A Neural Approach to Improve the Lee-Carter Mortality Density Forecasts

Mario Marino,[1], Susanna Levantesi,[2] and Andrea Nigri[3]

[1]*Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome, Rome, Italy*
[2]*Department of Statistical Sciences, Sapienza University of Rome, Rome, Italy*
[3]*Department of Social and Political Sciences, Bocconi University, Milan, Italy*

Several countries worldwide are experiencing a continuous increase in life expectancy, extending the challenges of life actuaries and demographers in forecasting mortality. Although several stochastic mortality models have been proposed in the literature, mortality forecasting research remains a crucial task. Recently, various research works have encouraged the use of deep learning models to extrapolate suitable patterns within mortality data. Such learning models allow achieving accurate point predictions, though uncertainty measures are also necessary to support both model estimate reliability and risk evaluation. As a new advance in mortality forecasting, we formalize the deep neural network integration within the Lee-Carter framework, as a first bridge between the deep learning and the mortality density forecasts. We test our model proposal in a numerical application considering three representative countries worldwide and for both genders, scrutinizing two different fitting periods. Exploiting the meaning of both biological reasonableness and plausibility of forecasts, as well as performance metrics, our findings confirm the suitability of deep learning models to improve the predictive capacity of the Lee-Carter model, providing more reliable mortality boundaries in the long run.

1. INTRODUCTION

Since the second half of the 20th century, mortality has exhibited notable improvements, despite country-specific behavior experienced in industrialized regions (Levantesi, Nigri, and Piscopo 2021; Nigri, Barbi, and Levantesi 2021), engaging attention from life insurers and pension systems, as well as from actuarial and demographic researchers. Principally, mortality reductions in modern populations arise from a continuous flow of social progress (Oeppen and Vaupel 2006). In fact, industrialized countries have made efforts to improve the socioeconomic development, health system, and lifestyle of their populations, impacting how mortality will vary in the future. Various factors move human longevity trends, and different mortality scenarios should be anticipated through predictive analysis. The need for accurate forecasting to address longevity risk and adequately price annuity products has led actuaries towards more sophisticated extrapolative methods; in a stochastic environment, see, for instance, Lee and Carter (1992), Brouhns, Denuit, and Vermunt (2002), Renshaw and Haberman (2006), Cairns, Blake, and Dowd (2006), Booth and Tickle (2008), Cairns et al. (2009), Plat (2009), Hunt and Blake (2014), and Currie (2017).

Demographers and actuaries have concentrated their efforts on a model's functional form and its parameterization in order to better explain the mortality structure. In most of these models, mortality projections arise from time-dependent parameters, modeled by time series analysis techniques, the class of autoregressive integrated moving average (ARIMA) processes among all. However, alternative mortality forecasting methods have been suggested in past literature. For instance, a P-spline based approach is proposed in Currie, Durban, and Eilers (2004), where forthcoming values are interpreted as missing values through smoothing procedures. A development of this model was presented in Camarda (2019), overcoming robustness forecasting problems. An innovative proposal was introduced in Mitchell et al. (2013), wherein the Lee-Carter (henceforth LC) time index was predicted through a normal inverse Gaussian distribution, attaining accuracy in the approximation of the observed force of mortality. Furthermore, new advances in mortality modeling, grounded in machine and deep learning models, have recently

Address correspondence to Mario Marino, Department of Methods and Models for Economics, Territory and Finance, Sapienza University of Rome, Rome 00185, Italy. E-mail: m.marino@uniroma1.it

appeared in the literature. The first insight based on machine learning tools was offered in Deprez , Shevchenko, and Wüthrich (2021), where regression trees algorithms were adopted to improve the estimation of death rates from canonical models, such as the LC and the Renshaw-Haberman models. These findings were extended in Levantesi and Pizzorusso (2019) and Levantesi and Nigri (2020) for predictive purposes. A neural network (henceforth NN) design for mortality analysis was initially scrutinized by Hainaut (2018), profitably aiming to extrapolate suitable non-linearity in the observed force of mortality. An NN vision within the LC framework was presented in Nigri et al. (2019), Perla et al. (2021), and Richman and Wüthrich (2021). The former employs a recurrent NN architecture, namely, long short-term memory (henceforth LSTM), to model the future LC time-dependent parameter values. For each country investigated and both genders, numerical experiments performed confirm greater LSTM accuracy w.r.t. the best ARIMA process. The latter proposed an NN representation for the multi population LC model, overcoming parameter optimization problems and achieving reliable forecasting performances. Following this, Perla et al. (2021) showed the remarkable accuracy achieved in a large-scale prediction of mortality. In particular, different NN structures were tested, such as LSTM and convolutional NN, engaging each of them to produce point forecasts of mortality rates simultaneously for many countries. Given the suitability of the recurrent network in forecasting, Nigri, Levantesi, and Marino (2020) considered an LSTM model to predict both life expectancy and life span disparity measures for various countries and both genders.

Deep learning models, especially recurrent NNs (RNNs), are gaining confidence in many forecasting tasks, as well as in mortality. They are dynamic systems stemming from the composition and superposition of non-linear functions, earning notable accuracy gains in predictive issues. Wanting to exploit the latter feature, we aim to investigate the suitability of deep NNs models within the LC framework to extrapolate the future mortality realizations. Contextualizing suggestions expressed in Makridakis, Spiliotis, and Assimakopoulos (2020), our approach pursues a model integrating deep learning techniques in the spirit of Nigri et al. (2019), representing an appropriate compromise between the interpretation of the mortality model and high accuracy in projections. Therefore, we freeze the LC age–period mortality representation, forecasting the mortality profile employing an RNN model.

It is worth to recalling that a proper forecasting model provides robust point predictions, outlining the future mortality trend, as well as confidence ranges of variability. Uncertainty measures associated with the expected values are necessary to sufficiently inspect the phenomenon and to judge both the model adequacy and the reliability of the results. As in actuarial assessments, uncertainty measures, such as prediction intervals, are imperative. This is a compelling topic, because learning models such as NNs furnish only point predictions. To this end, Khosravi et al. (2011) provided an extensive methodological review of the main approaches for calculating confidence and prediction intervals, concluding that no method beats other ones in each considered comparison metric. Procedures based on structural assumptions, such as the delta method (Wild and Seber 1989), mean–variance estimation (Nix and Weigend 1994), and the Bayesian approach (MacKay 1992), are relevant solutions but suffer computational problems that could be prohibitive. At the state of the art, the prevailing approach to forecast prediction intervals for NNs is based on coherent sampling techniques, favoring the estimation of a theoretical probability distribution through an empirical one; see, for instance, Tibshirani (1996), Heskes (1997), Khosravi et al. (2015), Mazloumi et al. (2011), Kasiviswanathan, Sudheer (2014), and K. Li et al. (2018). In particular, bootstrap procedures seem to represent the more tempting alternative because they do not require stringent sampling assumptions, allowing for accurate plug-in estimates (Efron and Tibshirani 1993). In fact, such an approach has become a common practice to measure uncertainty in stochastic mortality models, as emerged in Brouhns, Denuit, and van Keilegom (2005), Koissi, Shapiro, and Hognas (2006), Li et al. (2009), D'Amato et al. (2011), D'Amato, Haberman, and Russolillo (2012), and D'Amato, Haberman, Piscopo, and Russolillo (2012). In the framework of mortality uncertainty forecasting based on NNs, Schnürch and Korn (2021) adapted the boostrap procedure of Heskes (1997) and Carney, Cunningham, and Bhagwan (1999) to estimate prediction intervals for a two-dimensional convolutional NN representation of death rates.

The present work formalizes the integration of deep learning techniques in the LC model framework, in terms of both point estimates and prediction intervals for future mortality rates. We use an RNN with LSTM architecture to forecast the LC time index. The resulting integrated model, namely, LC-LSTM, and mortality boundaries it provides fill the gap between deep learning integrated mortality models and uncertainty estimation, obtaining suitable ranges of variability. This is a step forward in mortality forecasting.

