LL	CAE	Plat	LL	CAE	Plat	LL	CAE	Plat
1	**1.3327**	1.4275	1.3743	3.6259	3.0236	**2.7788**	1.4856	1.6027	**1.4515**
2	**0.9692**	1.2805	1.2509	2.0578	**1.8407**	1.8716	1.1946	1.2721	**1.1912**
3	0.7436	**0.7396**	0.8416	2.4992	**1.9319**	2.0930	0.9395	0.9666	**0.9221**
4	**0.6853**	0.8357	0.8987	2.8050	**2.1437**	2.3239	**0.8530**	0.9327	0.9533
5	**0.8433**	0.9518	1.0570	2.4233	2.1329	**1.8905**	**0.8928**	1.0593	1.0390
6	**0.5896**	0.6464	0.7379	2.6853	2.3437	**2.1192**	**0.7630**	0.8736	0.7749
7	**0.6317**	0.6708	0.7788	3.7611	3.6867	**2.8752**	0.7681	0.8645	**0.7284**
8	**0.7108**	0.7783	0.7892	3.4809	**2.4448**	2.6673	0.8677	1.0031	**0.8662**
9	**0.6903**	0.7570	0.8233	4.4645	**2.8037**	3.3212	0.8259	0.9085	**0.7876**
Sum	**7.1964**	8.0876	8.5518	27.8031	22.3517	**21.9406**	**8.5902**	9.4831	8.7140

Lowest values for specific sex in bold. Scale: 10^{-4}

Table 4 Mean squared errors for different Recurrent network models and groups

Deprivation	Female			Male			Combined		
Group	E_i^{LSTM3}	E_i^{LSTM2}	E_i^{LSTM1}	E_i^{LSTM3}	E_i^{LSTM2}	E_i^{LSTM1}	E_i^{LSTM3}	E_i^{LSTM2}	E_i^{LSTM1}
1	0.6477	**0.6242**	0.6495	1.1999	**1.1560**	1.2223	0.5309	**0.5071**	0.5362
2	0.5710	**0.5521**	0.5907	0.6545	**0.6475**	0.6853	0.4347	**0.4295**	0.4485
3	0.4006	**0.3876**	0.4139	0.8020	**0.7556**	0.8101	0.3449	**0.3294**	0.3451
4	0.3114	**0.2992**	0.3196	0.6739	**0.6364**	0.6783	0.2379	**0.2245**	0.2413
5	0.4202	**0.4039**	0.4270	0.8119	**0.7816**	0.8481	0.3581	**0.3511**	0.3709
6	0.3363	**0.3257**	0.3411	0.7937	**0.7621**	0.8044	0.3046	**0.2900**	0.3096
7	0.3213	**0.3123**	0.3215	0.9426	**0.8860**	0.9429	0.2753	**0.2599**	0.2771
8	0.2922	**0.2781**	0.3000	0.8838	**0.8420**	0.8981	0.2760	**0.2591**	0.2801
9	0.2979	**0.2827**	0.3007	0.7111	**0.6564**	0.7067	0.2772	**0.2580**	0.2766
Sum	3.5986	**3.4658**	3.6639	7.4734	**7.1236**	7.5963	3.0395	**2.9086**	3.0853
	E_i^{GRU3}	E_i^{GRU2}	E_i^{GRU1}	E_i^{GRU3}	E_i^{GRU2}	E_i^{GRU1}	E_i^{GRU3}	E_i^{GRU2}	E_i^{GRU1}
1	**0.4052**	0.4219	0.5124	**0.5439**	0.5636	0.8325	**0.3051**	0.3396	0.3979
2	**0.3478**	0.3599	0.4389	**0.3925**	0.4212	0.5785	**0.2744**	0.2982	0.3658
3	**0.2756**	0.2951	0.3586	**0.4108**	0.4216	0.5702	**0.2111**	0.2458	0.2905
4	**0.2633**	0.2686	0.2928	**0.3824**	0.3978	0.5464	**0.1955**	0.2099	0.2358
5	**0.3323**	0.3449	0.3901	**0.4831**	0.5321	0.6863	**0.2736**	0.3038	0.3401
6	**0.2736**	0.2774	0.3059	**0.4036**	0.4290	0.5686	**0.2219**	0.2499	0.2778
7	0.2577	**0.2562**	0.2752	**0.4324**	0.4461	0.6428	**0.2101**	0.2247	0.2450
8	**0.2291**	0.2327	0.2633	**0.4676**	0.4806	0.6703	**0.1826**	0.2038	0.2341
9	**0.2387**	0.2398	0.2612	**0.4237**	0.4296	0.5115	**0.1967**	0.2160	0.2408
Sum	**2.6232**	2.6966	3.0984	**3.9399**	4.1217	5.6069	**2.0710**	2.2918	2.6278

