POINT AND INTERVAL FORECASTS OF DEATH RATES USING NEURAL NETWORKS

BY

SIMON SCHNÜRCH AND RALF KORN

ABSTRACT

The Lee–Carter model has become a benchmark in stochastic mortality modeling. However, its forecasting performance can be significantly improved upon by modern machine learning techniques. We propose a convolutional neural network (NN) architecture for mortality rate forecasting, empirically compare this model as well as other NN models to the Lee–Carter model and find that lower forecast errors are achievable for many countries in the Human Mortality Database. We provide details on the errors and forecasts of our model to make it more understandable and, thus, more trustworthy. As NN by default only yield point estimates, previous works applying them to mortality modeling have not investigated prediction uncertainty. We address this gap in the literature by implementing a bootstrapping-based technique and demonstrate that it yields highly reliable prediction intervals for our NN model.

KEYWORDS

Mortality forecasting, neural networks, convolutional neural networks, uncertainty quantification, prediction intervals, Lee–Carter model, mortality of multiple populations.

JEL codes: J11, C45, C53, G22

1. INTRODUCTION

Lee and Carter (1992) propose a seminal stochastic mortality model, the Lee–Carter (LC) model, in which they decompose logarithmic death rates into an age-specific base level and a time-varying component (period effect) multiplied by an age-modulating parameter (age effect). Since then, many other stochastic mortality models have been introduced (Cairns *et al*. 2009). While, initially,

Astin Bulletin 52(1), 333–360. doi:10.1017/asb.2021.34

they were used to describe one population, there are situations in which it is useful or even necessary to model the mortality of multiple populations simultaneously. To this end, multi-population models such as the augmented common factor (ACF) model by Li and Lee (2005) have been proposed.

The structure of the LC model and many of its descendants is simple. They are easy to understand and implement but have the potential drawback of not being optimal with respect to forecasting performance, which is one of the main requirements for a mortality model. In fact, there is a broad consensus in the literature that LC-type models work well for some but certainly not for all mortality data. It seems promising to apply more sophisticated methods such as machine learning to mortality forecasting, which might, for example, be able to handle nonlinearities better than existing models. Here, we focus on neural networks (NN), in particular feed-forward neural networks (FFNN), recurrent neural networks (RNN) and convolutional neural networks (CNN).

FFNN have been applied to mortality forecasting by Shah and Guez (2009) and more recently Richman and Wüthrich (2021), who provide a review of existing stochastic multi-population mortality models and point out some of their drawbacks: Sometimes, they are difficult to calibrate, and some model structures are hard to justify and lack theoretical foundations. Thus, they propose to refrain from making any structural assumptions at all on mortality development and fully rely on an NN to learn mortality intensities from historical data. To this end, they train an FFNN and find that it outperforms the LC model and other stochastic models in an out-of-sample test on a data set comprised of 41 countries.

Nigri *et al*. (2019) base their approach on the LC model. Instead of the standard random walk or a general autoregressive integrated moving average (ARIMA) process, they use a certain type of RNN called long short-term memory (LSTM) network for projecting the period effects. They find that this leads to more accurate forecasts for several populations. Richman and Wüthrich (2019) provide an in-depth explanation on how to model and forecast death rates directly using LSTM or other RNN architectures. They evaluate their approach on Swiss data and find that their NN outperforms the LC model, but it is not very stable over different runs of the training algorithm. Therefore, they advocate the use of network ensembles to reduce the variance in the forecast under different random seeds.

Until very recently, CNN have not been considered in the mortality forecasting literature. Originally introduced by LeCun *et al*. (1989) for image recognition, they have become a key technology for modern computer vision. There are some possible advantages for convolution-based architectures over FFNN and RNN. CNN are designed to leverage local spatial relationships in the input data. For images, this means that each neuron only handles a small part of the whole image, whereas for mortality data one can, for example, think of neurons which are only activated by higher-age mortality and others which focus on lower-age mortality. Furthermore, the convolution parameters are shared, which can make CNN more parsimonious and their parameter

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

estimates more stable. Motivated by this, we apply two-dimensional CNN to mortality forecasting.

There have been similar efforts in parallel to our work. Perla *et al*. (2021) investigate the use of RNN and one-dimensional CNN for mortality forecasting and show that their NN, which can be interpreted as nonlinear extensions of the classical LC model, work very well on real data. They use one-dimensional convolutions, which means that the model performs the convolution operations only in the time and not in the age dimension. To capture the specific input structure of this application, in particular the correlation structure in the age dimension and interactions along the age–year plane (e.g., cohort effects), we believe that two-dimensional convolutions might be more appropriate. This is in line with the approach of Meier and Wüthrich (2020), who use two-dimensional CNN for detecting anomalies in mortality data. Wang *et al*. (2021) consider CNN with two-dimensional convolutions as well and show that they produce more accurate one-step point forecasts than classical stochastic mortality models.

Despite their typically stronger predictive performance, practitioners do not always prefer NN because they are hard to interpret. The ability to understand why a model makes a certain prediction or at least the conviction that the model creates a meaningful representation of the features is sometimes considered more important than the forecasting performance of the model. We will make some theoretical and practical efforts towards better explainability of our NN models. Even if a fully interpretable model is required for a particular application, it is still worthwhile to consider an NN as a benchmark, for example, to find out where the interpretable model could be improved in terms of forecasting performance.

In many applications, it is not only necessary to make accurate mortality forecasts but also useful or even required to quantify the uncertainty related to these forecasts. For example, insurance companies have to build up reserves based on risk measures such as the value-at-risk in order to be prepared for extreme mortality events. One of the main motivations for stochastic mortality models such as the LC model is that they provide estimates for dispersion measures and quantiles. It is a material shortcoming of all the NN approaches to mortality modeling described above that their output exclusively consists in point estimates. Therefore, we are also interested in a suitable method for uncertainty quantification and prediction interval estimation for our NN models. For this, we consider a bootstrapping-based technique. We show in an empirical study that it produces reliable and informative prediction interval estimates for the CNN, whereas the intervals obtained from the standard LC approach fail to contain the target values as often as required.

In total, our main contributions to the literature consist in

- proposing to train a CNN on the age–period mortality surface,
- comparing its forecasts to four benchmarks from the literature: two other types of NN (FFNN, RNN), the ACF model and the LC model,

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

- presenting and applying a bootstrapping approach for quantifying the uncertainty of NN forecasts.

The remainder of this article is structured as follows. In Section 2, we describe the NN methodology and architectures with a focus on the CNN. In Section 3, we explain how prediction uncertainty is quantified and prediction intervals are obtained. In Section 4, we perform an empirical comparison with respect to goodness-of-fit, forecasting performance and uncertainty quantification. Section 5 concludes.

2. Neural network models

We assume that a data set of death rates $m^i_{x,t}$ with ages $x \in \mathcal{X} := \{x_1, \dots, x_A\}$, populations $i \in \mathcal{P} := \{1, \dots, P\}$ and years $t \in \mathcal{T} := \{t_1, \dots, t_Y\}$ is given. Data availability by year usually depends on the population. We ignore this in our notation for simplicity. In our numerical experiments, we train the NN models on the whole available age range $x \in \mathcal{X} = \mathcal{X}_{\text{in}} := \{0, \dots, 100\}$ to make use of all data during training, and we usually evaluate them on the ages $x \in \mathcal{X}_{\text{out}} := \{60, \dots, 89\}$, which are most relevant for annuity payments and therefore often considered in actuarial mortality forecasting applications.

2.1. Convolutional neural networks

A two-dimensional CNN can pick up spatial relationships in the data which other models might not be able to exploit. Intermediate data representations calculated by this network topology are based on death rates adjacent to each other both in the time and age dimension. Thereby, the model can incorporate the correlation structure of mortality rates along the age dimension as well as a variety of age–year interaction effects. These are generally referred to as neighborhood effects by Wang *et al*. (2021), who provide some further motivation for the approach. Similar ideas have been investigated in the mortality modeling literature based on classical time series analysis, for example, by Denton *et al*. (2005), who find that correlations between the residuals of their ARIMA mortality models for adjacent age groups tend to be high. They propose a block bootstrap method for generating long-term mortality forecasts, using age–year matrices as inputs, an approach which aims to preserve the age correlation structure of mortality rate changes. A simple example for age–year interactions are cohort effects, which depend on the year of birth and are often present in mortality data (Renshaw and Haberman 2006).

The convolution operation is equivariant to shifts in the input data, which essentially means that similar patterns of mortality observed over different ages or at different points in time should lead to similar outputs. This could allow the model to learn general patterns of mortality development even though they might not appear exactly in the same age range or at the same time across

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

different populations. For more details on CNN, we refer to (Goodfellow *et al*. 2016, Chapter 9) and Section A.3 of the Online Supplementary Material.