We test the proposed model in a numerical application considering three countries worldwide, Australia, Japan, and Spain, for both genders, scrutinizing two different learning periods to deepen how they could affect the forecasting performances. Our results are assessed considering both qualitative and quantitative criteria. The former were well established in Cairns et al. (2011) and concern (a) the biological reasonableness of mortality forecasts, (b) the plausibility of projected uncertainty at different ages, (c) the robustness of predictions w.r.t. the historical mortality trend. The latter, like performance metrics, are used to assess the resulting mortality forecasts with a back-testing approach. Our findings confirm the LC-LSTMs ability to produce plausible mortality projections, improving the LC predictive capacity, in particular in the long run. The proposed framework might represent a prominent practice in the field of longevity forecasting, as for actuarial business tasks.

The remainder of the article is structured as follow. Section 2 presents the RNN model with LSTM architecture. Section 3 illustrates the LC-LSTM model formalization. Section 4 discusses the uncertainty framework within the LC-LSTM, highlighting the methodology to estimate prediction intervals. Section 5 describes the performance metrics to evaluate both point and interval forecasts. Section 6 collects the results and related comments on the LC-LSTMs application to mortality data. Finally, Section 7 provides concluding remarks.

2. THE NEURAL NETWORK MODEL

An NN model is a computational graph consisting of connected nodes, or neurons, located in consecutive layers. Connections among neurons are pondered by parameters, whose values are learned from data implementing efficient optimization procedures (Rumelhart, Hinton, and Williams 1986). Each neuron receives weighted information, namely, activation, and transforms it employing a differentiable function, the activation function. As a consequence, NN outputs descend from composition and superposition of differentiable functions, providing flexible data-driven tools that deeply gather data features and generalize them.

For forecasting purposes, RNNs are used to handle sequential data such as time series. In RNNs recurrent connections between neurons are added, so that the network processes data creating a dynamic memory. However, RNN learning optimization is tricky because of the vanishing or exploding gradient problems (Pascanu, Mikolov, and Bengio 2013). To address such a problem, Hocreiter and Schmidhuber (1997) proposed the LSTM architecture, whose more engineered structure relies both on a memory block and gates, essentials for controlling data elaborations. In the following, we will consider the RNN with LSTM architecture, referring the interested reader to Goodfellow, Bengio, and Courville (2016), Aggarwal (2018), and references therein for further details on RNNs and LSTM.

2.1. RNNs with LSTM Architecture

In order to define the general structure of the RNN with LSTM architecture, let N_0 be the number of neurons within the input layer, N_p the number of neurons of the pth hidden layer with $p \in \{1, ..., P\}$, and N_{P+1} the number of neurons of the output layer. We have $P, N_0, N_p, N_{P+1} \in \mathbb{N}$. Let $A^{(p)} : \mathbb{R}^{N_{p-1}} \to \mathbb{R}^{N_p}$ be an affine map defining the p^{th} hidden layer activation, given the output produced by the $(p-1)^{\text{th}}$ hidden layer, and let $\phi : \mathbb{R}^{N_p} \to \mathbb{R}^{N_p}$ be a differentiable activation function.

Definition 1. *The output of an LSTM neuron at time t in the* pth *hidden layer is:*

$$H_t^{(p)} = \boldsymbol{o}_t^{(p)} \odot \tanh\left(\boldsymbol{c}_t^{(p)}\right), \tag{1}$$

where the components of the element-wise product stem from the LSTM neuron internal forward flow described by the following equations:

$$\begin{aligned}
&\text{Forget gate :} && \boldsymbol{f}_t^{(p)} = \sigma_f \circ A^{(p)} = \sigma\left(\langle \boldsymbol{W}_f^{(p)}, H_t^{(p-1)} \rangle + \langle \boldsymbol{U}_f^{(p)}, H_{t-1}^{(p)} \rangle + \boldsymbol{b}_f^{(p)}\right),\\
&\text{Input gate :} && \boldsymbol{i}_t^{(p)} = \sigma_i \circ A^{(p)} = \sigma\left(\langle \boldsymbol{W}_i^{(p)}, H_t^{(p-1)} \rangle + \langle \boldsymbol{U}_i^{(p)}, H_{t-1}^{(p)} \rangle + \boldsymbol{b}_i^{(p)}\right),\\
&\text{Output gate :} && \boldsymbol{o}_t^{(p)} = \sigma_o \circ A^{(p)} = \sigma\left(\langle \boldsymbol{W}_o^{(p)}, H_t^{(p-1)} \rangle + \langle \boldsymbol{U}_o^{(p)}, H_{t-1}^{(p)} \rangle + \boldsymbol{b}_o^{(p)}\right),\\
&\text{Memory state :} && \boldsymbol{c}_t^{(p)} = \boldsymbol{f}_t^{(p)} \odot \boldsymbol{c}_{t-1}^{(p)} + \boldsymbol{i}_t^{(p)} \odot \tanh\left(\langle \boldsymbol{W}_c^{(p)}, H_t^{(p-1)} \rangle + \langle \boldsymbol{U}_c^{(p)}, H_{t-1}^{(p)} \rangle + \boldsymbol{b}_c^{(p)}\right),
\end{aligned} \tag{2}$$

where $\sigma(x) = (1 + e^x)^{-1}$ *is the sigmoid function,* $\tanh(x) = (e^x - e^{-x})(e^x + e^{-x})^{-1}$ *is the hyperbolic tangent function,* $\left(\boldsymbol{W}_l^{(p)}, l = f, i, o, c\right)$ *are the weight matrices for gates, feedforward connections,* $\left(\boldsymbol{U}_l^{(p)}, l = f, i, o, c\right)$ *are the weight matrices for gates, recurrent connections, and* $\left(\boldsymbol{b}_l^{(p)}, l = f, i, o, c\right)$ *are the bias terms.*

Definition 2. *Let* $\mathcal{D} = \left\{(\boldsymbol{x}_t, \boldsymbol{y}_t), \boldsymbol{x}_t \in \mathbb{R}^{N_0}, \boldsymbol{y}_t \in \mathbb{R}^{N_{P+1}}\right\}$ *be a dataset wherein* $\boldsymbol{x}_t$ *is the input variable at time t and* $\boldsymbol{y}_t$ *the associated response. An RNN with LSTM architecture is a function* $f_{LSTM} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{P+1}}$ *such that*

$$\boldsymbol{y}_t = f_{LSTM}(\boldsymbol{x}_t; \mathcal{W}) + \gamma_t = \psi \circ \left(H_t^{(P)} \circ H_t^{(P-1)} \circ \cdots \circ H_t^{(1)}\right)(\boldsymbol{x}_t; \mathcal{W}) + \gamma_t, \tag{3}$$

where $\psi : \mathbb{R}^{N_P} \to \mathbb{R}^{N_{P+1}}$ is the output layer activation function, $\mathcal{W}$ is the set of all NN parameters, and γ_t is a noise term, with zero mean and variance σ^2_γ and independent of f_{LSTM}.

Starting from Equation (3), our proposal aims at creating a bridge between deep learning and mortality forecasting structured on the LC model. The following sections formalize such a proposal, in terms of both point and interval forecasts.

3. THE LC-LSTM MODEL

Let us consider the LC Poisson model proposed in Brouhns, Denuit, and Vermunt (2002) as the reference model describing the behavior of the age–period mortality rates. Hence, for ages $x \in \mathcal{X} = \{0, 1, ..., \omega\}$ and calendar years $t \in \mathcal{T} = \{t_0, t_1, ..., t_n\}$, we assume that the observed number of deaths, $D_{x,t}$, follows a Poisson distribution:

$$D_{x,t} \sim Poi(E^c_{x,t} m_{x,t}), \tag{4}$$

where $E^c_{x,t}$ is the central exposure to the death risk and $m_{x,t} = \mathbb{E}\left(\frac{D_{x,t}}{E^c_{x,t}}\right)$ is the central death rate. The LC model structure associated to assumption (4) is defined by following log-bilinear equation:

$$\log m_{x,t} = \alpha_x + \beta_x k_t, \tag{5}$$

where α_x and β_x are age-dependent parameters illustrating the mortality age patterns and k_t is the time index parameter representing the mortality behavior over time. To calibrate such a model, parameters constraints must be satisfied to ensure model identification—that is, $\sum_{t=t_0}^{t_n} k_t = 0$ and $\sum_{x=0}^{\omega} b_x = 1$—and the maximum likelihood procedure is employed to achieve the estimates $\hat{\alpha}_x$, $\hat{\beta}_x$, and $\hat{k}_t$ (Brouhns, Denuit, and Vermunt 2002). To introduce the network model for forecasting purposes, let $\boldsymbol{\kappa}_{\mathcal{T}} = (k_{t-1}, k_{t-2}, ..., k_{t-j})$ be the time lagged k_t series, where $j \in \mathbb{N}$ is the time lag. According to Equation (3), we model the LC time index as follows:

$$k_t = f_{LSTM}(\boldsymbol{\kappa}_{\mathcal{T}}; \mathcal{W}) + \gamma_t = \psi \circ \left(H^{(P)}_t \circ H^{(P-1)}_t \circ \cdots \circ H^{(1)}_t\right)(\boldsymbol{\kappa}_{\mathcal{T}}; \mathcal{W}) + \gamma_t. \tag{6}$$