Lowest values for specific sex and network type in bold. Scale: 10^{-4}

Table 5 Mean squared errors for specifically tailored age ranges, females (54–86), and males (50–82), SMM models

Deprivation	Female			Male		
Group	LL	CAE	Plat	LL	CAE	Plat
1	0.0996	**0.0982**	0.1025	**0.1113**	0.1117	0.1147
2	**0.0855**	0.0904	0.0925	0.1004	**0.0969**	0.1026
3	**0.0795**	0.0851	0.0828	0.1006	**0.0985**	0.1035
4	**0.0859**	0.0875	0.0886	0.0899	**0.0869**	0.0922
5	0.0915	**0.0884**	0.0885	0.1052	**0.1036**	0.1059
6	**0.0770**	0.0792	0.0800	0.1178	**0.1169**	0.1189
7	**0.0790**	0.0806	0.0825	0.1081	**0.1067**	**0.1067**
8	**0.0804**	0.0867	0.0853	0.1098	**0.1059**	0.1090
9	**0.0782**	0.0793	0.0802	**0.1072**	0.1097	0.1115
Sum	**0.7566**	0.7755	0.7830	0.9503	**0.9367**	0.9650

Scale: 10^{-4}

From the results in Table 3, we observe that it is not possible to draw a decisive conclusion about the best performing model, since for different groups and sex there is a mixed evidence about the best performing model. When the data for both males and females are combined, then we observe that the Plat model has a better performance in the two extremes of the deprivation group, or in other words, shows a better performance for the least and most deprived county groups.

The two RNN models seem to sensibly improve the in-sample fit compared to the three competing stochastic models. In more detail, the two-layer LSTM outperforms the other two LSTM models over all socio-economic groups for males, females, and combined. On the other hand, the GRU models with three layers turn out to perform even better. For female and both sexes combined mortality rates, the most deprived groups show higher in-sample errors compared to less deprived ones. A similar evidence can be observed for the LL, CAE, and Plat models. This may be indicative of the issues of mortality models in general when fitting more deprived subpopulations via a multi-population approach. This can also be the effect of the created index in Sect. 2 and of the underlying data. Nevertheless, we note that the difference between females and males is smaller for the RNN approach compared to the competing models.

In conclusion, the multi-population RNN models outperform the well established stochastic mortality models when analysing county-based socio-economic subgroups of the Italian population, based on the mean squared error. The mean squared errors for males are higher compared to females and both sexes combined.

For a deeper inspection of these results we restrict the MSE analysis to a shorter age range for males and females, in a way such that they both have the same life expectancy. These ranges were chosen based on remaining life expectancy of Italian males at age 50 in 2017 (32.08 years) and Italian females at age 54 in year 2017 (approx. 32 years). Therefore, based on the Italian life tables from the Human Mortality Database [13], we restrict the analysis to the male population aged 50 to 82 and to the females aged 54 to 86.

Table 6 Mean squared errors for specifically tailored age ranges, females (54–86), and males (50–82), selected RNN models

Deprivation	Female			Male		
Group	GRU3	GRU2	LSTM1	GRU3	GRU2	LSTM1
1	0.0322	0.0318	0.0327	0.0287	0.0289	0.0293
2	0.0186	0.0185	0.0195	0.0193	0.0199	0.0189
3	0.0209	0.0211	0.0211	0.0173	0.0178	0.0178
4	0.0163	0.0164	0.0167	0.0159	0.0165	0.0154
5	0.0169	0.0168	0.0171	0.0171	0.0173	0.0180
6	0.0153	0.0151	0.0153	0.0184	0.0187	0.0181
7	0.0188	0.0185	0.0195	0.0208	0.0207	0.0199
8	0.0122	0.0122	0.0124	0.0157	0.0164	0.0162
9	0.0174	0.0173	0.0173	0.0167	0.0173	0.0172
Sum	0.1689	0.1677	0.1715	0.1700	0.1735	0.1707

Scale: 10^{-4}

Again, we recalculated the deprivation group-specific MSE for males and females for the restricted age-ranges, still based on the model fitted to the original dataset (males and females aged 50 to 95). The results are shown in Table 5 (LL, CAE, and Plat models) and Table 6 (for RNN).

5.2 Mean squared errors for specific ages

For the selected age ranges, we observe that for the three competing stochastic mortality models, the mean squared error is less than 0.1 of the mean squared error for all ages, even if these reduced age ranges cover 72% of the modelled years.

Again, we observe that for the female population the MSE is larger for the most deprived socio-economic groups for both the competing models as well as for the RNN-based ones. A similar evidence is obtained for the males at a lesser extent. For the models fitted using female subpopulation data, we still face more difficulties when considering the most deprived groups compared to less deprived ones, especially for SMM models. It is also evident, that all models within one model class (SMM or RNN) have similar mean squared errors. This means, the models perform very well for these ages through all selected models. Let us also mention the fact that mean squared errors in the RNN case are by far smaller than in the SMM case.