We begin by fixing $\tau \in \mathbb{N}$, the maximal length of a historical time window influencing predictions. There is a tradeoff: larger values allow for a longer history of observations to be used for prediction, taking into account the possibility of longer serial correlations in mortality data, but they also reduce the number of available training data and increase the number of network parameters. For our numerical studies, we follow Perla *et al*. (2021) and set $\tau = 10$. We have done some experiments with $\tau = 20$, for which we observed a decline in forecasting accuracy. Then, for all populations $i \in \mathcal{P}$, we arrange the death rate data in age–time matrices via the rolling window approach

$$\left(m^i_{x,t}\right)_{x\in\mathcal{X}_{\text{in}},t=t_1,\ldots,t_\tau}, \ldots, \left(m^i_{x,t}\right)_{x\in\mathcal{X}_{\text{in}},t=t_{Y-\tau},\ldots,t_{Y-1}}. \tag{2.1}$$

These matrices are standardized element-wise over the training data set and then passed as inputs to the CNN. The corresponding outputs which the net is trained to forecast are, respectively, given by

$$\left(m^i_{x,t_{\tau+1}}\right)_{x\in\mathcal{X}_{\text{out}}}, \ldots, \left(m^i_{x,t_Y}\right)_{x\in\mathcal{X}_{\text{out}}}.$$

In other words, the network is trained on the death rates of the past τ years (matrices of dimension $A_{\text{in}} \times \tau$) to give a prediction for the next year (a vector of dimension A_{out}). We obtain forecasts for multiple years ahead by recursive 1-year predictions, using the forecast for year 1 as an input for the forecast of year 2, and so on, an approach which is well-established in time series analysis and has, for example, also been applied by Perla *et al*. (2021).

An example for the input data of the CNN is shown in Figure 1. In the demographic literature, such a two-dimensional arrangement is called a Lexis diagram. It is a classical method of visualizing mortality dynamics (see Pitacco *et al*. 2008, p. 94). We could additionally include country and gender information via embedding layers (Guo and Berkhahn 2016). However, the network might achieve good forecasts solely based on historical death rates as inputs. This could lead to more stable forecasts for populations for which just a small amount of training data or only test data is available. Therefore, we do not provide the network with such information.

CNN contain three different types of layers. Convolutional layers create several representations of their inputs where different properties characterizing these inputs are amplified. Pooling layers are then used for reducing the redundancy introduced by these multiple representations of the same input by extracting the most dominant output signals of the convolutional layers. Dense layers, which are basic feed-forward layers also used in FFNN, often follow after some repetitions of convolutional and pooling layers. Therefore, we can interpret the convolutional and pooling layers of a CNN as sophisticated feature extractors, whose outputs are passed on to an FFNN which learns how to translate these special features into mortality rate predictions. Figure 2 shows a schematic illustration of a CNN with four layers.

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

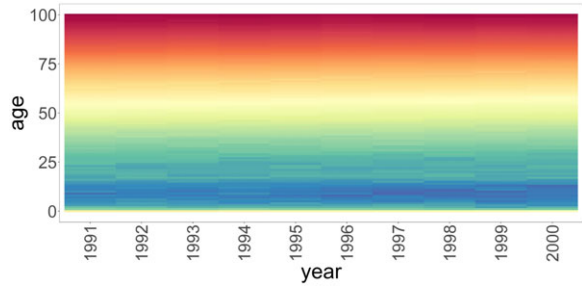

FIGURE 1: Illustration of an input matrix for the CNN with $x_1 = 0$, $x_A = 100$, $\tau = 10$, as a heat map displaying the log-transformed death rates of English and Welsh females. The colors range from blue (low death rates) to red (high death rates).

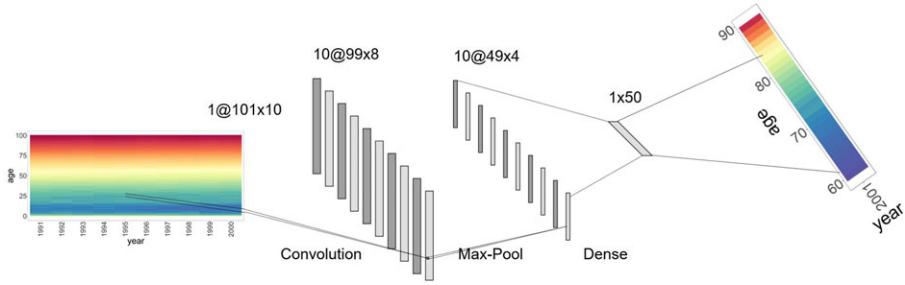

FIGURE 2: CNN with input size (1,101,10) consisting of a convolutional layer of size (10,99,8), a pooling layer of size (10,49,4), a dense layer of size 50 and a dense output layer of size 30. Figure produced using the tool by LeNail (2019).

We can interpret our CNN in terms of an LC-type modeling approach along the lines of Perla *et al*. (2021). Taking into account our hyperparameter choices (linear activation for the output layer), for fixed $x \in \mathcal{X}_{\text{out}}$, $t \in \mathcal{T}$, $i \in \mathcal{P}$ the model reads

$$\log m^i_{x,t} = b_x + \left\langle \left(\mathbf{W}_{x,j}\right)_{j=1,\dots,k}, \mathbf{Z}^i_t \right\rangle, \tag{2.2}$$

where $\mathbf{Z}^i_t \in \mathbb{R}^k$ is a nonlinear function of $\left(m^i_{\tilde{x},\tilde{t}}\right)_{\tilde{x}\in\mathcal{X}_{\text{in}},\, \tilde{t}=t-\tau,\dots,t-1}$. In this sense, our model generalizes the common age effect model considered by Wen *et al*. (2021),

$$\log m^i_{x,t} = \alpha_x + \sum_{j=1}^{k} \beta^j_x \kappa^{i,j}_t, \tag{2.3}$$

where both the base mortality level α_x (corresponding to the CNN bias term b_x) and the age effects $\beta^1_x, \dots, \beta^k_x$ (corresponding to a row in the CNN weight matrix $\mathbf{W}$) are assumed to be identical over all considered populations. However, we use a sequence of convolutional, pooling and dense layers,

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

which can capture complex, nonlinear interaction effects across the age and time dimension of the death rates, to create a powerful generalization $\mathbf{Z}_t^i$ of the period effects $\kappa_t^{i,1}, \ldots, \kappa_t^{i,k}$.

2.2. Hyperparameter selection and model training

As usual, we perform a hyperparameter tuning before starting the training of the NN. We initially divide the available data into a training set (containing all available years up to 2006) and a test set (containing the years from 2007 to 2016). The latter is exclusively used for a final evaluation and comparison of the chosen models in Section 4. On the training set, we perform threefold cross-validation and choose the hyperparameter configuration which minimizes the obtained cross-validation mean-squared error. Compared to the use of a single validation set at the end of the training data, this approach ignores the existing temporal dependence structure of the data to some extent. However, it has been found that this theoretical shortcoming of applying random cross-validation for time-dependent data usually does not have significant practical consequences for sufficiently large and flexible models (Bergmeir *et al*. 2018). On the contrary, it makes better use of the available data and can therefore lead to a more robust model selection.

We have evaluated over 8500 hyperparameter combinations by threefold cross-validation. The chosen network architecture is similar to the one depicted in Figure 2 except for an additional sequence of a convolutional and a pooling layer before the dense layers. Details can be found in Section A.3 of the Online Supplementary Material. Note that there are more sophisticated hyperparameter selection strategies (Goodfellow *et al*. 2016, Chapter 11), which could yield further improvements in performance.

As the weights of an NN are initialized randomly and stochastic gradient descent is used for optimization, training the same network multiple times yields different parameters and predictions. It is a popular approach to train multiple networks and average their outputs to obtain more robust forecasts (see, for example, Richman and Wüthrich 2020), and we have found this to improve performance for our application as well. Therefore, once we have fixed the values for the hyperparameters, we train such a model ensemble of 1000 CNN. To allow for an additional source of randomness and make the model more robust with respect to the choice of training data, each CNN is trained on a bootstrap sample of the original training data. This approach has been proposed by Breiman (1996), where it is termed a bagging (**b**ootstrapping and **agg**regating) ensemble.

The ordinary bootstrapping approach does not explicitly account for dependence structures in the covariates. However, as can be seen from (2.1), the input data for our CNN model are $A_{\text{in}} \times \tau$-sized blocks of mortality rates. In this sense, the bootstrapping procedure preserves the age–time dependence structure. The dependence between populations is ignored here because we have

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

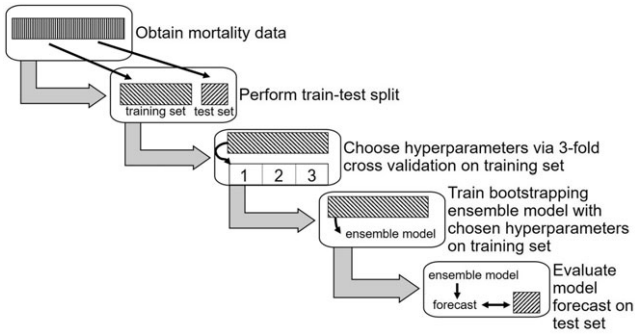

FIGURE 3: The process for choosing hyperparameters, training and evaluating models.

empirically found the model to achieve better performance when the population does not explicitly enter as a feature. Improvements might be possible by using modifications of the bootstrap procedure such as stratified or Sieve bootstrap (D'Amato *et al*. 2011; 2012). We leave a deeper investigation of this matter and a potential further improvement of the models in this direction for future research.

Figure 3 summarizes the process we follow from data acquisition up to the test set evaluation of the final model.

2.3. Feed-forward and recurrent neural networks

We consider two other NN architectures from the literature as benchmarks for our CNN model. FFNN for mortality forecasting have been investigated by Richman and Wüthrich (2021). We adopt their method and calibrate an FFNN on the year as a continuous input and on age, country and gender as categorical inputs, which are fed to the network via embedding layers. RNN are particularly suited for modeling and forecasting sequential data. Here, we adapt the LSTM mortality forecasting approach of Richman and Wüthrich (2019) to multiple populations.

For both FFNN and LSTM, hyperparameters are chosen based on three-fold cross-validation. Afterwards, bagging ensembles of 100 FFNN and 10 LSTM, respectively, are trained. For further details, we refer to Sections A.1 and A.2 of the Online Supplementary Material and to textbooks such as Denuit *et al*. (2019).