Integrating Equation (6) within the LC structure in Equation (5), the LSTM will act as a predictor over the forecasting horizon $\mathcal{T}' = \{t_n + 1, t_n + 2, ..., t_n + s\}$, so that the LC-LSTM model expression is

$$\log m_{x,t} = \hat{\alpha}_x + \hat{\beta}_x (f_{LSTM}(\boldsymbol{\kappa}_{\mathcal{T}}; \mathcal{W}) + \gamma_t), \quad \forall t \in \mathcal{T}'. \tag{7}$$

The meaning of the proposed model integration is the following. Because the mortality dynamic over time stems from a continuous evolution of various social and demographic factors (Oeppen and Vaupel 2006), a coherent mortality profile investigation suggests an autoregressive approach to the time index modeling. From a general perspective, the LC time index values should be interpreted as the realization of the following process:

$$k_t = \varphi(\boldsymbol{\kappa}_{\mathcal{T}}) + \gamma_t, \tag{8}$$

where the unknown function $\varphi : \mathbb{R}^j \to \mathbb{R}$ maps the series $\boldsymbol{\kappa}_{\mathcal{T}}$ to k_t, without considering the noise component. Referring to RNNs' universal functional approximation property (Schäfer and Zimmermann 2007), the proposed model integration is based on the approximation of the unknown map $\varphi(\boldsymbol{\kappa}_{\mathcal{T}})$ through an RNN with LSTM architecture, whose functional form is shaped according to the available time index history. As the RNN model approximates the map $\varphi(\boldsymbol{\kappa}_{\mathcal{T}})$, it also estimates the conditional mean of response variable (Geman, Bienenstock, and Doursat 1992); that is,

$$\hat{k}_t = \hat{f}_{LSTM}(\boldsymbol{\kappa}_{\mathcal{T}}; \hat{\mathcal{W}}) = \hat{\mathbb{E}}(k_t | \boldsymbol{\kappa}_{\mathcal{T}}), \tag{9}$$

where $\hat{f}_{LSTM}$ is the fitted function composition and $\hat{\mathcal{W}}$ is the NN parameters estimate. Consequently, the LC-LSTM model provides the following point predictions:

$$\log \hat{m}_{x,t} = \hat{\alpha}_x + \hat{\beta}_x \hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}; \hat{\mathcal{W}}\right), \quad \forall t \in \mathcal{T}'. \tag{10}$$

We stress that point predictions do not describe the uncertainty arising from the estimates of mortality rates. Therefore, a methodology for building prediction intervals is necessary in order to provide a measure of prediction uncertainty.

4. PREDICTION INTERVALS FOR THE LC-LSTM MODEL

Prediction intervals (henceforth PIs) outline a probabilistic range suitable for incorporating various forecasting scenarios and then probing uncertainty on the future mortality realizations. Stochastic mortality models forecast PIs, whose estimates act as uncertainty measures linked to the expected future mortality (see, for instance, Booth and Tickle 2008; Cairns et al. 2009, 2011; Dowd et al. 2010). Thus, in a proper forecasting process, PIs are meaningful in supporting both risk evaluation and the model estimate's reliability. Referring to NNs, the construction of PIs is a challenging task because different uncertainty sources impact the learning process and conditioning the NN generalization performance. From a broader perspective, NN models are exposed to a learning uncertainty depending on both the data and the NN functioning. Because the data employed in the learning process are realizations of an underlying stochastic process, a training data uncertainty looms. Indeed, an input variation could involve different function compositions. In addition, a certain variability could arise from the optimization procedure necessary to learn NN parameter values from data. Because the cost function could exhibit many local minima, the NN parameters take on different values entailing variability in estimates. In this case, a parameter uncertainty emerges. Nevertheless, also model uncertainty could occur for possible structural model misspecification.

Addressing the measurement of uncertainty sources separately is a complex problem, because they are closely connected and no information is available about the input–output relation. However, PIs account for all uncertainty sources, embracing the overall variability around NN point predictions. Therefore, we proceed to define PIs for the LC-LSTM mortality rates in order to estimate the total uncertainty produced by the proposed model integration. More specifically, the uncertainty in death rates over time concerns the time index dynamic described in Equation (6), wherein $\boldsymbol{\kappa}_{\mathcal{T}}$ follows some distribution $\mathbb{K}$. Because the network model draws a predictor $\hat{f}_{LSTM}$ for the purpose of approximating the future k_t values, a natural way to approach the PI construction involves the bias–variance trade-off principle. According to Geman, Bienenstock, and Doursat (1992), a measure to depict both accuracy and variability of $\hat{f}_{LSTM}$ as a predictor of the (unseen) k_t is given by the mean squared error of prediction, defined in terms of bias–variance decomposition (see Appendix A for the proof):

$$\mathbb{E}\left[\left(k_t - \hat{k}_t\right)^2\right] = \mathbb{E}\left[\left(\mathbb{E}\left(\hat{k}_t\right) - \hat{k}_t\right)^2\right] + \mathbb{E}\left[\left(\varphi(\boldsymbol{\kappa}_{\mathcal{T}}) - \mathbb{E}\left(\hat{k}_t\right)\right)^2\right] + \sigma^2_{\gamma_t}, \tag{11}$$

where $\sigma^2_{\hat{k}_t} := \mathbb{E}\left[\left(\mathbb{E}\left(\hat{k}_t\right) - \hat{k}_t\right)^2\right]$ and $\mathbb{E}\left[BIAS^2\left(\hat{k}_t\right)\right] := \mathbb{E}\left[\left(\varphi(\boldsymbol{\kappa}_{\mathcal{T}}) - \mathbb{E}\left(\hat{k}_t\right)\right)^2\right]$ are, respectively, the variance and the expected squared bias related to the LSTM. The former stems from the network calibration process, hence including uncertainty due to both training data and learned weights; the latter summarizes the effectiveness of $\hat{f}_{LSTM}$ in approximating the true (regression) function φ in Equation (8). In compliance with the bias–variance principle, both bias and variance contribute to the NN prediction error and the NN model suitability is based on the balancing of both. Finally, the variance σ^2_{γ} constitutes an irreducible term of uncertainty, because it refers to the random noise component. We emphasize that expectations in Equation (11) range over different realizations of $\boldsymbol{\kappa}_{\mathcal{T}}$ by the virtue of the unknow distribution $\mathbb{K}$.

4.1. Estimating $\sigma^2_{\hat{k}_t}$

To compute the variance related to the NN output, the conditional distribution $\mathbb{P}(\hat{k}_t | \boldsymbol{\kappa}_{\mathcal{T}})$ should be known. However, it is not available, and we could either hypothesize some distribution or extract it from the data grasped. Considering the latter, our approach to estimate the variance $\sigma^2_{\hat{k}_t}$ refers to the NN ensemble paradigm, which is based on the joint use of multiple NNs (Zhou, Wu, and Tang 2002). Exploiting a bootstrap procedure, multiple training data samples are generated in order to develop an empirical distribution constitute by different NN point predictions. The final estimates are then obtained by aggregating the average of various NN projections. Such a procedure, namely, bootstrap aggregating or bagging (Breiman 1996), is an ensemble technique producing an unbiased estimation and favoring an adequate variance measurement. Therefore, the expected bias in Equation (11) is seen as a negligible component affecting the overall time index uncertainty (Geman, Bienenstock, and Doursat 1992; Khosravi et al. 2011). The bagging scheme proposed in the present work is described in the following steps:

*Step 1.*Conforming to Equation (9), we firstly train the LSTM model on the training data, $\boldsymbol{\kappa}_{\mathcal{T}}$, to obtain the point estimates $\hat{f}_{LSTM}(\boldsymbol{\kappa}_{\mathcal{T}};\hat{\mathcal{W}})$ over the forecast horizon $\mathcal{T}'$.