Furthermore, it is interesting to observe an alignment of female and male rate fitting (as it was aimed through the analysis of different age ranges): the SMM mean squared error for the male population were on average approximately 3.02 times compared to females. When restricting the age range observation, this factor reduces to about 1.23, which means that for our data both male and female rates are almost fitted equally well on average for an age range with similar remaining life expectation. For the three selected RNN models from above, the factors are 1.52 and 1.01 respectively, that is the alignment of female and male rates fitting ability is almost perfect in the RNN case.

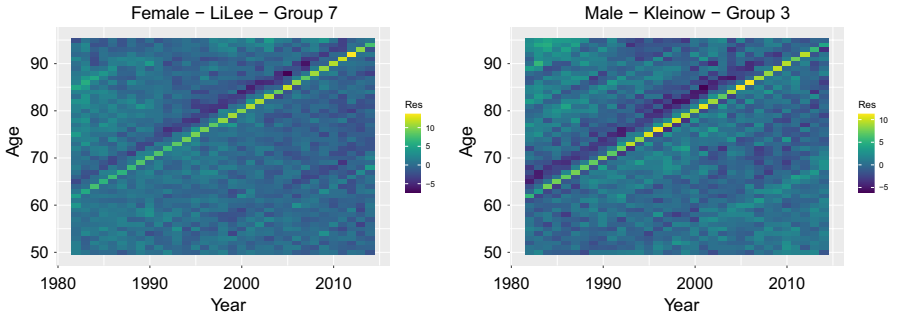

Fig. 5 Selected residual heatmap plots per group

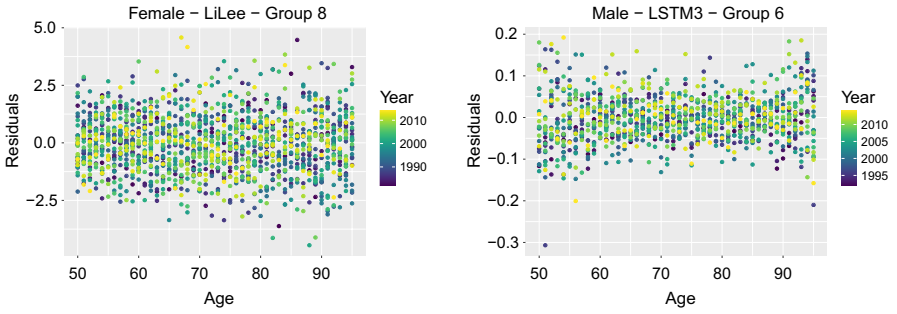

Fig. 6 Selected residual plots function of age, per group, both sexes

Concluding, it seems that the oldest people (of age 87–95) drive most of the mean squared error. In addition, one should be aware of looking at different age ranges for females and males based on the evidence that life expectancy significantly differs for these two groups. The point in time for older ages, where mortality is harder to fit due to more volatility in death and exposure data, starts earlier for males which could explain a higher mean squared error when looking at all ages.

Another possible reason for larger errors in the male population through all models may come from the difference in population sizes of the underlying deprivation groups. These are higher in the female case, since in the observed age ranges females outlive males. A larger sample size could give the statistical model more stability when estimating parameters. However, the difference is not very large[15].

5.2.1 Standardized residuals

We perform a graphical analysis of these results by investigating the standardized residuals. When a model fits the data reasonably well, standardized residuals should not exhibit any pattern based on years or ages. To quote Cairns et al. [4], *"if the model fits the data well, then the standardized residuals should be independent of*

[15] Of the population in Italy aged 50–95 on January 1st 2021, 46.31% were males (Istat Statbase).

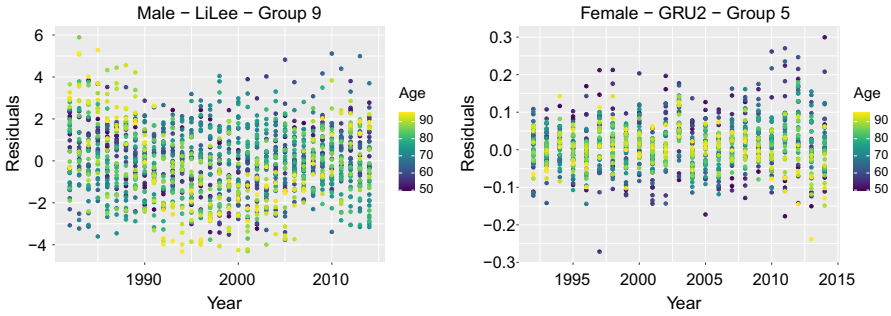

Fig. 7 Selected residual plots function of year, per group, both sexes

each other, meaning that the heat plot should exhibit a high degree of randomness, with no discernible patterns".