3. PREDICTION UNCERTAINTY

Uncertainty in forecasting $m^i_{x,t}$ can be quantified by calculating a lower bound $\hat{m}^{i,\,\text{lower}}_{x,t}$ and an upper bound $\hat{m}^{i,\,\text{upper}}_{x,t}$ such that

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

$$\mathbb{P}\left(\hat{m}_{x,t}^{i,\,\text{lower}} \le m_{x,t}^{i} \le \hat{m}_{x,t}^{i,\,\text{upper}}\right) \ge a \tag{3.1}$$

for some given threshold $a \in (0,1)$. The interval $[\hat{m}_{x,t}^{i,\,\text{lower}}, \hat{m}_{x,t}^{i,\,\text{upper}}]$ is called a *prediction interval* for $m_{x,t}^{i}$ at level a. Prediction interval calculation for stochastic mortality models has been extensively dealt with in the literature. We present details on how we obtain them in Section B of the Online Supplementary Material.

For calculating prediction intervals in our NN models, we rule out several existing methods from the literature because they would require substantial structural changes, such as the insertion of dropout layers for Monte Carlo dropout (Gal and Ghahramani 2016), a change of the loss function for lower upper bound estimation (Khosravi *et al* 2011b) or a substantial increase of the output dimension for mean–variance estimation (Nix and Weigend 1994). We prefer not to change these hyperparameters because doing so might decrease the forecasting performance, the optimization of which is still our main goal. Khosravi *et al*. (2011a) provide a survey of multiple approaches for calculating prediction intervals. They refer to bootstrapping as the most commonly used technique and find it to achieve the largest variability in prediction interval width compared to other methods. This shows it is able to respond to differing levels of uncertainty in the data, an important quality for mortality forecasting applications. Therefore, we adapt the bootstrapping approach of Heskes (1997) to our setup, for which no change of our existing model structure is necessary.

We assume the process generating logarithmic death rates is given by

$$y(z) = f(z) + \varepsilon(z), \tag{3.2}$$

where $\varepsilon(z)$ is zero-mean noise, f is the true, unobservable input–output relationship and $y(z) := \log m_{x,t}^{i}$ is the noisy observation of $f(z)$. Note that we consider logarithmic death rates here in order to make the normal assumption we are going to use at a later point more appropriate. The input z depends on the modeling framework, for example, we consider $z := (x, t, i)$ with age x, year t and population i for FFNN, while we use previous death rate observations as inputs for RNN and CNN as well.

NN aim to learn an estimator $\hat{f}(z)$ of the true value $f(z)$ based on some training data, which typically do not contain z. The natural question arises how well $\hat{f}(z)$ predicts the actual target value, the realization of the random variable $y(z)$. Under the assumption that $f(z) - \hat{f}(z)$ and $\varepsilon(z)$ are uncorrelated, this is addressed by the well-known bias-variance decomposition

$$\mathbb{E}\left(\left(y(z) - \hat{f}(z)\right)^2\right) = \text{Bias}\left(\hat{f}(z)\right)^2 + \text{Var}\left(\hat{f}(z)\right) + \text{Var}\left(\varepsilon(z)\right). \tag{3.3}$$

We call $\sigma^2(z) := \text{Var}\left(\hat{f}(z)\right)$ model uncertainty and $\sigma_\varepsilon^2(z) := \text{Var}\left(\varepsilon(z)\right)$ noise variance. They sum up to the total variance $s^2(z) := \text{Var}\left(y(z) - \hat{f}(z)\right)$.

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

As described in Section 2, we train NN $\hat{f}_m$, $m = 1, \ldots, N_E$, on bootstrap samples of the training data and obtain the ensemble estimator by averaging,

$$\hat{f}(z) := \frac{1}{N_E} \sum_{m=1}^{N_E} \hat{f}_m(z). \tag{3.4}$$

From this, assuming Bias $\left(\hat{f}_m(z)\right) \approx 0$ for all $m = 1, \ldots, N_E$, we estimate model uncertainty $\sigma^2(z)$ for FFNN and for one-step forecasts of RNN and CNN by the ensemble variance

$$\hat{\sigma}^2(z) := \frac{1}{N_E - 1} \sum_{m=1}^{N_E} \left(\hat{f}_m(z) - \hat{f}(z)\right)^2. \tag{3.5}$$

This mainly accounts for variance arising from the randomness in the initialization and calibration of the NN (by considering an ensemble of multiple such networks) and for the uncertainty of the model parameters with respect to the training data (by bootstrapping).

As stated in Section 2.1, we perform recursive one-step predictions to obtain multi-step forecasts for RNN and CNN. Directly applying the one-step formula (3.5), a natural way to estimate model uncertainty σ_h^2 in a recursive h-step forecast for fixed $h \in \{1, 2, \ldots\}$ would be via

$$\left(\hat{\sigma}_h^{\mathrm{E}}\right)^2 := \frac{1}{N_E - 1} \sum_{m=1}^{N_E} \left(\hat{f}_m\left(z_h\right) - \frac{1}{N_E} \sum_{p=1}^{N_E} \hat{f}_p\left(z_h\right)\right)^2. \tag{3.6}$$

The input of the one-step forecast, $z_1 \in \mathbb{R}^{A_{\mathrm{in}} \times \tau}$, consists entirely of data available at the beginning of the forecasting period. For $h \geq 2$, we drop the first column of z_{h-1}, denote the resulting matrix by ${}_{-1}z_{h-1} \in \mathbb{R}^{A_{\mathrm{in}} \times (\tau - 1)}$ and then set $z_h := \left({}_{-1}z_{h-1}, \hat{f}(z_{h-1})\right) \in \mathbb{R}^{A_{\mathrm{in}} \times \tau}$. This means the input z_h for the h-step forecast depends on the previous forecasts $\hat{f}(z_1), \ldots, \hat{f}(z_{h-1})$, which are themselves subject to uncertainty. Therefore, it would be plausible for model uncertainty to increase with h, mirroring the subjective belief that mortality rates further in the future are more uncertain.

To achieve this, we consider the following heuristic modification of (3.6):

$$\left(\hat{\sigma}_h^{\mathrm{P}}\right)^2 := \frac{1}{N_E - 1} \sum_{m=1}^{N_E} \left(\hat{f}_m\left(z_{h,m}\right) - \frac{1}{N_E} \sum_{p=1}^{N_E} \hat{f}_p\left(z_{h,p}\right)\right)^2. \tag{3.7}$$

The input of the one-step forecast, $z_{1,m} := z_1$, is the same for all $m = 1, \ldots, N_E$. For $h \geq 2$, using analogous notation as above, we set $z_{h,m} := \left({}_{-1}z_{h-1,m}, \hat{f}_m(z_{h-1,m})\right)$. This means that we recursively supply to each ensemble member only its *own* past forecasts instead of the averaged forecasts of *all*

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

ensemble members, which we normally use as the forecast of our entire model and which also appears in (3.6). We have found in our numerical studies that this approach indeed works as $\left(\hat{\sigma}_h^{\mathrm{P}}\right)^2$ tends to increase with h (see Section 4.3). Therefore, we use $\left(\hat{\sigma}_h^{\mathrm{P}}\right)^2$ to estimate model uncertainty.

We estimate noise variance σ_ε^2 via an additional FFNN with exponential output activation which is fit to the floored residuals

$$r^2(\xi) := \left(\left(y(\xi) - \hat{f}(\xi)\right)^2 - \hat{\sigma}^2(\xi)\right)^+ \tag{3.8}$$

by maximum likelihood, that is, minimizing

$$L := \frac{1}{2}\sum_{j=1}^{N}\left(\log\left(\hat{\sigma}_\varepsilon^2(\xi_j)\right) + \frac{r^2(\xi_j)}{\hat{\sigma}_\varepsilon^2(\xi_j)}\right) \tag{3.9}$$

over all available training data. Apart from that, we use the hyperparameters as described in Section A.1 of the Online Supplementary Material. We have also experimented with RNN and CNN for noise variance prediction but found them to be less numerically stable and yield less plausible uncertainty estimates than an FFNN.

Finally, once model uncertainty and noise variance have been estimated, we set $\hat{s}^2(z) := \hat{\sigma}^2(z) + \hat{\sigma}_\varepsilon^2(z)$ and obtain prediction interval bounds under a normal assumption (cf. Carney *et al*. 1999) by

$$\hat{y}^{\mathrm{lower|upper}}(z) := \hat{f}(z) \pm \Phi^{-1}\left(\frac{1+a}{2}\right)\hat{s}(z). \tag{3.10}$$

These bounds for the logarithmic death rates are easily transformed to yield prediction intervals for the death rates themselves.

4. Empirical model comparison

For the numerical studies in this section, we use death rates from the Human Mortality Database (2019, HMD) with zeros replaced by a small, positive number because we often work with logarithmic death rates. We calibrate the three NN models on the death rates $m_{x,t}^i$ for ages $x = 0, \ldots, 100$, almost all available populations and all the years $t \leq 2006$ for which data are available. The reason why we use all available years is that NN generally need many training data.