*Step 2.*We generate $B \in \mathbb{N}$ samples of the training data through a proper bootstrap procedure, getting $\left(\boldsymbol{\kappa}_{\mathcal{T}}^{(b)}, b = 1, ..., B\right)$. In particular, we consider the residual bootstrap strategy proposed in Koissi, Shapiro, and Hognas (2006), whose technical details are reported in Appendix B.

*Step 3.*For each sample $\boldsymbol{\kappa}_{\mathcal{T}}^{(b)}$, we re-optimize the weights of the function composition $\hat{f}_{LSTM}$ defined in Step 1, so that only the NN weights will change given the new training data. Hence, the created NN ensemble will include uncertainty for both training data and parameters.

*Step 4.*For each retrained NN in Step 3, we predict the associate point estimate on $\mathcal{T}'$, producing a bootstrap distribution consisting of B point predictions; that is,

$$\hat{\mathbb{P}}\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}}\right) = \left(\hat{k}_t^{(b)} = \hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}^{(b)}, \hat{\mathcal{W}}^{(b)}\right), b = 1, ..., B\right). \tag{12}$$

*Step 5.*From $\hat{\mathbb{P}}\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}}\right)$, we can determine the bagged estimate for the variance $\sigma^2_{\hat{k}_t}$; that is,

$$\hat{\sigma}^2_{\hat{k}_t} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}^{(b)}, \hat{\mathcal{W}}^{(b)}\right) - \bar{k}_t\right), \tag{13}$$

where $\bar{k}_t = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}^{(b)}, \hat{\mathcal{W}}^{(b)}\right)$ is the bagged estimate for the conditional expectation $\mathbb{E}\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}}\right)$.

We emphasize that when using an ensemble technique for estimating the NN output variance, the expected bias component is irrelevant. Thus, the ensemble technique could associate high uncertainty to the NN predictions, as the bias–variance tradeoff states. However, if the employed bootstrap technique fits the density estimation problem and the trained NN model is robust, then the estimated variance does not induce explosive prediction intervals behavior over time.

4.2. Estimating σ^2_γ

The noise variance represents an irreducible risk, reflecting the randomness in predicting k_t values through a deterministic function (the LSTM) applied to the past realizations. Thus, the mortality profile incorporates an intrinsic randomness not taken into account by the network model, although such noise does not affect, on average, the point predictions (see Equation [9]). Indeed, properly trained NNs learn the key input–output data schemes, skimming noisy examples and avoiding overfitting occurrences. Considering the training set interval $\mathcal{T}$, we can observe both the available time index values and the predictions provided by the network. Hence, dealing with the series $\hat{\gamma}_t = (k_t - \hat{k}_t, t \in \mathcal{T})$ as realization of the unwrapped noise by NN, we consider the sample variance, $\hat{\sigma}^2_\gamma = \frac{1}{(t_n - t_0) - 1}\sum_{t=t_0}^{t_n}\hat{\gamma}_t$, as an estimate of the time index irreducible uncertainty over $\mathcal{T}$. For the purposes of PI construction, we can finally set the variance estimate for the time index as $\hat{\sigma}^2_{k_t} = \hat{\sigma}^2_{\hat{k}_t} + \hat{\sigma}^2_\gamma$, because of the independence between the network function and the noise. It worth noting that if $\hat{\gamma}_t$ shows gaussianity features, we can spread the noise component over the forecast horizon through a Gaussian random walk, and the estimated PI boundaries for a confidence level α are

$$\left[\hat{k}_t - z_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2_{\hat{k}_t} + \hat{\sigma}^2_\gamma}\, , \, \hat{k}_t + z_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2_{\hat{k}_t} + \hat{\sigma}^2_\gamma}\right] \tag{14}$$

where z_α is the α-quantile of a standard normal distribution.

5. PERFORMANCE METRICS OF FORECASTING

To quantitatively assess the LC-LSTM projections over the forecast horizon, we refer to performance metrics for both point and interval forecasts. In the former case, the root mean square error (henceforth RMSE) is acknowledged as an accuracy measure for both the time index and mortality rates, respectively, as

$$RMSE_{(k)} = \sqrt{\frac{\sum_{t=t_n+1}^{t_n+s}\left(k_t - \hat{k}_t\right)^2}{s-1}}, \quad RMSE_{(m)} = \sqrt{\frac{\sum_{t=t_n+1}^{t_n+s}\left(\log m_{x,t} - \log \hat{m}_{x,t}\right)^2}{s-1}}. \tag{15}$$

To judge the PI quality and effectiveness, we jointly examine PI coverage probability and PI width. In analytical terms, we consider two indicators, namely, the prediction interval coverage probability (henceforth PICP) and the mean prediction interval width (henceforth MPIW). The former inspects the PI coverage counting how many values are wrapped in the probabilistic range, given a confidence level. In other words, the PICP estimates the probability that the mortality rate values fall within the PI provided by the mortality model. Let $\hat{k}_t^L$ be the estimated time-index lower bound and $\hat{k}_t^U$ the estimated time index upper bound. Then, the PICP for the k_t series is defined as follows:

$$PICP_{(k)} = \frac{1}{s-1}\sum_{t=t_n+1}^{t_n+s} \mathbf{1}_{\left\{\hat{k}_t \in \left[\hat{k}_t^L, \hat{k}_t^U\right]\right\}}, \tag{16}$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function such that $\mathbf{1}_{\{\cdot\}} = 1$ if $\hat{k}_t \in [\hat{k}_t^L, \hat{k}_t^U]$, and $\mathbf{1}_{\{\cdot\}} = 0$ otherwise.

The MPIW indicates the PI mean width over the forecasting horizon; that is,

$$MPIW_{(k)} = \frac{1}{s-1}\sum_{t=t_n+1}^{t_n+s} \hat{k}_t^U - \hat{k}_t^L. \tag{17}$$

We also calculate PICP and MPIW on the log-mortality rates by a given age x. Let $\log \hat{m}_{x,t}^L$ be the estimated mortality rates' lower bound and be $\log \hat{m}_{x,t}^U$ the estimated mortality rates' upper bound. Then, we specify the PICP and MPIW as follows:

$$PICP_{(m)} = \frac{1}{s-1}\sum_{t=t_n+1}^{t_n+s} \mathbf{1}_{\left\{\log \hat{m}_{x,t} \in \left[\log \hat{m}_{x,t}^L,\ \log \hat{m}_{x,t}^U\right]\right\}}, \tag{18}$$

where $\mathbf{1}_{\{\cdot\}} = 1$ if $\log \hat{m}_{x,t} \in [\log \hat{m}_{x,t}^L,\ \log \hat{m}_{x,t}^U]$ and $\mathbf{1}_{\{\cdot\}} = 0$ otherwise, and

$$MPIW_{(m)} = \frac{1}{s-1}\sum_{t=t_n+1}^{t_n+s} \log \hat{m}_{x,t}^U - \log \hat{m}_{x,t}^L. \tag{19}$$

A higher PICP value indicates PIs having a greater probability to cover the true mortality realizations. High MPIW values are desirable in order to provide a suitable uncertainty portrayal. Nonetheless, an explosive demeanor in variability is reflected by greater MPIW levels, jeopardizing the biological plausibility of mortality forecasts. The latter qualitative criterion is valuable because it concerns the predicted uncertainty levels consistency w.r.t. the historical volatility at different ages (Cairns et al. (2011)).

6. EMPIRICAL INVESTIGATION AND RESULTS

In the following, we illustrate the empirical analysis carried out to test our model proposal. The results and considerations presented will also take into account the forecasts obtained from the LC Poisson model (Brouhns, Denuit, and Vermunt 2002) as a term of comparison. The equations defining the LC Poisson predictions are reported in Appendix C. Our analysis was performed using R software (R Core Team 2020, version 3.6.3) and the packages StMoMo (Villegas, Kaishev, and Millossovich 2018, version 0.4.1), forecast (Hyndman and Khandakar 2008, version 8.13), Keras (Chollet 2017, version 2.2.5), and Tensorflow (Abadi et al. 2015, version 1.13.1).