Figure 5 plots the heatmaps of the standardized residuals for the fitted rates under the Li and Lee model for group 7 and for the group 3 of the CAE model. The plots for the other groups under the three SMMs are similar. These are calculated for model M as

$$Z(x,t,i)^M = \frac{d(x,t,i) - E^c(x,t,i) \cdot \hat{m}(x,t,i)^M}{\sqrt{E^c(x,t,i) \cdot \hat{m}(x,t,i)^M}}, \tag{5.4}$$

where $\hat{m}(x,t,i)^M$ denotes the fitted mortality rate under model M.

The two plots show a diagonal line, which is indicative of a cohort effect for those individuals born around 1918, which corresponds to the end of the first world war and the Spanish flu pandemic. We discuss these points in more detail in Appendix A.2. For all other ages and years, the standardized residuals appear to lack any specific pattern (Figs. 6, 7).

For RNN models, the standardized residuals shown in Fig. 8 (for the other groups we have similar evidences) seem to be spread homogeneously across the ages and years. There are no visible cohort effects in residuals as in earlier models and residuals seem to be much smaller than in the previous models. These observations suggest that RNN models used in this work are able to capture cohort effects in the data and yield a better in-sample fit for observed years. In any case, we observe that for ages 70+/80+ in 2003, residuals are large and positive, meaning that the observed mortality is higher than expected. Conversely, for 2004 mortality rates seem to be over-estimated by the models. A possible explanation for this under-estimation in 2003 could be the massive European heat wave in summer 2003, see Johnson et al. [15], the hottest summer in Europe for centuries. ISTAT reported over 18,000 deaths in that summer compared to the year before (+ 11.6%). As reported in Mattone [21], 91.8% of these were aged 75 and older. This could explain in part the large number of negative residuals at older ages in 2004. These extraordinary circumstances may explain why the models could not fully detect such a sudden increase in the number

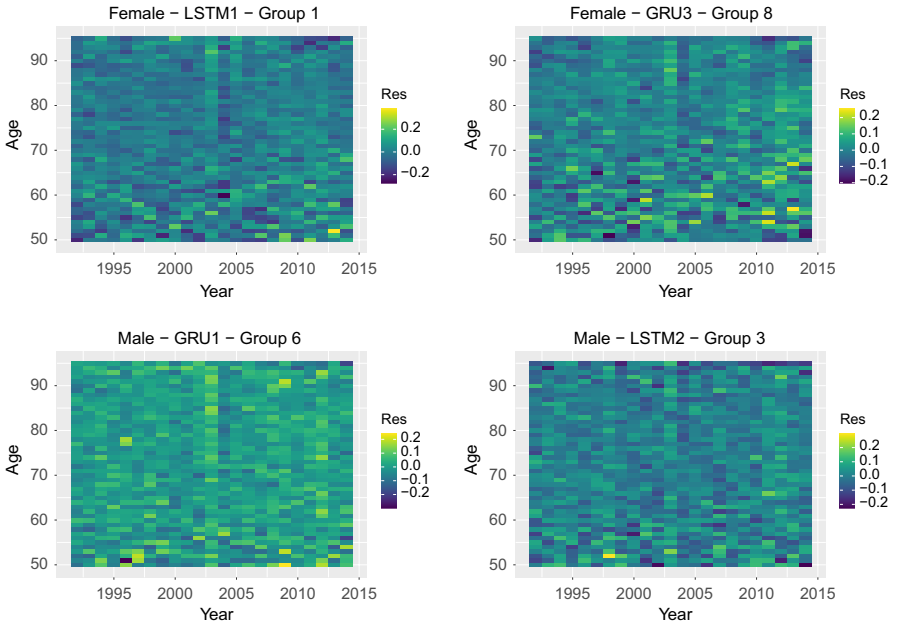

Fig. 8 Randomly selected residual heatmap plots per group

Table 7 Out-of-sample mean squared errors for different models and groups

Deprivation	Female			Male			Both Sexes		
Group	E_i^{LL}	E_i^{KL}	E_i^{Plat}	E_i^{LL}	E_i^{KL}	E_i^{Plat}	E_i^{LL}	E_i^{KL}	E_i^{Plat}
1	1.0038	**0.9955**	2.1453	7.7344	**4.2412**	4.8872	1.1059	**1.0278**	2.5211
2	**0.6560**	0.7621	1.4115	2.3270	**0.7531**	1.4893	**0.7326**	5.1101	1.1578
3	**0.7433**	1.1019	1.9106	3.2611	2.1388	**1.7976**	**1.0087**	9.9613	1.4781
4	**0.5366**	0.6045	16.3794	9.7053	**1.8354**	16.1711	8.8701	**0.8926**	13.5179
5	1.0720	**0.9237**	2.3226	**2.8685**	2.9456	3.2342	**0.8967**	1.1394	2.0517
6	1.4249	**1.3634**	2.7108	3.5877	4.0032	**2.2568**	**1.3117**	1.5394	2.0675
7	**0.9823**	1.9311	2.0970	2.8236	6.2054	**2.3788**	1.0659	**0.9724**	1.7426
8	1.2298	**1.1849**	2.3117	3.9409	3.5925	**1.6364**	**1.0616**	1.4791	1.7348
9	**1.3697**	2.2630	2.1508	5.4839	2.3254	**1.3812**	1.2547	**1.1114**	1.5911
Sum	**9.0183**	11.1301	33.4397	41.7324	**28.0406**	35.2326	**17.3079**	23.2334	27.8627

Lowest values for specific sex in bold. Scale: 10^{-4}

of deaths. A similar result might be observed in 2020 or later following the COVID-19 pandemic.