The Poisson LC model proposed by Brouhns *et al*. (2002) is trained on the death rates of the ages $x = 60, \ldots, 89$, the 54 populations for which there are data available for every year between 1987 and 2016 and on 10 (LC10) or 20 (LC20) years of data. One might argue that these calibration periods are too short for a fair comparison. However, the LC model does not necessarily benefit from more training data because the longer the calibration period, the higher the risk that its underlying assumptions are violated (see Booth *et al*. 2002). In

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

fact, we will see that the LC10 model produces slightly lower out-of-sample errors than the LC20 model, and we have found that it even achieves slightly superior out-of-sample performance compared to an LC model calibrated on 30 years of data. Therefore, and in order to have as many populations as possible available for performance evaluation, we restrict ourselves to a maximum calibration period of 20 years for the LC model. As an additional classical benchmark model, we consider the ACF model (Li and Lee 2005), using 20 years of data for calibration and allocating countries to regions as proposed by (Richman and Wüthrich 2021, Appendix A).

With each model, we produce forecasts $\hat{m}^i_{x,t}$ for ages $x = 60, \dots, 89$, years $t = 1997, \dots, 2006$ to evaluate the in-sample errors and years $t = 2007, \dots, 2016$ to evaluate the out-of-sample errors by comparing the forecasts with the corresponding observations $m^i_{x,t}$. The years for the in-sample evaluation are chosen as the intersection of the training sets of all models. We calculate the error measures

- mean squared error $\text{MSE} = \frac{1}{N} \sum_{x,t,i} \left(\hat{m}^i_{x,t} - m^i_{x,t}\right)^2$,
- mean absolute error $\text{MAE} = \frac{1}{N} \sum_{x,t,i} \left|\hat{m}^i_{x,t} - m^i_{x,t}\right|$,
- median absolute percentage error $\text{MdAPE} = \text{median}_{x,t,i} \left\{ \frac{|\hat{m}^i_{x,t} - m^i_{x,t}|}{m^i_{x,t}} \right\}$,
- mean Poisson deviance $\text{Dev} = \frac{2}{N} \sum_{x,t,i} D^i_{x,t} \left(\log \frac{m^i_{x,t}}{\hat{m}^i_{x,t}} + \frac{\hat{m}^i_{x,t}}{m^i_{x,t}} - 1 \right)$,

where ages x, years t and populations i range over all observations considered for evaluation and N denotes the number of these observations. The *MAE* and particularly the *MSE* penalize forecasting errors more when the target death rate is higher. They are strongly influenced by errors in forecasting high-age mortality because $m^i_{x,t}$ usually increases with age. To prevent this, weighted versions of these measures could be considered. Alternatively, evaluations could focus on the *MdAPE* as a relative measure. We also include mean Poisson deviance as an error measure, which assigns more weight to errors in larger populations or age groups as it depends on the death counts $D^i_{x,t}$.

For implementing the networks, we use the R interface to the package Keras (R Core Team 2019; Falbel *et al*. 2019). For producing figures, we use the data visualization package ggplot2 by Wickham (2016).

4.1. Goodness-of-fit

Table 1 contains the in-sample error measures. The LC10 model achieves the highest goodness-of-fit, which is no surprise since it is evaluated exactly on the years it was calibrated on, while all the other models were trained on more data. Intuitively, they will not fit as well on subsets of their training data as a model that was trained specifically on that subset, but they might be able

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

TABLE 1

GOODNESS-OF-FIT MEASURES FOR 54 POPULATIONS, AGES 60–89, YEARS 1997–2006 (MODELS TRAINED ON YEARS UP TO 2006). THE BEST VALUE IN EACH COLUMN IS MARKED IN BOLD.

Model	MSE$\times 10^5$	MAE$\times 10^3$	MdAPE (%)	Dev
LC10	**1.9**	**2.0**	**2.0**	**1.8**
LC20	2.3	2.3	2.4	2.4
ACF	2.1	2.2	2.4	2.3
FFNN	2.8	3.0	4.3	11.7
RNN	6.0	4.5	6.0	19.1
CNN	3.7	3.4	4.3	12.3

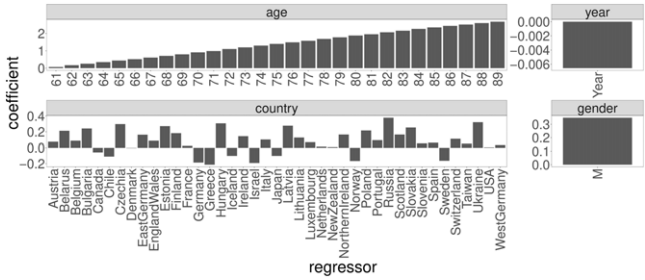

FIGURE 4: Coefficients of a log-linear global surrogate model for the CNN.

to generalize better. We will see in the following subsection that the good in-sample performance of the LC models contrasts with their weaker predictive performance, indicating some degree of overfitting for these models.

In order to check whether our CNN is an appropriate model despite its comparably high in-sample errors, we consider a log-linear global surrogate model. The basic idea lies in fitting a simple, interpretable model to the fitted values of a complex model in order to better understand what the complex model is doing by interpreting the simple model. More precisely, we calibrate a linear model to the fitted logarithmic death rates of the CNN using age, gender and country as categorical regressors and year as a numerical regressor. Its coefficients are displayed in Figure 4. They suggest that the network has learned an overall plausible internal representation of the training data.

We observe, as expected, a log-linear pattern for the age dependency of mortality forecasts, a decrease of predicted mortality rates with time, a tendency to predict higher death rates for males than for females and country-specific adjustments with respect to the reference country (Australia) which look mostly sensible. Some countries with long mortality history which have experienced relatively high mortality rates in earlier times such as France, Italy or Switzerland are assigned somewhat high country-specific coefficients compared to their current mortality levels. This is not problematic for forecasts with the CNN because it only receives current mortality levels as input and

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

TABLE 2

OUT-OF-SAMPLE ERROR MEASURES FOR 54 POPULATIONS, AGES 60–89, YEARS 2007–2016 (MODELS TRAINED ON YEARS UP TO 2006). THE BEST VALUE IN EACH COLUMN IS MARKED IN BOLD.

Model	MSE $\times 10^5$	MAE $\times 10^3$	MdAPE (%)	Dev	% of pop. with lower MSE than LC10	% of pop. with lower MdAPE than LC10
LC10	4.9	3.7	5.7	**19.5**	0.0	0.0
LC20	5.5	4.0	5.8	21.9	51.9	50.0
ACF	3.4	3.3	5.5	19.7	61.1	61.1
FFNN	**2.6**	3.0	5.9	32.5	64.8	53.7
RNN	5.9	4.2	6.1	22.9	46.3	46.3
CNN	2.9	**3.0**	**4.9**	30.7	**79.6**	**75.9**

no other information on the country. However, it shows that the information obtained from a surrogate model depends on the data it is calibrated on – for example, one could calibrate another surrogate model only on a subset of fitted CNN predictions for years after 1970 or 1980 to get an impression of its behavior on more recent data.

We have checked how well the global surrogate model works by calculating its coefficient of determination $R^2 = 0.9698$. This shows that a simple regression model describes the in-sample predictions of the CNN well and the insights obtained from Figure 4 should be reasonably reliable. With respect to out-of-sample forecasts of death rates, we have found the surrogate model to perform very poorly. This is not surprising due to its limited model structure. We emphasize that surrogate models are *not* meant to describe the data and their future development but to make the global in-sample behavior of black-box models more interpretable. In this regard, it is also important to observe that the surrogate model is trained on different input features (age, country, gender, year) than the CNN (matrices of death rates) for better interpretability.

4.2. Forecasting performance

Table 2 contains out-of-sample error measures. The ACF and LC models achieve low Poisson deviances. FFNN and CNN have larger Poisson deviances because their forecasting performance is suboptimal for the USA and Japan, respectively, which are very populous countries and therefore heavily influence this measure. Except for Poisson deviance, both LC models and the RNN perform similarly and noticeably worse than the FFNN (except for the MdAPE) and the CNN. The rather high errors of the RNN might be explained by the fact that it is the hardest to train as it takes a lot of computational resources, which is why we can only build a small ensemble consisting of 10 models with relatively few neurons per model. The ACF model yields better results than

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

the LC model but is still outperformed by the CNN. The FFNN achieves good performance with respect to absolute measures such as MSE and MAE but has a rather high MdAPE, which will be investigated in Figure 5 below. The CNN minimizes MAE and MdAPE. To get a better impression of the performance for different populations, we have evaluated the percentage of populations for which each model achieves a lower MSE (MdAPE) than the LC10. Here, the CNN performs best as it produces lower MSE (MdAPE) than the LC10 for 43 (41) out of 54 populations, which corresponds to 79.6% (75.9%).

Figure 5 shows the MdAPE of the 10-year forecast of the CNN by age, year and population compared to the FFNN, RNN, ACF and LC20 models. This is a more detailed evaluation of the corresponding column in Table 2, which can, for example, help us understand why the FFNN has comparably high MdAPE in spite of having the lowest MSE. We display the MdAPE because it is a relative error measure, which is especially useful to compare performance at different ages. All models tend to yield lower MdAPE for higher ages (75 and above). In particular, the FFNN produces quite large MdAPE for younger ages and seems to be more recommendable for ages above 80. For the forecast error in the time dimension, there is an unsurprising increase with the length of the forecasting period, which is more pronounced for the RNN and less pronounced for the FFNN and CNN. Looking at the population-wise errors, none of the models performs best for all the populations. For the female populations of Spain, France and especially Japan, the MdAPE of the CNN model is noticeably larger than that of the other models. For many of the remaining populations, however, the CNN performs better than or at least similarly to the other models.