6.1. Data

Aiming to portray heterogeneous longevity scenarios, we performed our numerical experiment for three countries worldwide, Australia, Japan, and Spain, analyzed by gender. We consider such countries representative in terms of both demographic transition and population structure. Data were downloaded from the Human Mortality Database (2018) and refer to the age range $\mathcal{X} = \{0, 1, ..., 99\}$. We consider two calendar year sets, 1950–2018 and 1960–2018, to assess both accuracy and variability of the LC-LSTM outcomes with respect to the historical time chunks. This allows us to verify the effect on the learning process of shortening the NN training set; that is, the network's robustness to changes in the training set length.

6.2. Neural Network Tuning, Training, and Ensembling

To apply the LSTM model, we first need to calibrate the LC structure in Equation (5) on the age–period mortality data, estimating both age-dependent and time-dependent parameters. The latter constitutes the series $(k_t, t = t_0, ..., 2018)$, with $t_0 = \{1950, 1960\}$, to implement the network learning process. We tune and train the LSTM model splitting the time index series into distinct datasets by a hierarchical procedure. In particular, setting $T = 2000$ as the forecasting year for all countries investigated, we define the training set and the testing set as follows

$$\begin{array}{ll} \text{TRAINING SET}: & \mathcal{TR} = (k_t, t = t_0, ..., 2000) \\ \text{TESTING SET}: & \mathcal{TS} = (k_t, t = 2001, ..., 2018). \end{array} \tag{20}$$

In addition, to validate the LSTM model, we divide the training set into a training subset and a validation set, considering the splitting rule $80\% - 20\%$. Hence, denoting by T^{sub} the last year in the training subset we set:

$$\begin{array}{ll} \text{TRAINING SUBSET}: & \mathcal{TR}^{\text{sub}} = \left(k_t, t = t_0, ..., T^{\text{sub}}\right) \\ \text{VALIDATION SET}: & \mathcal{VS} = \left(k_t, t = T^{\text{sub}} + 1, ..., 2000\right) \end{array}. \tag{21}$$

According to Equation (6), we consider a lag $j = 1$ so that $k_t = f_{LSTM}(k_{t-1}; \mathcal{W}) + \gamma_t$; that is, the LSTM network sifts the mortality profile at annual paces.

We use the sets $\mathcal{TR}^{\text{sub}}$ and $\mathcal{VS}$ to tune the NN structure through a grid search technique. Thus, we set a bounded discrete parametric space whose possible values are arbitrarily chosen, acting as network hyperparameter. Fixing a hyper-parameters combination, the learning process begins minimizing the mean square error cost function over the set $\mathcal{TR}^{\text{sub}}$. We select as an optimal NN structure the one identified by the hyperparameter combination returning the minimum error on the validation set $\mathcal{VS}$. In doing so, the function composition, $\hat{f}_{LSTM}$, is built according to the data. Such an NN architecture is then employed on the training set, $\mathcal{TR}$, to generate point predictions over the testing set horizon. Therefore, we compare the NN forecasts, $\hat{k}_t$, with the available time index values in $\mathcal{TS}$ as a back-testing exercise. We highlight that, for each country and for both genders, the LSTM model is characterized by $p = 1$ hidden layer, considering the ReLu function (Nair and Hinton 2010) as a feed-forward activation function, the tangent hyperbolic function as a recurrent activation function, and the linear function as the output layer activation function ψ. The number N_p of hidden neurons varies depending on both country and gender. The depicted learning process suggests the minimum learning period length to produce robust predictions. Shortening the training dataset, our experiment highlights that training periods beginning after the 1960s generate predictions sensitive to small variations in the data. Therefore, we need at least 40 observations to adequately tune the network model.

For the purposes of PI construction, the tuned network architecture acts as reference model in Step 1 of the proposed bagging scheme in Subsection 4.1. Consequently, following the bootstrap strategy proposed in Koissi, Shapiro, and Hognas (2006), we generate $B = 1000$ bootstrap samples of the training set $\mathcal{TR}$. Maintaining the tuned network function composition, $\hat{f}_{LSTM}$, we re-optimize its weights on the bth training set producing the related forecasts over the testing horizon. Therefore, the bootstrap distribution is obtained allowing for the bagged variance calculation as in Equation (13).

6.3. Results

In the following, we provide the results of our numerical application, recalling the performance metrics presented in Section 5. We first refer to the RMSE metric to evaluate the point forecast accuracy, considering the error of the LC projections as benchmark. To appreciate the PIs' quality by PICP and MPIW indicators, after the bagging scheme, we assess the noise variance in order to estimate PI boundaries. We consider the sample variance of $\hat{\gamma}_t = (k_t - \hat{k}_t, t \in \mathcal{TR})$ as the noise variance estimate over the training set. To project the noise and its uncertainty over the testing horizon, we inspect its possible random walk behavior. To this end, the augmented Dickey-Fuller (ADF) test is implemented. In addition, we test normality features of the noise realizations through statistical tests, such as Shapiro-Wilk, D'Agostino-Pearson, and Jarque-Bera. For all investigated countries and for both genders, the noise analysis confirms the ability of a random walk representation with Gaussian innovations for the noise component (see Appendix D). Therefore, the LC-LSTM time index values are embedded within the following PI, for a confidence level α:

$$\left[\hat{k}_t^L, \hat{k}_t^U\right] = \left[\hat{k}_t - z_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2_{\hat{k}_t} + \hat{\sigma}^2_{\gamma}}, \hat{k}_t + z_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2_{\hat{k}_t} + \hat{\sigma}^2_{\gamma}}\right],$$

where z_α is the α-quantile of a standard normal distribution.

We then calculate the performance metrics for the LC-LSTM and the LC model. Their values for the time index are provided in Table 1, comparing the LSTM performances in the LC-LSTM, with the ARIMA in the LC model.

For all countries considered, the time index series observed since the 1960s exhibits a markable linear decline over time. In particular, mortality reductions accelerated over the period 1950–1960, and an approximately constant rate of degrowth characterized the interval 1960–2000. Such a behavior was driven by a decline in infant mortality, as well as reductions in mortality at older ages after World War II (see, for instance, Rau et al. (2008)).

As a general statement about prediction accuracy, our analysis confirms the ability of the ARIMA to represent linear evolution in mortality. On the other hand, the LSTM seems to be advisable for linear, noisy, or non-linear series. Scrutinizing the uncertainty results, the LSTM always offers greater probability coverage, in most cases due to the PI width. Because the LSTM point predictions present low bias, their variance tends to be increasing and to be higher than those of ARIMA.

The majority of cases promote the LSTM model's usefulness in affording a more accurate mortality trend, as well as for uncertainty estimation. The best example concerns Australian males, presenting the lower RMSE for the period 1960–2000. Considering the training period 1950–2000, the NN allows the simultaneous presence of total coverage of the future k_t realizations and a proper PI width. This situation appears also when reducing the training set length; that is, considering the interval 1960–2000. A suitable mortality dynamic for the ARIMA model is offered by Japanese females. In fact, their mortality behavior presents a strong linear decrease over time, also when observed from 1950. In this circumstance, the LSTM learns a too steep trend of mortality reductions, as opposed to ARIMA. However, switching to the training period 1960–2000, the network performance improves significantly. We observe a gain of 67.7% in RMSE terms, maintaining at the same time both a total probability coverage and a coherent MPIW value. On the other hands the ARIMA model does not favor a reliable uncertainty estimation in both periods. Its coverage probability is around 50%, indicating that the predictive model fails to anticipate, on average, half of the future realizations. An analogous result holds for Spanish males, whose time index dynamic shows a noisier series over both training periods. Indeed, the ARIMA coverage probability for Spanish males remains stable around 33%.