We further compare the residuals as function of the age (Fig. 6), and as a function of the calendar year (Fig. 7), where groups and sex have been chosen randomly and

where we exclude the data for the cohorts born in 1916, 1917, and 1920, based on observations in Appendix A.2.

Patterns are not observable if not for yellow points which indicate age 90+. In general, residuals for the Li and Lee model (the same holds for the other models of type SMM) spread on much higher levels than those of RNN models, by a factor larger than ten. However, both model types suggest evenly distributed residuals with exception for the oldest years which seem harder to model. Presumably, this latter observation is driven by the smaller exposure for older ages.

5.2.2 Out-of-sample measures

We analyse the out-of-sample performance of the implemented models by using a forecast period of four years, with $n = 46$ and $T = 4$. All figures have been obtained on the basis of the procedure described in Sect. 3 for RNN and Sect. 4 for the SMMs. Results are shown in Tables 7 and 8.

Table 8 Out-of-sample mean squared errors for different RNNs and groups

Deprivation	Female			Male			Both sexes		
Group	E_i^{LSTM3}	E_i^{LSTM2}	E_i^{LSTM1}	E_i^{LSTM3}	E_i^{LSTM2}	E_i^{LSTM1}	E_i^{LSTM3}	E_i^{LSTM2}	E_i^{LSTM1}
1	1.0916	0.9873	1.0225	2.0643	2.1807	2.0877	1.3548	1.3108	1.2765
2	0.6805	0.6004	0.6265	1.0510	1.1554	1.1369	0.7402	0.7240	0.7030
3	0.7193	0.6268	0.6695	0.9951	1.0019	0.9764	0.6845	0.6732	0.6642
4	0.7020	0.6073	0.6596	1.1020	1.1364	1.0489	0.7590	0.7397	0.6967
5	0.7966	0.6839	0.7631	1.1623	1.2082	1.0256	0.8945	0.8979	0.8581
6	1.2054	1.0858	1.1674	1.3938	1.5168	1.4414	1.1915	1.1963	1.1657
7	0.8209	0.7249	0.8058	1.1760	1.2602	1.1435	0.7999	0.8020	0.7914
8	1.0526	0.9342	0.9979	1.2587	1.2837	1.1041	1.0766	1.0483	0.9947
9	1.0315	0.8979	0.9617	1.2335	1.3099	1.2287	1.0079	0.9829	0.9457
Sum	8.1005	**7.1485**	7.6741	11.4367	12.0532	**11.1932**	8.5090	8.3750	**8.0959**
	E_i^{GRU3}	E_i^{GRU2}	E_i^{GRU1}	E_i^{GRU3}	E_i^{GRU2}	E_i^{GRU1}	E_i^{GRU3}	E_i^{GRU2}	E_i^{GRU1}
1	0.9718	0.8977	1.1457	2.1442	2.2496	2.2774	1.1988	1.1511	1.2861
2	0.5922	0.5266	0.7010	1.1787	1.1653	1.1986	0.6606	0.6301	0.7223
3	0.6240	0.5631	0.7322	1.0778	1.0036	1.0708	0.6095	0.5845	0.6802
4	0.5709	0.5095	0.6293	1.0881	1.1330	1.1358	0.6224	0.5863	0.6684
5	0.6083	0.5181	0.7028	1.3581	1.4528	1.2105	0.6910	0.6191	0.7283
6	1.0879	1.0308	1.1805	1.5759	1.6477	1.6592	1.1063	1.0329	1.1135
7	0.6631	0.5915	0.6908	1.1824	1.2642	1.2480	0.6614	0.6132	0.6786
8	0.8777	0.7958	0.9119	1.3260	1.4557	1.3781	0.9272	0.8538	0.9174
9	0.8758	0.8304	0.8947	1.3269	1.4161	1.3662	0.9129	0.8486	0.8722
Sum	6.8717	**6.2636**	7.5890	**12.2580**	12.7880	12.5448	7.3902	**6.9196**	7.6669

Lowest values for specific sex and network type in bold. Scale: 10^{-4}

First of all, we observe that the RNN-based approaches show a better out-of-sample performance, since their MSE are sensibly lower compared to their SMM counterparts. In more detail, for females and both sexes combined it can be observed that the GRU networks tend to produce a lower error on average than LSTM networks. Overall, it seems that LSTM models produce better forecasts when they are two layered in the female case and one layered in the male case, whereas for GRU models this holds for two layered in the female case and three layered in the male case. Hence, we cannot draw a simple conclusion on which number of layers uniformly provides better forecasts for both sexes. However, it can be seen that for a specific sex and network type, the fitted values closest to those observed are generated by the same layered model through most groups.