To better understand the forecast errors of the CNN for Japanese females, we plot its forecasts for some ages and compare them to the ground truth and the LC20 benchmark model in Figure 6. They increase over time at all ages, which is neither plausible nor in accordance with the real development. Curiously, for most ages even the prediction for the first out-of-sample year is too high. This wrong prediction is then used as an input for the second out-of-sample year forecast and so on, possibly leading to a propagation of errors. This observation illustrates the need for an evaluation of the forecasts of any mortality model, but of course especially of black-box models, with respect to biological reasonableness (see Cairns *et al*. 2006). A potential reason for this behavior might be the fact that mortality rates of Japanese females are very low. Therefore, the CNN receives input matrices with death rates which are smaller than most of – or for some ages even all – the death rates it was trained on. It has to extrapolate in the sense of making a prediction for an input which lies on the boundary or even outside of the range of the training data. Other types of NN are structurally better suited to deal with this issue, either by learning a decreasing dependence on the year (FFNN) or by extrapolating an observed falling trend (RNN), whereas caution has to be taken when applying a CNN for the prediction of populations with very low mortality rates. If

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

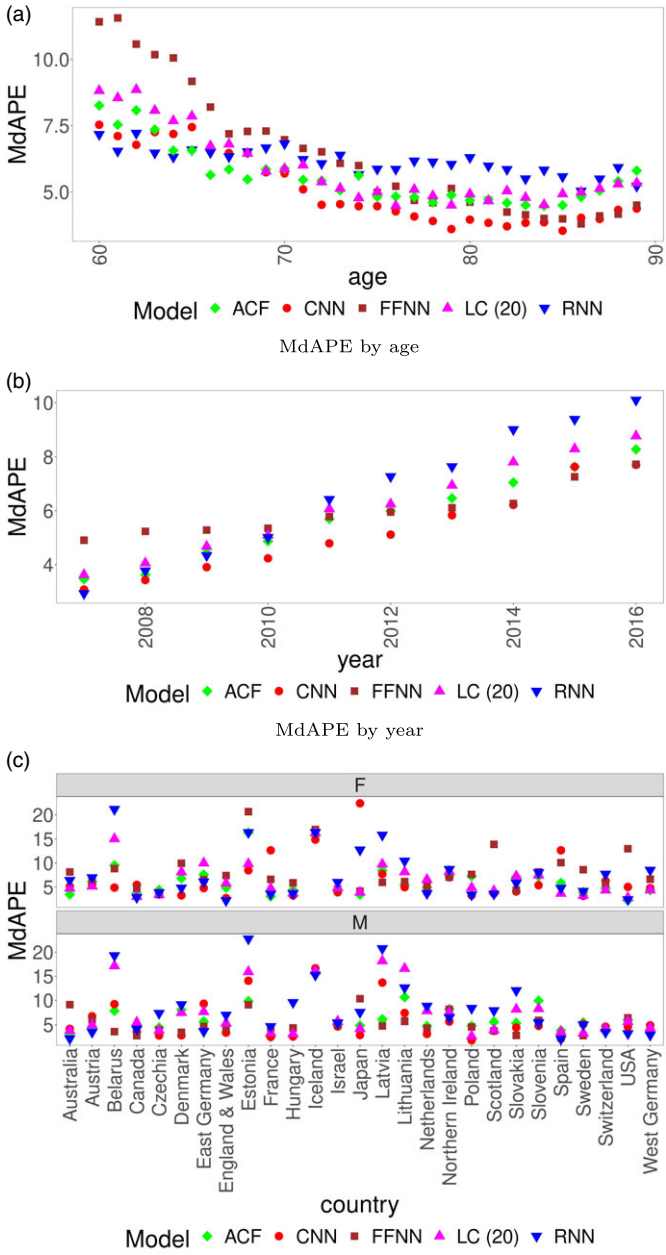

FIGURE 5: MdAPE by age, year and population of CNN (red circles), FFNN (brown squares), RNN (blue inverted triangles), ACF (green diamonds) and LC20 (magenta triangles).

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

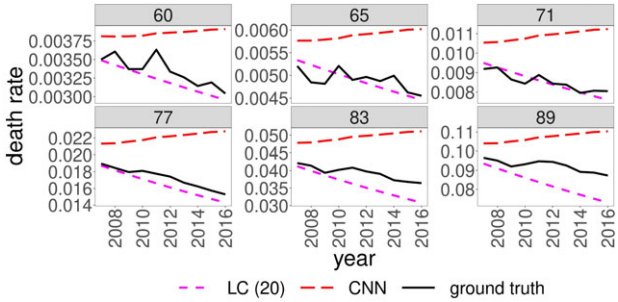

FIGURE 6: CNN (red, long dash) and LC20 (magenta, dash) forecasts and ground truth (black, solid) from 2007 to 2016 for Japanese females aged 60, 65, 71, 77, 83, 89.

one is interested in forecasting mortality in such populations, the training data should possibly be adjusted to include further low-mortality populations so that more data in the mortality range of interest are shown to the NN during training.

4.3. Prediction uncertainty

For measuring the quality of prediction intervals, Khosravi *et al.* (2011a) consider the prediction interval coverage probability (PICP)

$$\text{PICP} := \frac{1}{N}\sum_{x,t,i} \mathbb{1}_{\left\{m^i_{x,t} \in \left[\hat{m}^{i,\text{lower}}_{x,t}, \hat{m}^{i,\text{upper}}_{x,t}\right]\right\}}, \tag{4.1}$$

where $\mathbb{1}$ denotes an indicator function. A large PICP only ensures reliability (true values lie in the interval sufficiently often) but not informativeness of the intervals (intervals are as narrow as possible). For this aspect, Khosravi *et al.* (2011a) propose the mean prediction interval width (MPIW)

$$\text{MPIW} := \frac{1}{N}\sum_{x,t,i}\left(\hat{m}^{i,\text{upper}}_{x,t} - \hat{m}^{i,\text{lower}}_{x,t}\right). \tag{4.2}$$

Good prediction intervals should minimize MPIW under the constraint that PICP is at or above the specified threshold a. Here, we set $a = 0.95$.

In Table 3, we show an evaluation of prediction interval performance for all models. The prediction intervals for the LC, ACF and RNN models ignore some uncertainty as their realized coverage probabilities are significantly smaller than 95%. Both the FFNN (97.0%) and the CNN (98.0%) fulfil the requirement that PICP ≥ 0.95. As a look at the MPIW shows, this comes at the cost of a slightly higher prediction interval width, where the prediction intervals of the FFNN are slightly more informative (narrow) on average compared to the CNN.

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

TABLE 3

PREDICTION INTERVAL MEASURES OVER 54 POPULATIONS, AGES 60–89, YEARS 2007–2016 (MODELS TRAINED ON YEARS UP TO 2006).

Model	PICP (%)	MPIW
LC10	74.0	0.012
LC20	74.3	0.011
ACF	77.2	0.010
FFNN	97.0	0.017
RNN	86.0	0.015
CNN	98.0	0.019

We provide some details regarding the dependence of the PICP on age, year and population in Figure 7. The ACF and LC models are very unreliable at the boundaries of the considered age range, while the NN approaches are more stable across ages. There is also a dependence on the length of the forecasting horizon. The LC model improves from a PICP of around 60% in the first year to around 75% in the last year, while the RNN gets less reliable with increasing forecasting horizon. Both FFNN and CNN have high PICPs over the whole forecasting horizon. However, the reliability of the FFNN slightly decreases over time. Finally, the PICP also varies by population, and the mortality rates of some populations are very hard to anticipate for the ACF, LC and RNN models. For some male populations (Belarus, Estonia, Latvia, Lithuania), the CNN is outperformed by the FFNN, but apart from that its prediction intervals are remarkably reliable.

The MPIW of the NN models is entirely determined by their central forecasts and their variance estimates. Therefore, it is equally informative to directly consider these estimated variances instead of MPIW. We show model uncertainty $\hat{\sigma}^2$, noise variance $\hat{\sigma}_\varepsilon^2$ and total variance $\hat{s}^2$ by age, year and population in Figure 8. For all three models, we observe a decrease both in model uncertainty and noise variance with age. On the other hand, there is an increase with the length of the forecasting horizon, which is particularly strong for the model uncertainty of the CNN. This indicates that the estimation method for multistep prediction intervals outlined in Section 3 leads to an increase of the estimated uncertainty in long-term forecasts as required. For the FFNN, total variance is also increasing, but the slope is rather modest. The dependence of the variances on the population is dominated by some populations with high estimated noise variance, very notably Iceland but also Estonia, Northern Ireland and Slovenia. The noise variances of some populations are estimated quite differently by the three models. For example, compared to the other two models, the FFNN seems to overestimate the noise variance of Danish, Scottish and US females and the CNN seems to overestimate the noise variance of Japanese females. This is interesting considering the failure of the CNN to accurately forecast mortality rates of this population, see Section 4.2.