We also depict the mortality profile for both genders considering ages 45, 65, and 85. To explore these results, we display the performance metrics in Table 2 and the PI graphs in Figure 1 and Figure 2. We can highlight the estimated PIs for the LC-LSTM model in terms of both point and interval estimates. Looking at the Japanese population, we endorse the findings in Table 1 for ages 45 and 65. The LC-LSTM provides boundaries properly shaped according to death rates, whereas the LC model presents the narrowest ranges of variability lacking uncertainty information. For example, over the training period 1960–2000 for Japanese females age 65, the PIs for the LC model show a coverage probability around 33%, whereas the LC-LSTM provides $PICP_{(m)} = 1$ with a similar interval width. For age 85, where mortality reductions present slower linear changes over time, the LC fits the future mortality profile. For the Spanish population, the LC-LCTM seems to be the best-fitting model for predictive purposes. As reported in Table 1, for this country, as the training period shifts, the MPIW value for k_t identifies a significant reduction in the PI width (–20.56% for males and –53.38% for females), although full probability coverage is maintained. Such a reduction affects the uncertainty measurement in the LC-LSTM model, although the PI width wider than that of the LC model. Finally, we stress how both the LC and the LC-LSTM model fail to catch the non-linear mortality

TABLE 1

k_t Performance Metrics Values for Each Training Period. Forecasting Years: 2001–2018

Country	Model	Training period 1950–2000						Training period 1960–2000					
		Male			Female			Male			Female		
		RMSE	$PICP_{(k)}$	$MPIW_{(k)}$	*RMSE*	$PICP_{(k)}$	$MPIW_{(k)}$	*RMSE*	$PICP_{(k)}$	$MPIW_{(k)}$	*RMSE*	$PICP_{(k)}$	$MPIW_{(k)}$
Australia	ARIMA	9.514	1	53.503	3.861	1	25.195	5.138	1	47.485	3.637	1	25.089
	LSTM	4.280	1	32.865	3.790	1	39.478	1.970	1	28.143	2.659	1	37.433
Japan	ARIMA	3.743	1	21.503	10.084	0.556	20.767	4.647	1	17.392	9.790	0.500	12.409
	LSTM	2.228	1	43.784	18.014	1	53.431	2.069	1	28.209	5.818	1	30.701
Spain	ARIMA	14.038	0.333	19.354	6.215	1	21.394	13.071	0.333	17.343	5.805	1	20.747
	LSTM	8.625	1	35.424	7.471	1	60.373	9.983	0.778	23.340	4.357	1	28.141

TABLE 2

$\log m_{x,t}$ Performance Metrics Values for each Training Period, Forecasting Years: 2001–2018

x = 45

Country	Model	Training period 1950-2000						Training period 1960-2000					
		Male			Female			Male			Female		
		$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$
Australia	LC	0.227	**1**	***0.534***	***0.091***	0.944	0.267	0.175	**1**	***0.478***	***0.084***	0.944	0.265
	LC-LSTM	***0.110***	0.944	0.295	0.142	0.944	***0.407***	***0.116***	0.944	0.280	0.097	**1**	***0.394***
Japan	LC	0.071	0.667	***0.180***	0.255	0	0.173	***0.063***	0.722	0.150	0.155	0.056	0.105
	LC-LSTM	***0.062***	***0.722***	0.143	***0.077***	***0.444***	***0.254***	0.073	0.944	0.243	***0.061***	***0.667***	***0.115***
Spain	LC	0.200	0.333	0.153	***0.104***	0.611	0.179	0.228	***0.333***	0.136	***0.067***	0.722	0.174
	LC-LSTM	***0.161***	***0.556***	***0.276***	0.502	***0.944***	***0.489***	***0.205***	0.278	***0.215***	0.073	***0.944***	***0.259***

x = 65

Country	Model	Training period 1950-2000						Training period 1960-2000					
		Male			Female			Male			Female		
		$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$
Australia	LC	0.157	1	***0.672***	0.061	0.944	0.283	0.106	1	0.623	0.058	1	0.293
	LC-LSTM	***0.056***	1	0.371	0.061	**1**	***0.431***	***0.043***	1	0.365	0.052	1	0.436
Japan	LC	0.054	**1**	***0.177***	0.160	0.444	0.178	0.063	0.833	0.161	0.151	0.333	0.128
	LC-LSTM	***0.035***	0.944	0.141	0.077	1	0.262	0.029	1	0.261	0.028	1	0.141
Spain	LC	0.097	0.278	0.157	0.079	0.778	0.206	0.106	0.222	0.158	0.073	0.889	0.229
	LC-LSTM	***0.060***	**1**	***0.285***	0.066	1	0.568	0.080	0.889	0.249	0.068	0.944	0.340

x = 85

Country	Model	Training period 1950-2000						Training period 1960-2000					
		Male			Female			Male			Female		
		$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$
Australia	LC	0.053	0.944	0.344	0.032	1	0.191	0.039	0.944	0.319	0.033	1	0.194
	LC-LSTM	0.056	0.944	0.190	0.033	1	0.292	0.049	0.944	0.187	0.026	1	0.289
Japan	LC	0.030	0.889	0.134	0.050	0.778	0.142	0.040	0.944	0.133	0.071	0.444	0.115
	LC-LSTM	0.034	0.778	0.107	0.171	0.500	0.209	0.029	0.944	0.215	0.080	0.444	0.126
Spain	LC	0.082	0.333	0.113	0.059	0.611	0.122	0.086	0.278	0.116	0.057	0.833	0.150
	LC-LSTM	0.052	1	0.204	0.447	1	0.335	0.066	0.944	0.183	0.048	1	0.223

pattern characterizing age 45 over the testing horizon. Starting from the 2000s, Spanish males aged 45 have experienced a notable acceleration in the rate of mortality reduction. Because we use $T = 2000$ as the forecasting year, the extrapolation approach underlying both the LC and the LC-LSTM induces misleading projections. Finally, we appreciate the LC model's uncertainty estimation performance for Australian males. We highlight the LC model’s greatest probability coverage and interval width. Nevertheless, the latter hints at some questions about the LC model prediction’s suitability in the long run. See, for instance, Figure 3, which displays a 50-year prediction for the Australian males aged 65 for both training periods.

Given the observed mortality up to the forecasting year, the LC model seems to propose uncertainty levels not consistent with the historical mortality dynamics. Looking at the training period 1960–2000, we observe an overall reduction in death rates of about 61%. In the following 40 years of projection, the LC model estimates a further reduction in death rates of around 96% in the case of the PI lower bound or a possible increase of 68% for the PI upper bound. For the training period 1950–2000, this evidence is strengthened. Referring to the LC-LSTM model, the mortality estimates assume greater consistency with historical observations. In particular, the LC-LSTM produces a 40-year decrease in mortality between 82% for the PI lower bound and 46% for the PI upper bound.

Moreover, inspecting Figure 3, we stress how the learning period length impacts the long-run network forecasts. As aforementioned, the two learning periods considered show different accelerations in mortality decline. Fitting the LSTM model on