When analysing socio-economic groups, we find that the out-of-sample performance is similar, except for few outliers. When looking at SMM, we note that their MSE is considerably larger when forecasting the mortality rates for group 4 under the Plat model, and the Li & Lee model only for males. In these cases, for some reason mortality is overestimated over all ages.

We conclude the section with two remarks: first, as noted through the analysis of their in-sample performance, NN models have the advantage to capture cohort effects, while other models would require the inclusion of specific cohort parameters. This may affect the precision of point forecasts. Second, a downside of NN models is that they produce only point forecasts. The strength of the stochastic models is their ability to provide a prediction interval which gives more reliability when forecasting mortality rates. To overcome this problem, an average over 40 forecasts has been taken in the NN case to prevent outliers in predictions. However, this solves the problem only in part.

6 Conclusions and outlook

The main contribution of the present case study can be divided in two major themes. On one hand, a NN approach has been implemented for the analysis of multiple populations based on socio-economic characteristics. On the other hand, an index of multiple deprivation was created that is linked to the life expectancy of the Italian population across different counties.

These overarching aims have been achieved by subdividing the underlying data of the Italian population into nine socio-economic groups. Then, various (multi-) population models have been estimated to some decades of Italian mortality data. The original hypothesis, following the work of Wen et al. [39], was to detect differences in mortality in the underlying deprivation groups. This effect was only partially observed in the Italian data using the methods we implemented. Whereas the two most deprived groups clearly exhibited higher mortality rates, the other groups could not be separated as accurately as it was possible for the British population.

Comparing the individual models, it turned out that all implemented approaches showed a very good fit to the data; both in-sample and out-of-sample. In general, the implemented NN models had a slightly more accurate fit compared to the classical statistical models. One reason could be the fact that the latter ones

do not include cohort effect parameters, which has been observable in residual plots, whereas NNs are able to capture cohort effects. As a downside, however, NN models produce only point forecasts and face robustness problems when used to extrapolate longer into the future. These robustness issues could be partly overcome by averaging several outcomes. Further steps may include the analysis of prediction intervals for the NN models as proposed in Pearce et al. [25] or in Mancini et al. [20], which was beyond the scope of this work. An advantage of the classical statistical models by Li & Lee, Kleinow, and Plat is the interpretability of their parameters, these followed reasonable patterns over time for each deprivation group.

Determining which model is best suited for a certain group can be done based on the explanation ratio, the AIC criterion, mean squared errors, error plots, or out-of-sample performance. Generally speaking, female mortality data was best described by the Li & Lee model, whereas male data was best fitted by Kleinow's model (based on explanation ratios and AIC). Looking at in-sample errors, NNs outperform the other models, presumably due to including cohort effects. Within the group of NN-models, three layered GRU models tend to produce the lowest in-sample errors. In general, old age mortality prediction was challenging and the models' precision went down for higher ages.

In forecasting via ARIMA, there was not one model approach uniformly outperforming the others. When forecasting via NN models, it appeared that the three layered GRU model does not necessarily outperform the others. Indeed, the two layered version performed better when forecasting, and in some cases even the one layered LSTM delivered really good predictions.

Concluding, the models of Li and Lee, Kleinow, Plat, and NNs of type LSTM and GRU all detected time trends in the underlying data, provided a satisfying fit for all deprivation groups and sexes, and allowed for forecasting with different methods; at least for short periods into the future.

Upcoming steps could include additional fine tuning, for example by adding cohort effects, trying different time series structures, adding further input features (the flexibility of RNNs allow for several features to be included, on top or in place of socioeconomic indicators), conducting more tuning of hyper-parameters, or adding further models for comparison, for example the Age-Period-Cohort-Improvement (APCI) models of Richards et al. [34] or CNN from Wüthrich and Meier [41]. As an outlook for applications of the developed tools and models, it would be interesting using these (with socioeconomic based feature) when modelling mortality for pension plans. This might allow a discussion about social (un-) fairness and private insurance companies could address and explain the problem of adverse selection (since mortality rates within insurance portfolios are normally smaller due to the selection effect). During the review process, we became aware of the recent work Scognamiglio [37] that proposes a NN approach for the creation of clusters of countries with similar mortality patterns. This might be an alternative to the clustering via the index we proposed.