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

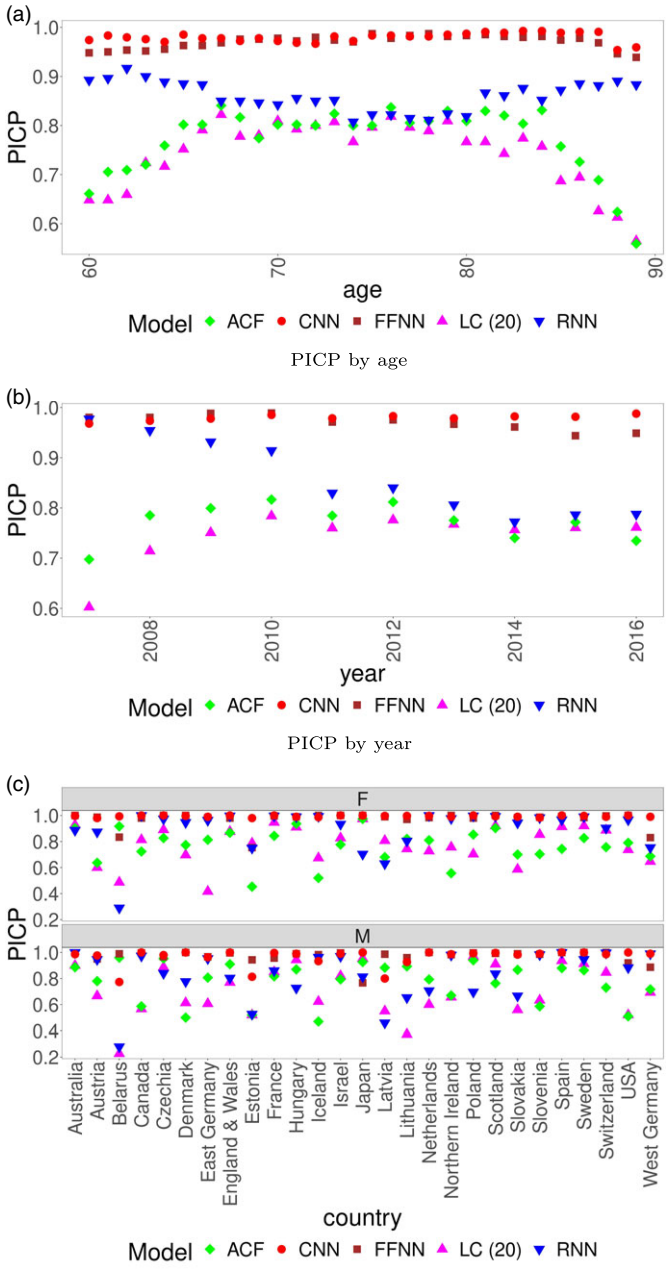

FIGURE 7: PICP by age, year and population of CNN (red circles), FFNN (brown squares), RNN (blue inverted triangles) ACF (green diamonds) and LC20 (magenta triangles).

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

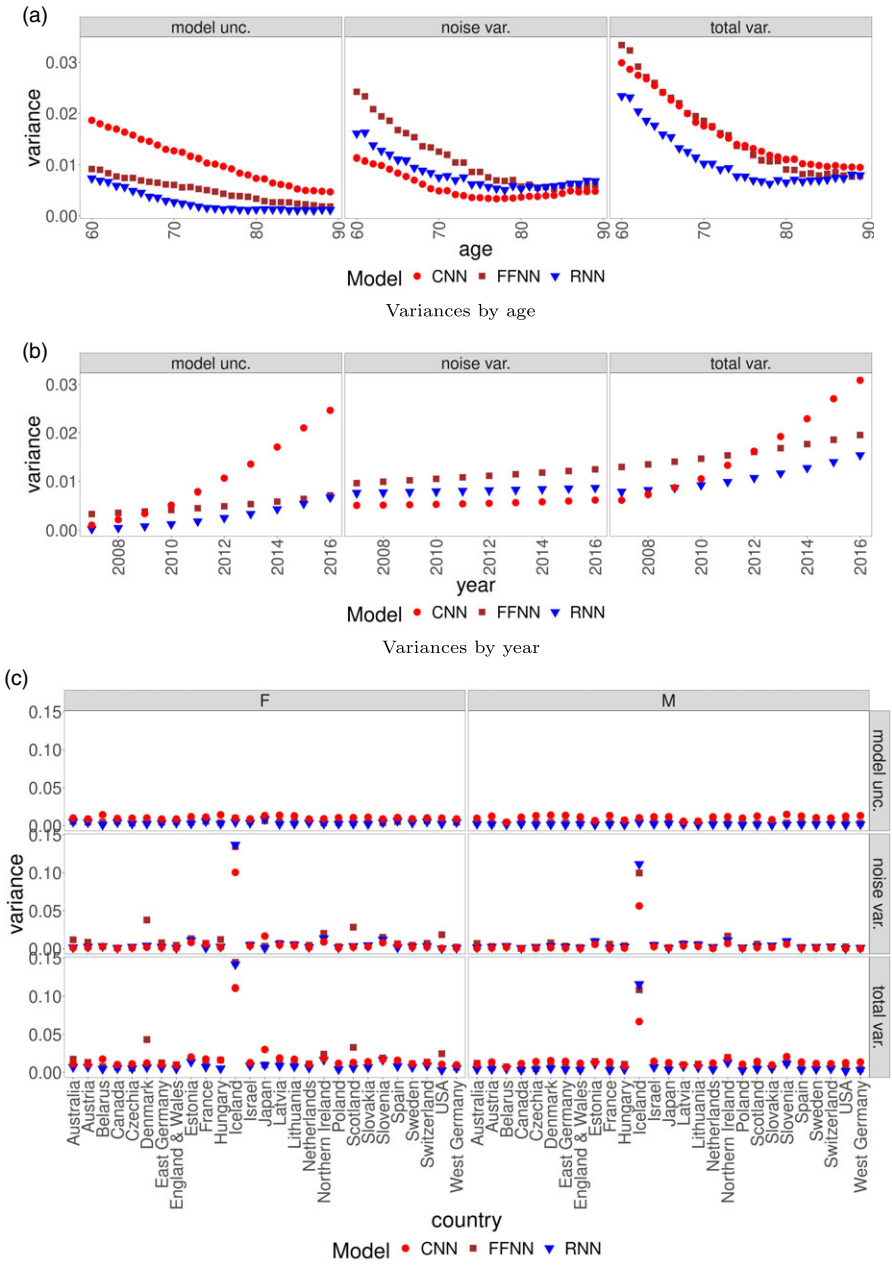

FIGURE 8: Estimated variances by age, year and population for the forecasts of CNN (red circles), FFNN (brown squares) and RNN (blue inverted triangles).

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

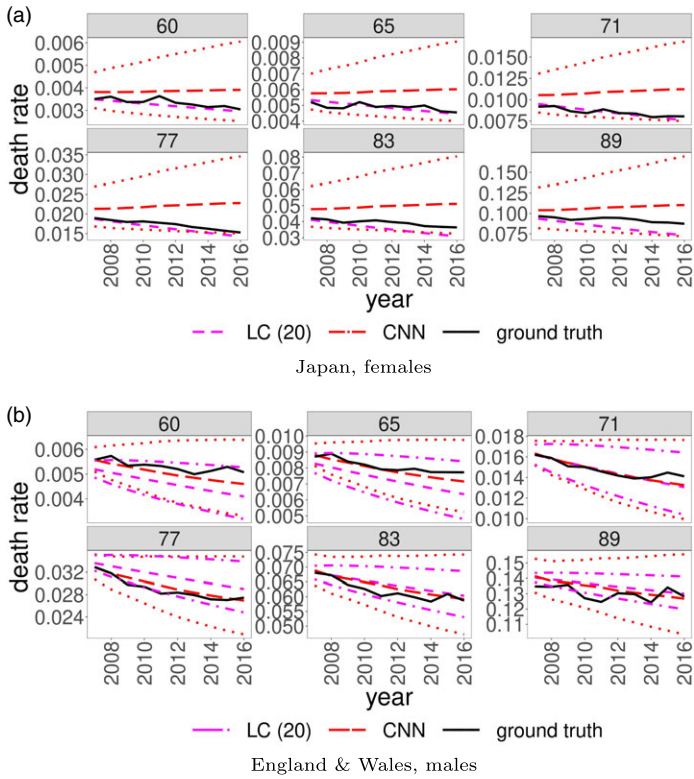

FIGURE 9: CNN (red, long dash) and LC20 (magenta, dash) forecasts along with prediction intervals ($a = 0.95$; dot for CNN, in (b) also dot and dash for LC20) and ground truth (black, solid) from 2007 to 2016 for ages 60, 65, 71, 77, 83, 89.

In fact, the availability of prediction intervals somewhat ameliorates this failure as we see in Figure 9(a). Even though the central forecast of the CNN for Japanese females is both erroneous and biologically implausible, all the observations lie within the prediction intervals. In particular, when relying on the lower bound we would never have overestimated the true death rates. Figure 9(b) shows the forecasts of the CNN and the LC20 model for the English and Welsh females with the prediction intervals included. Here, we mainly observe that the CNN intervals can be considerably wider than the LC intervals. One should keep in mind the results in Table 3 and Figure 7, which show that the LC prediction intervals are often too narrow – compared to this, prediction intervals which are sometimes too wide, that is, too conservative would be preferable in many applications.

4.4. Robustness check

To check the robustness of our forecasting performance results with respect to different training data and a larger test set, we have trained the models

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

TABLE 4

ROBUSTNESS CHECK. OUT-OF-SAMPLE ERROR MEASURES FOR 50 POPULATIONS, AGES 60–89, YEARS 1997–2016 (MODELS TRAINED ON YEARS UP TO 1996). THE BEST VALUE IN EACH COLUMN, WHERE APPLICABLE, IS MARKED IN BOLD.

Model	MSE$\times 10^5$	MAE$\times 10^3$	MdAPE (%)	Dev	PICP (%)	MPIW
LC20	18.9	7.8	11.1	69.0	67.6	0.024
ACF	14.9	7.4	11.0	64.3	57.8	0.016
FFNN	14.4	6.9	10.6	**59.5**	85.0	0.027
RNN	21.2	8.8	12.4	83.6	60.3	0.020
CNN	**9.1**	**5.7**	**8.2**	94.7	**92.5**	0.030

on the years up to 1996 and evaluated them on 1997–2016. In doing so, we use some out-of-sample information at training time for this particular evaluation because the optimal hyperparameters for the NN models were chosen based on data containing the years 1997–2006. However, as this holds true for all three NN models, at least a fair comparison between them should be possible. The resulting out-of-sample error measures are shown in Table 4. We find that the RNN still performs worst, while the CNN is the best model with respect to all error measures except for the Poisson deviance (Dev). This indicates that the performance of the CNN could become clearly superior when longer forecasting horizons are considered. As in Table 2, its high deviance is mostly driven by Japanese females and would equal 49.5 when excluding this population. The second-best model is the FFNN, followed by ACF and LC20.