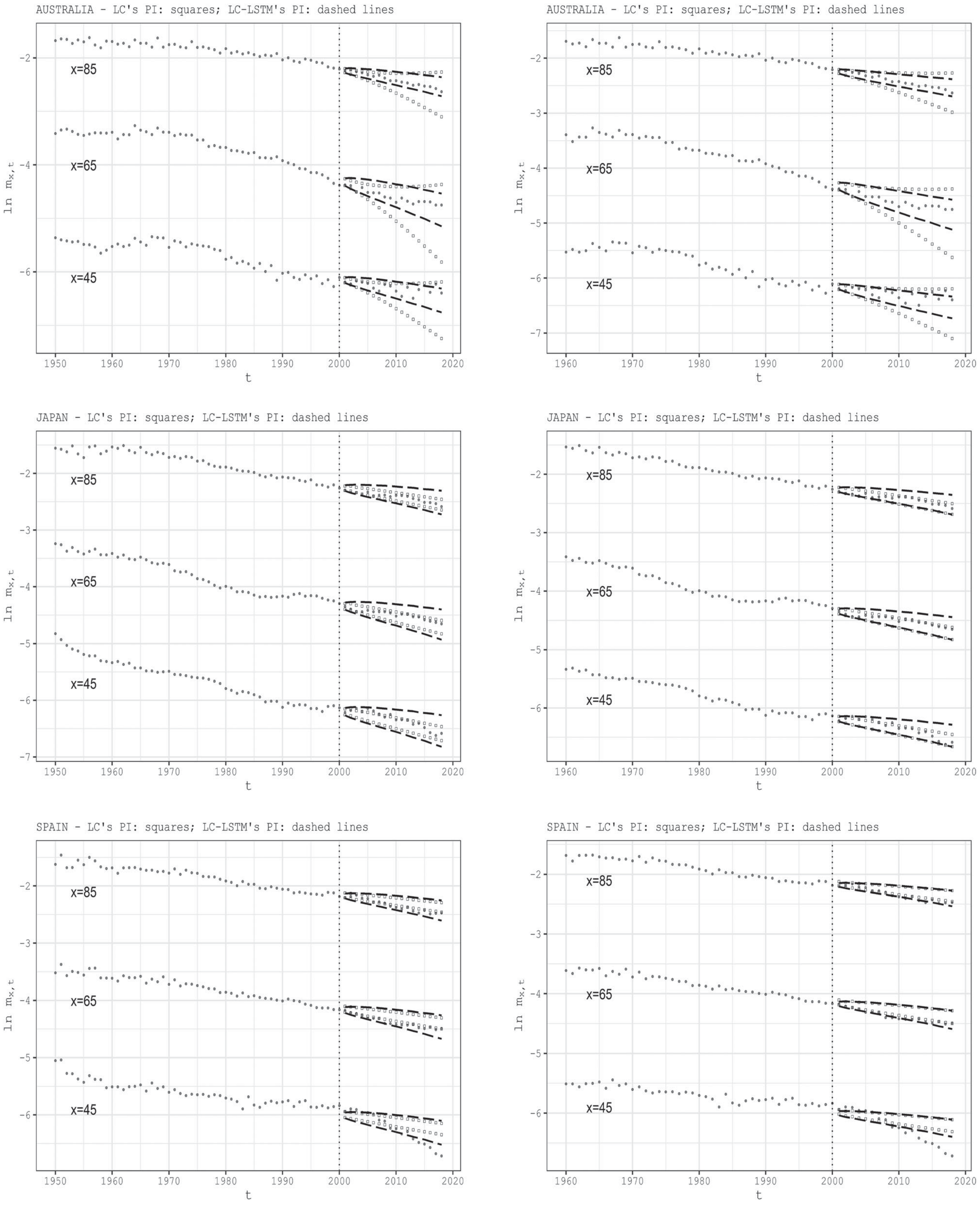

FIGURE 1. Male PI ($\alpha = 5\%$). Forecasting Period: 2001–2018. *Note:* Training period: 1950–2000 (left), 1960–2000 (right).

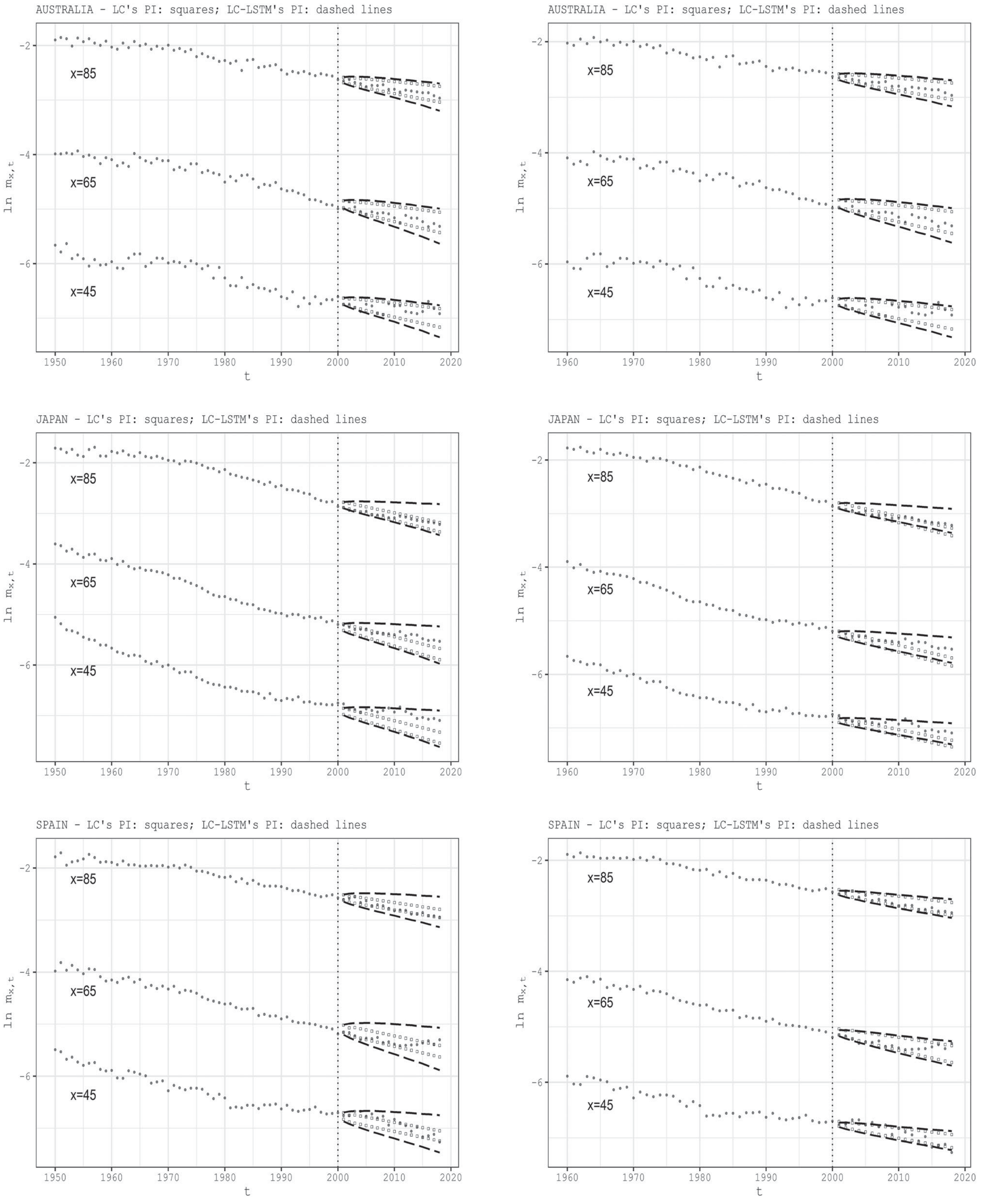

FIGURE 2. Female PI ($\alpha = 5\%$). Forecasting Period: 2001–2018. *Note:* Training period: 1950–2000 (left), 1960–2000 (right).

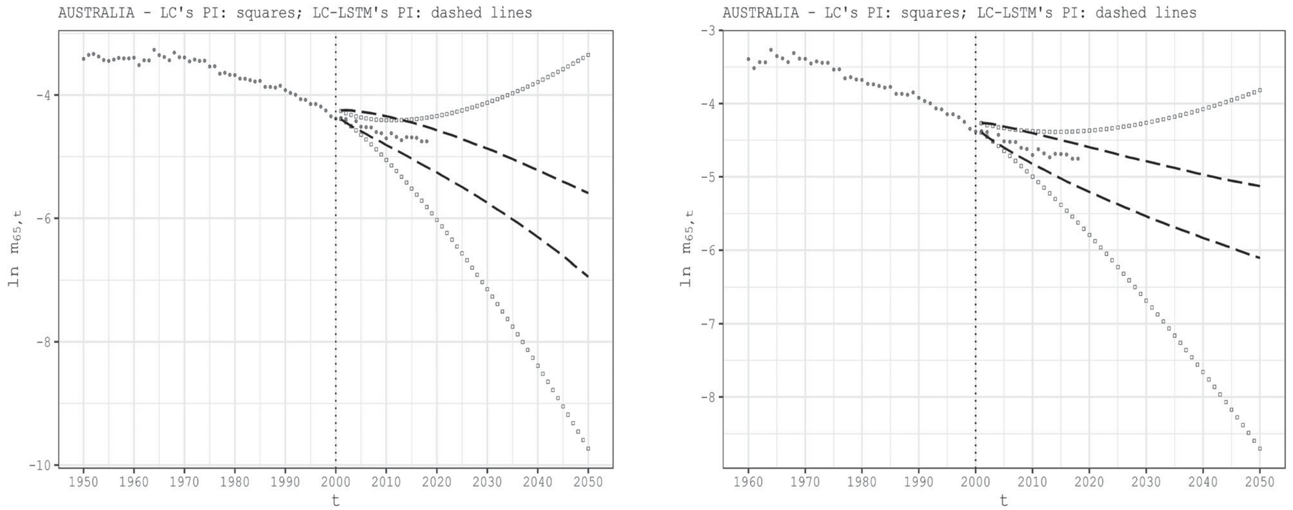

FIGURE 3. Australian Males. PI ($\alpha = 5\%$) for $x = 65$, Forecasting Period: 2001–2050. *Note:* Training period: 1950–2000 (left), 1960–2000 (right).