A. Appendix: Additional details

A.1. Number of parameters

Note: For the Li and Lee, Kleinow, and Plat model, the number of parameters is 1204, 1100, and 1008, respectively, if one takes into account the parameters $\alpha(x, i)$. These are not included in the table because, unlike all other parameters, they are not part of the optimization process.

A.2. Cohort inspection 1916, 1917, and 1920

When analyzing in-sample errors for the models of Li and Lee, Kleinow, and Plat, we noted how poorly these fit the data for individuals born in 1916, 1917, and 1920. We give three possible explanations for this observation:

(i) Data could be flawed, as either death counts or exposure (or both) may have been miscounted;
(ii) Cohort effects may be present in the data. These cohorts may experience some fundamental differences from other cohorts in death or exposure counts, and hence in mortality;
(iii) Third, this may come from a demographic issue (which in part could be solved by cohort effects). Thus, investigation of exposures of several cohorts for the whole Italian population has been performed, under the assumption that this could be representative for each deprivation group since female-male-partition should be approximately the same through all groups which were chosen according to homogeneous population sizes. As explained in Michelozzi et al. [22], during the years of World War I there was a huge decline in birth rates. In 1914, when World War I started (Italy entered one year later), approximately 1,023,210 people were born. In 1919, the number of births nearly halved, having observed only 576,042. Such a huge decline in birth rates has not been observed in any other cohort included in our data. For example, in World War

Table 9 Number of parameters of each model

Model	Number of parameters
Li and Lee	790
Kleinow	686
Plat	594
LSTM1	2102
LSTM2	4257
LSTM3	5292
GRU1	1582
GRU2	3197
GRU3	3972

II, the decline in birth rates was a lot smaller: From 968,282 in 1939 births decreased to 757,251 resp. 866,081 in 1945/1946. This large decrease during and after World War I leads to smaller exposures for this cohort through all years and hence more inaccuracy in mortality rate fitting, since for older ages mortality behaviour is approximately the same as in other cohorts. In addition to World War I, the above cohorts also faced the threat of the Spanish Flu of 1918, which caused an enormous loss of life.

In conclusion, these reasons may emphasize the need for SMMs accounting for a cohort effect, especially for those which experienced a massive deviation in exposure size with respect to others, especially for the fast decline over few years. For example, today there are even less births than during and after World War I, but this decline has happened over years and decades in a slow process which seems to be easier captured by mortality models without cohort effects.

Acknowledgements Francesco Ungolo acknowledges financial support from the ERGO Center of Excellence in Insurance. We would like to thank the anonymous referees and the handling editor for very constructive comments that helped to improve our manuscript and to Mario Wüthrich for pointing us to some recent references that we were not aware of.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

1. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. Akademiai Kiado:267-281
2. Brouhns N, Denuit M, Vermunt J (2002) A Poisson Log–Bilinear regression approach to the construction of projected life tables. Insur Math Econ 31:373–393
3. Cairns AJG, Blake D, Dowd K (2006) A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. J Risk Insur 73(4):687–718. https://www.macs.hw.ac.uk/~andrewc/papers/jri2006e.pdf
4. Cairns AJ, Blake D, Dowd K, Coughlan G, Khalaf-Allah M (2011) Bayesian stochastic mortality modelling for two populations. ASTIN Bull 41(1):29–59
5. Cairns AJG, Blake D, Dowd K, Kessler AR (2016) Phantoms Never die: living with unreliable population data. J R Stat Soc Ser A Stat Soc 179(4):975–1005. https://doi.org/10.1111/rssa.12159
6. Cairns AJ, Kallestrup-Lamb M, Rosenskjold C, Blake D, Dowd K (2019) Modelling socio-economic differences in mortality using a new affluence index. ASTIN Bull 49(3):555–590
7. Cho K, van Merrienboer B, Gulcehre C, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. CoRR. arXiv:1406.1078
8. Chollet F, Allaire J, et al (2017) R interface to keras. https://github.com/rstudio/keras