We refer to Section C.2 of the Online Supplementary Material for an additional robustness check with respect to the evaluation age range, which shows that the CNN is superior to the other models when evaluated over all available ages as well.

4.5. Long-term forecasts

For actuarial applications, mortality models should produce plausible forecasts even for longer time horizons. Figure 10 shows the development of age-specific mortality forecasts for the male and female populations of England and Wales by the three NN models, the LC model trained on 30 years of data (LC30) and the ACF model over a 1-year, 10-year, 20-year and 30-year horizon. For the 1-year and the 10-year forecasts, which correspond to the years 2007 and 2016, we also provide the ground truth, that is, the realized death rates during these years. We consider LC30 here because according to an often-used rule of thumb, the training horizon for an LC model should be at least as long as the forecasting horizon (see Janssen and Kunst 2007).

All models predict an approximately linear dependence of log death rates on age even in the far future and, with the exception of the RNN, age-specific improvements which tend to decrease over age. This is in line with

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

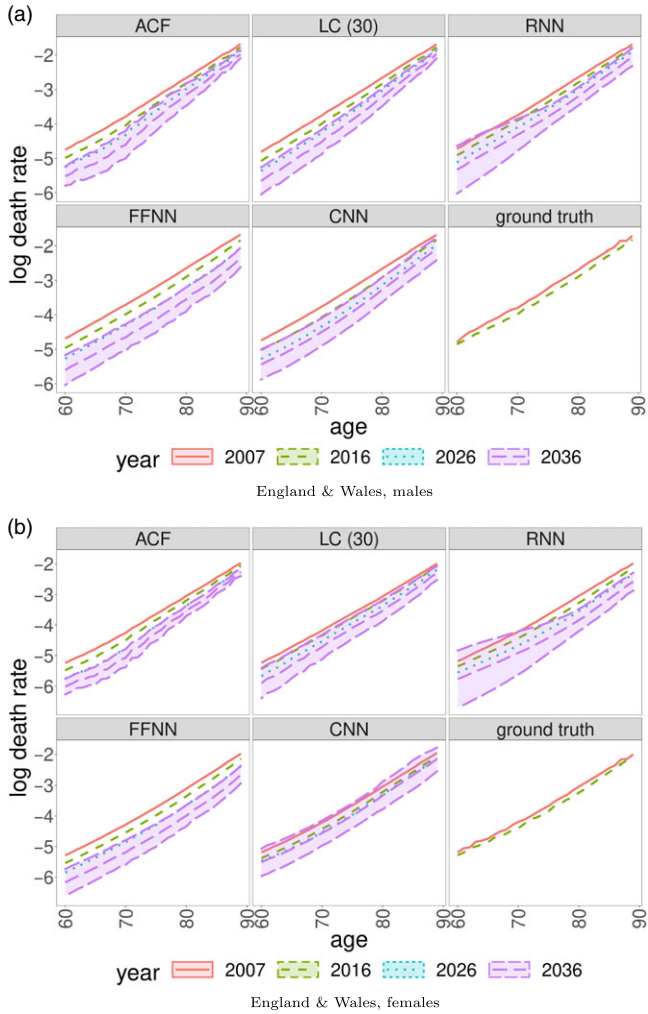

FIGURE 10: Log death rate forecasts of different mortality models and, where available, ground truth for the years 2007 (red, solid), 2016 (lime green, dash), 2026 (blue, dot) and 2036 (violet, long dash; including prediction intervals) for England and Wales.

the demographic literature (see Li *et al*. 2013). The FFNN forecasts substantial mortality improvements at all ages, which might turn out to be somewhat optimistic considering the real development between 2007 and 2016. The CNN and the LC model forecast stronger improvements for males than for females. While ACF, LC and particularly RNN model yield narrowing intervals for higher ages, FFNN and CNN prediction interval widths are similar over the considered age range. The notable differences between the forecasts of the models indicate the existence of a nonnegligible amount of model risk. This model risk can to some degree be quantified numerically by calculating implied

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

present values of annuities, which we present in Section C.3 of the Online Supplementary Material.

5. Conclusion

We have proposed a CNN for mortality forecasting. It outperforms the ACF, LC and RNN models in an out-of-sample evaluation with respect to all considered error measures except for Poisson deviance, which heavily depends on population size. An FFNN achieves lower out-of-sample quadratic errors than our CNN, but with respect to relative errors the CNN outperforms the FFNN as well. Checking the robustness of our results on a longer time period (20 years), we confirm the convincing forecasting performance of our CNN.

There is no single model which performs best for all ages, years or populations. In particular, a look at population-specific errors illustrates the need for a careful investigation whether a model works well for a population of interest. For example, we observe unrealistic forecasts of the CNN for a few populations with low mortality rates so that we recommend either not to use this model for populations which lie at the boundary of the training data distribution or to extend the training data accordingly.

Generally, NN models have the limitation of being conceptually more complex and less interpretable than, for example, the LC model. We strongly advise for preliminary studies similar to those in Section 4 before basing any decisions on a black-box model to ensure that the forecasts are biologically reasonable. Looking at a global surrogate model and at long-term forecasts produced by the CNN, we gain confidence that this is the case for many of the populations we consider. Based on our results, we believe that black-box models can be (at least) a helpful addition to classical approaches such as the LC model or deterministic life tables in demographic and actuarial modeling. This is supported by the results we obtain on prediction uncertainty. Our CNN model yields reliable prediction intervals as well as a plausible increasing development of model uncertainty with the length of the forecasting horizon.

NN models are computationally expensive. Training our models takes approximately 3–4 days (on a machine with 28 CPU cores, 2.6 GHz per core and 192 GB RAM), with an additional 3 days for the noise variance FFNN if prediction uncertainty is to be quantified. As re-training a model is only necessary when new mortality data are available, which can be expected to occur with yearly rather than weekly frequency, and obtaining forecasts from an already trained model is achieved very quickly, this does not impede the practical applicability of these models.

There are several ways to improve or extend the considered models:

- As proposed by Perla *et al*. (2021), combinations of model architectures could be considered, for example, by connecting the outputs of

https://doi.org/10.1017/asb.2021.34 Published online by Cambridge University Press

an FFNN and a CNN to a further dense layer and in this sense training an FFNN–CNN ensemble. This is an application of stacking (Wolpert, 1992) with the dense layer as meta learner. Averaging multiple different models in this way could also reduce model risk.

- We have trained both the RNN and the CNN model to make one-step forecasts and then obtained multistep forecasts by recursion. Although this is a standard approach in time series forecasting, there are other strategies which are worth investigating, see Ben Taieb and Atiya (2016) and the references therein.
- For estimating prediction uncertainty we make two assumptions, namely that the noise $\varepsilon(z)$ in (3.2) follows a normal distribution and that the out-of-sample bias of the NN estimators equals zero. The convincing results in Section 4.3 justify these assumptions ex post. Nevertheless, the quality of the uncertainty estimates if they are violated and possible remedies should be investigated.

ACKNOWLEDGMENTS

We are indebted to two anonymous reviewers for their suggestions which have substantially improved the article. We further wish to thank Andreas Wagner and Mario Wüthrich for valuable advice. S. Schnürch is grateful for the financial support from the Fraunhofer Institute for Industrial Mathematics ITWM.

SUPPLEMENTARY MATERIALS

For supplementary material for this article, please visit http://dx.doi.org/10.1017/asb.2021.34

REFERENCES

BEN TAIEB, S. and ATIYA, A.F. (2016) A bias and variance analysis for multistep-ahead time series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, **27** (1), 62–76. doi: 10.1109/TNNLS.2015.2411629.

BERGMEIR, C., HYNDMAN, R.J. and KOO, B. (2018) A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, **12**, 70–83. ISSN 01679473. doi: 10.1016/j.csda.2017.11.003.

BOOTH, H., MAINDONALD, J. and SMITH, L. (2002) Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, **56** (3), 325–336. ISSN 0032-4728. doi: 10.1080/00324720215935.

BREIMAN, L. (1996) Bagging predictors. *Machine Learning*, **24** (2), 123–140. ISSN 1573-0565. doi: 10.1023/A:1018054314350.

BROUHNS, N., DENUIT, M. and VERMUNT, J.K. (2002) A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, **31** (3), 373–393. doi: 10.1016/S0167-6687(02)00185-3.

CAIRNS, A.J.G., BLAKE, D.P. and DOWD, K. (2006) Pricing death: Frameworks for the valuation and securitization of mortality risk. *ASTIN Bulletin*, **36** (1), 79–120. ISSN 0515-0361. doi: 10.1017/S0515036100014410.

CAIRNS, A.J.G., BLAKE, D.P., DOWD, K., COUGHLAN, G.D., EPSTEIN, D.P., ONG, A. and BALEVICH, I. (2009) A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, **13** (1), 1–35. doi: 10.1080/10920277.2009.10597538.

CARNEY, J.G., CUNNINGHAM, P. and BHAGWAN, U. (1999) Confidence and prediction intervals for neural network ensembles. *IJCNN'99 International Joint Conference on Neural Networks, Washington, DC, July 10–16, 1999*, pp. 1215–1218, Piscataway, NJ. Institute of Electrical and Electronics Engineers. ISBN 0-7803-5529-6. doi: 10.1109/IJCNN.1999.831133.