9. Enchev V, Kleinow T, Cairns A (2015) Multi-population mortality models: fitting, forecasting and comparisons. Heriot-Watt University, Edinburgh. http://www.macs.hw.ac.uk/~andrewc/papers/Enchev2015.pdf
10. Euthum M (2021) Multi-population mortality models—a comparison via a socio-economic index of deprivation on italian population. In: Master thesis of Maximilian Euthum at the department of mathematics of Technical University of Munich (TUM) under the supervision of Prof. Dr. rer. nat. Matthias Scherer. Available at the mathematics department at TUM or by request to the author
11. Hainaut D (2018) A neural-network analyzer for mortality forecast. ASTIN Bull J IAA 48(2):481–508
12. Hsu D (2017) Time Series Forecasting Based on Augmented Long Short-Term Memory. https://arxiv.org/pdf/1707.00666v2.pdf
13. Human Mortality Database T (HMD)) Max Planck institute for demographic research and University of California, berkeley. https://www.mortality.org/
14. Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: the forecast package for R. J Stat Softw 27(3):1-22. https://www.jstatsoft.org/index.php/jss/article/view/v027i03
15. Johnson H, Kovats S, Mcgregor G, JR S, Gibbs M, Walton H (2005) The impact of the 2003 heat wave on daily mortality in England and Wales and the use of rapid weekly mortality estimates. Euro surveillance : Bulletin européen sur les maladies transmissibles Eur Commun Dis Bull 10:168–71
16. Kleinow T (2015) A common age effect model for the mortality of multiple populations. Insur Math Econ 63:147-152. https://www.sciencedirect.com/science/article/pii/S0167668715000554
17. Lee RD, Carter LR (1992) Modeling and forecasting US mortality. J Am Stat Assoc 87(419):659–671
18. Li N, Lee R (2005) Coherent mortality forecasts for a group of populations: an extension of the Lee-Carter mehtod. Demography 42:575–594. https://doi.org/10.1353/dem.2005.0021
19. Lindholm M, Palmborg L (2022) Efficient use of data for LSTM mortality forecasting. Eur Actu J 12:749–778
20. Mancini T, Calvo-Pardo H, Olmo J (2021) Prediction intervals for deep neural networks
21. Mattone A (2017) Estate 2003, la strage degli anziani noi abbiamo superato la Francia. https://ricerca.repubblica.it/repubblica/archivio/repubblica/2004/08/02/estate-2003-la-strage-degli-anziani-noi.html
22. Michelozzi P, De' Donato F, Scortichini M, Sario MD, Asta F, Agabiti N, Guerra R, de Martino A, Davoli M (2016) 'Sull'incremento della mortalità in Italia nel 2015: analisi della mortalità stagionale nelle 32 città del sistema di sorveglianza della mortalità giornaliera'. Epidemiol Prev. 40(1):22–8
23. Nigri A, Levantesi S, Marino M, Scognamiglio S, Perla F (2019) A deep learning integrated lee-Carter model. Risks 33(1). https://www.mdpi.com/2227-9091/7/1/33
24. Osservatorio della salute (2019) Indice di deprivazione nazionale e mortalita generale. https://osservatoriosullasalute.it
25. Pearce T, Zaki M, Brintrup A, Neely A (2018) High-quality prediction intervals for deep learning: a distribution-free, ensembled approach
26. Perla F, Scognamiglio S (2022) Locally-coherent multi-population mortality modelling via neural networks. Decis Econ Fin:1-20
27. Perla F, Richman R, Scognamiglio S, Wüthrich MV (2021) Time-series forecasting of mortality rates using deep learning. Scand Actu J 7:572–598. https://doi.org/10.1080/03461238.2020.1867232
28. Plat R (2009) On stochastic mortality modeling. Insur Math Econ. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1362487
29. Poulain M, Herm A, Pes G (2013) The blue zones: areas of exceptional longevity around the world. Vienna Yearb Popul Res 11:87–108. http://www.jstor.org/stable/43050798
30. R Core Team (2013) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
31. Renshaw A, Haberman S (2006) A cohort-based extension to the Lee-Carter model for mortality reduction factors. Insur Math Econ 38:556–570
32. Richman R, Wüthrich M (2018) A neural network extension of the Lee-Carter model to multiple populations. https://ssrn.com/abstract=3270877
33. Richman R, Wüthrich M (2019) Lee and Carter go machine learning: Recurrent neural networks. https://ssrn.com/abstract=3441030

34. Richards SJ, Currie ID, Kleinow T, Ritchie GP (2019) A stochastic implementation of the APCI model formortality projections. Br Act J 24(13). https://www.researchgate.net/publication/332040614_A_stochastic_implementation_of_the_APCI_model_for_mortality_projections
35. Schnürch S, Korn R (2022) Point and interval forecasts of death rates using neural networks. ASTIN Bull 52(1):333–360
36. Scognamiglio S (2022) Calibrating the Lee–Carter and the poisson Lee-Carter models via neural networks. ASTIN Bull 52(2):519–561
37. Scognamiglio S (2022) A multi-population locally-coherent mortality model. In: Corazza M, Perna C, Pizzi C, Sibillo M (eds) Mathematical and statistical methods for actuarial sciences and finance. Springer International Publishing, Cham, pp 423–428
38. Wang C-W, Zhang J, Zhu W (2021) Neighbouring prediction for mortality. ASTIN Bull 51(3):689–718
39. Wen J, Cairns A, Kleinow T (2020) Fitting multi-population mortality models to socio-economic groups. Ann Actu Sci
40. Wilson C (2001) On the scale of global demographic convergence 1950–2000. Popul Dev Rev 27(1):155–171
41. Wüthrich MV, Meier D (2020) Convolutional neural network case studies: (1) anomalies in mortality rates (2) image recognition. In: Actuarial Data Science, Swiss Association of Actuaries. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3656210