D'AMATO, V., DI LORENZO, E., HABERMAN, S., RUSSOLILLO, M. and SIBILLO, M. (2011) The Poisson log-bilinear Lee-Carter model. *North American Actuarial Journal*, **15** (2), 315–333. doi: 10.1080/10920277.2011.10597623.

D'AMATO, V., HABERMAN, S., PISCOPO, G. and RUSSOLILLO, M. (2012) Modelling dependent data for longevity projections. *Insurance: Mathematics and Economics*, **51** (3), 694–701. ISSN 0167-6687. doi: 10.1016/j.insmatheco.2012.09.008.

DAV, D. (2018) Herleitung der DAV-Sterbetafel 2004 R für Rentenversicherungen. Fachgrundsatz der Deutschen Aktuarvereinigung e. V.

DENTON, F.T., FEAVER, C.H. and SPENCER, B.G. (2005) Time series analysis and stochastic forecasting: An econometric study of mortality and life expectancy. *Journal of Population Economics*, **18** (2), 203–227. ISSN 0933-1433. doi: 10.1007/s00148-005-0229-2.

DENUIT, M., HAINAUT, D. and TRUFIN, J. (2019) *Effective Statistical Learning Methods for Actuaries. III, Neural Networks and Extensions*. *Springer Actuarial*, pp. 2523–3289. Cham: Springer. ISBN 9783030258276.

FALBEL, D., ALLAIRE, J.J., CHOLLET, F., TANG, Y., VAN DER BIJL, W., STUDER, M. and KEYDANA, S. (2019) R Interface to Keras. URL https://github.com/rstudio/keras.

GAL, Y. and GHAHRAMANI, Z. (2016) Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Preprint. URL https://arxiv.org/pdf/1506.02142.

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016) *Deep Learning*. *Adaptive Computation and Machine Learning*. Cambridge, MA: MIT Press. ISBN 9780262035613. URL https://www.deeplearningbook.org/.

GUO, C. and BERKHAHN, F. (2016) Entity Embeddings of Categorical Variables. Preprint. URL https://arxiv.org/abs/1604.06737.

HESKES, T. (1997) Practical confidence and prediction intervals. In *Advances in Neural Information Processing Systems* (eds. M.I. JORDAN, M.C. MOZER and T. PETSCHE), vol. 9, pp. 176–182, Cambridge, MA: MIT Press. ISBN 0262100657.

Human Mortality Database. (2019) University of California, Berkeley (USA), and Max Planck Institute for Demographic Research, Rostock (Germany). Data downloaded on July 2. URL www.mortality.org.

IOFFE, S. and SZEGEDY, C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Preprint. URL http://arxiv.org/pdf/1502.03167v3.

JANSSEN, F. and KUNST, A. (2007) The choice among past trends as a basis for the prediction of future trends in old-age mortality. *Population Studies*, **61** (3), 315–326. ISSN 0032-4728. doi: 10.1080/00324720701571632.

KHOSRAVI, A., NAHAVANDI, S., CREIGHTON, D. and ATIYA, A.F. (2011a) Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on Neural Networks*, **22** (9), 1341–1356. ISSN 1045-9227. doi: 10.1109/TNN.2011.2162110.

KHOSRAVI, A., NAHAVANDI, S., CREIGHTON, D. and ATIYA, A.F. (2011b) Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE Transactions on Neural Networks*, **22** (3), 337–346. ISSN 1045-9227. doi: 10.1109/TNN.2010.2096824.

KINGMA, D.P. and BA, J. (2014) Adam: A Method for Stochastic *Optimization*. Preprint. URL http://arxiv.org/pdf/1412.6980v9.

KLEINOW, T. and RICHARDS, S.J. (2016) Parameter risk in time-series mortality forecasts. *Scandinavian Actuarial Journal*, **2016** (9), 804–828. ISSN 0346-1238. doi: 10.1080/03461238.2016.1255655.

KOISSI, M.-C., SHAPIRO, A.F. and HÖGNÄS, G. (2006) Evaluating and extending the Lee–Carter model for mortality forecasting: Bootstrap confidence interval. *Insurance: Mathematics and Economics*, **38** (1), 1–20. doi: 10.1016/j.insmatheco.2005.06.008.

LECUN, Y., BOSER, B., DENKER, J.S., HENDERSON, D., HOWARD, R.E., HUBBARD, W. and JACKEL, L.D. (1989) Backpropagation applied to handwritten zip code recognition. *Neural Computation*, **1** (4), 541–551. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541.

LECUN, Y., BOTTOU, L., BENGIO, Y. and HAFFNER, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86** (11), 0 2278–2324. ISSN 00189219. doi: 10.1109/5.726791.

LEE, R.D. and CARTER, L.R. (1992) Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, **87** (419), 659–671. ISSN 0162-1459. doi: 10.1080/01621459.1992.10475265.

LENAIL, A. (2019) NN-SVG: Publication-ready neural network architecture schematics. *Journal of Open Source Software*, **4** (33), 747. doi: 10.21105/joss.00747.

LI, N. and LEE, R.D. (2005) Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, **42** (3), 575–594. doi: 10.1353/dem.2005.0021.

LI, N., LEE, R.D. and GERLAND, P. (2013) Extending the Lee-Carter method to model the rotation of age patterns of mortality decline for long-term projections. *Demography*, **50** (6), 2037–2051. doi: 10.1007/s13524-013-0232-2.

MEIER, D. and WÜTHRICH, M.V. (2020) Convolutional neural network case studies: (1) anomalies in mortality rates (2) image recognition. Tutorial, SSRN. URL https://ssrn.com/abstract=3656210.

NIGRI, A., LEVANTESI, S., MARINO, M., SCOGNAMIGLIO, S. and PERLA, F. A deep learning integrated Lee–Carter model. *Risks*, **7** (1), 33, 2019. doi: 10.3390/risks7010033.

NIX, D.A. and WEIGEND, A.S. (1994) Estimating the mean and variance of the target probability distribution. *In Neural Networks*, vol. **1**, pp. 55–60. IEEE. ISBN 0-7803-1901-X. doi: 10.1109/ICNN.1994.374138.

PERLA, F., RICHMAN, R., SCOGNAMIGLIO, S. and WÜTHRICH, M.V. (2021) Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*, **7**, 572–598. ISSN 0346-1238. doi: 10.1080/03461238.2020.1867232.

PITACCO, E., DENUIT, M., HABERMAN, S. and OLIVIERI, A. (2008) *Modelling Longevity Dynamics for Pensions and Annuity Business*. Oxford: Oxford University Press. ISBN 9780199547272.

R Core Team. (2019) R: A language and environment for statistical computing. Vienna, Austria. URL http://www.R-project.org.

RENSHAW, A.E. and HABERMAN, S. (2006) A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, **38** (3), 556–570. doi: 10.1016/j.insmatheco.2005.12.001.

RICHMAN, R. and WÜTHRICH, M.V. (2019) Lee and Carter go Machine Learning: Recurrent Neural Networks. Tutorial, SSRN. URL https://ssrn.com/abstract=3441030.

RICHMAN, R. and WÜTHRICH, M.V. (2020) Nagging predictors. *Risks*, **8** (3), 83. doi: 10.3390/risks8030083.

RICHMAN, R. and WÜTHRICH, M.V. (2021) A neural network extension of the Lee-Carter model to multiple populations. *Annals of Actuarial Science*, **15** (2), 346–366. ISSN 1748-4995. doi: 10.1017/S1748499519000071.

SHAH, P. and GUEZ, A. (2009) Mortality forecasting using neural networks and an application to cause-specific data for insurance purposes. *Journal of Forecasting*, **28** (6), 535–548. ISSN 02776693. doi: 10.1002/for.1111.

TIELEMAN, T. and HINTON, G.E. (2012) Lecture 6.5 - Rmsprop: Divide the gradient by a running average of its recent magnitude. In *Coursera: Neural Networks for Machine Learning*.

WANG, C.-W., ZHANG, J. and ZHU, W. (2021) Neighboring prediction for mortality. *ASTIN Bulletin*, **51** (3), 689–718. ISSN 0515-0361. doi: 10.1017/asb.2021.13.

WEN, J., CAIRNS, A.J.G. and KLEINOW, T. (2021) Fitting multi-population mortality models to socio-economic groups. *Annals of Actuarial Science*, **15** (1), 144–172. ISSN 1748-4995. doi: 10.1017/S1748499520000184.
WICKHAM, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer. ISBN 9783319242774. URL https://ggplot2.tidyverse.org.
WOLPERT, D.H. (1992) Stacked generalization. *Neural Networks*, **5** (2), 241–259.ISSN 08936080. doi: 10.1016/S0893-6080(05)80023-1.

SIMON SCHNÜRCH (CORRESPONDING AUTHOR)
Department of Financial Mathematics
Fraunhofer Institute for Industrial Mathematics ITWM
Fraunhofer-Platz 1
67663 Kaiserslautern, Germany

Department of Mathematics
University of Kaiserslautern
Gottlieb-Daimler-Straße 48
67663 Kaiserslautern, Germany
E-Mail: simon.schnuerch@itwm.fraunhofer.de

RALF KORN
Department of Financial Mathematics
Fraunhofer Institute for Industrial Mathematics ITWM
Fraunhofer-Platz 1
67663 Kaiserslautern, Germany

Department of Mathematics
University of Kaiserslautern
Gottlieb-Daimler-Straße 48
67663 Kaiserslautern, Germany