the interval 1960–2000, the network learns the fundamental linear decrease of mortality such that a coherent PI shape is predicted over the forecasting horizon. In contrast the interval 1950–2000 points to a nonlinear behavior due to the longevity accelerations in the period 1950–1960. In this case, the LSTM is able to extrapolate a coherent mortality range with the historical observation, allowing for biological plausibility but a more marked increase in longevity. In light of this, we do not question the robustness of the model; rather, we emphasize its ability to extrapolate the fundamental pattern from the observed data. The selection of the historical sample on which to fit the mortality model depends on the aware modeler's expert judgment, given the population under investigation. As suggested by Cairns et al. (2011), it is crucial to evaluate qualitative ex ante criteria, such as biological reasonableness, the plausibility of predicted levels of uncertainty, and model robustness. At the same time, ex post quantitative criteria, such as performance metrics in Section 5, are indispensable to address forecasts in a back-testing exercise (see, for instance, Dowd et al. 2010). Following both qualitative and quantitative criteria, our analyses demonstrate how, overall, both models are biologically regular in projecting mortality. The discriminating factor between the two models is the plausibility of foreseen uncertainty levels, especially for long-term forecasts. Hence, our model improves the prediction level of the LC model, as proven in most cases by the performance indicators. Finally, we suggest the interval 1960–2000 as the most proper training period for the LSTM calibration on mortality data. In fact, it is plausible to believe that the reduction in mortality will continue to occur in a fairly linear way over time and at different ages, properly reflecting the demographic trend observed since the 1960s.

7. CONCLUSIONS

Mortality forecasting is still a major challenge for actuaries and demographers. Obviously, different populations might show diverse mortality scenarios, and a well-performing mortality model for one population might be not adequate for another one. Not surprising, the collection of mortality models in the literature is extensive. Recently, new methodological advances in mortality forecasting have been proposed, grounded on machine and deep learning techniques, mainly based on NN models. The present work formalizes a deep learning integration of the LC model, in terms of both point prediction and prediction interval. Our proposal allows representing the mortality surface through a canonical age–period model and predicting the future mortality realizations by extrapolating the temporal mortality dynamics from data. The resulting LC-LSTM model provides a compromise between the interpretation of the mortality phenomenon and high precision in anticipating its future realizations. Moreover, exploiting both the NN ensemble paradigm and noise analysis, we are able to produce a mortality density forecast. From our empirical investigation, we highlight the LC-LSTMs capacity to produce forecasts both biologically consistent and plausible in uncertainty levels w.r.t. the historical observations, also in the long run. The latter feature is crucial in actuarial assessments, especially in the evaluation of annuity products or to appraise pension systems sustainability. Therefore, our proposal establishes a reliable improvement of the LC model in terms of predictive ability posing an innovative approach within mortality literature. The proposed framework might represent a prominent practice in the field of longevity forecasting and for actuarial business tasks.

ACKNOWLEDGMENTS

The authors thank the reviewers for their insightful comments.

FUNDING

Mario Marino was supported by the funded project Actuarial and Financial Risk Management Solutions in a Pandemic Mortality Framework (Sapienza University Funding No. RM120172B84AD5EC).

REFERENCES

Abadi, M., A. Aggarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. 2015. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. https://www.tensorflow.org/

Aggarwal, C. C. 2018. *Neural networks and deep learning. A textbook*. Cham: Springer

Booth, H., and L. Tickle. 2008. Mortality modelling and forecasting: A review of methods. *Annals of Actuarial Science* 3 (1–2):3–43. doi:10.1017/S1748499500000440

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24:123–40. doi:10.1007/BF00058655

Brouhns, N., M. Denuit, and I. van Keilegom. 2005. Bootstrapping the Poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal* 3:212–24. doi:10.1080/03461230510009754

Brouhns, N., M. Denuit, and J. Vermunt. 2002. A Poisson log-bilinear regression approach to the construction of projected life tables. *Insurance: Mathematics and Economics* 3:373–93.

Cairns, A. J. G., D. Blake, and K. Dowd. 2006. A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance* 73:687–718. doi:10.1111/j.1539-6975.2006.00195.x

Cairns, A. J. G., D. Blake, K. Dowd, G. D. Coughlan, D. Epstein, and M. Khalaf-Allah. 2011. Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics* 48 (3):355–67. doi:10.1016/j.insmatheco.2010.12.005

Cairns, A. J. G., D. Blake, K. Dowd, G. D. Coughlan, D. Epstein, A. Ong, and I. Balevich. 2009. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal* 13:1–35. doi:10.1080/10920277.2009.10597538

Camarda, C. G. 2019. Smooth constrained mortality forecasting. *Demographic Research* 41:1091–130. doi:10.4054/DemRes.2019.41.38

Carney, J. G., P. Cunningham, and U. Bhagwan. 1999. Confidence and prediction intervals for neural network ensembles. *In Proceeding of the International Joint Conference on Neural Networks*, 1215–18. Washington DC, U.S.: IEEE Computer Society Press.

Chollet, F. 2017. *R Interface to Keras*, GitHub, https://github.com/rstudio/keras

Currie, I. D. 2017. On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal* 4:356–83. doi:10.1080/03461238.2014.928230

Currie, I. D., M. Durban, and P. H. C. Eilers. 2004. Smoothing and forecasting mortality rates. *Statistical Modelling* 4:279–98. doi:10.1191/1471082X04st080oa

D'Amato, V., E. Di Lorenzo, S. Haberman, M. Russolillo, and M. Sibillo. 2011. The Poisson log-bilinear Lee-Carter model. *North American Actuarial Journal* 15 (2):315–33. doi:10.1080/10920277.2011.10597623

D'Amato, V., S. Haberman, G. Piscopo, and M. Russolillo. 2012. Modelling dependent data for longevity projections. *Insurance: Mathematics and Economics* 51 (3):694–701. doi:10.1016/j.insmatheco.2012.09.008

D'Amato, V., S. Haberman, and M. Russolillo. 2012. The stratified sampling bootstrap for measuring the uncertainty in mortality forecasts. *Methodology Computing in Applied Probability* 14:135–48. doi:10.1007/s11009-011-9225-z

Deprez, P., P. V. Shevchenko, and M. V. Wüthrich. 2017. Machine learning techniques for mortality modeling. *European Actuarial Journal* 7:337–52. doi:10.1007/s13385-017-0152-4

Dowd, K., D. Blake, A. J. G. Cairns, G. D. Coughlan, D. Epstein, and M. Khalaf-Allah. 2010. Backtesting stochastic mortality models: An ex-post valuation of multi-period-ahead density forecasts. *North American Actuarial Journal* 14:281–98. doi:10.1080/10920277.2010.10597592

Efron, B., and R. Tibshirani. 1993. *An introduction to the bootstrap*. New York: Chapman & Hall.

Geman, S., E. Bienenstock, and R. Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural Computation* 4 (1):1–58. doi:10.1162/neco.1992.4.1.1

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep learning*. Cambridge, MA, U. S.: MIT Press.

Hainaut, D. 2018. A neural-network analyzer for mortality forecast. *ASTIN Bulletin* 48 (2):481–508. doi:10.1017/asb.2017.45

Heskes, T. 1997. Practical confidence and prediction intervals. In *Advances in neural information processing systems*, vol. 9, 176–82. Cambridge, MA, U. S.: MIT Press.

Hocreiter, S., and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9 (8):1735–80. doi:10.1162/neco.1997.9.8.1735

Human Mortality Database. 2018. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). https://www.mortality.org.

Hunt, A., and D. Blake. 2014. A general procedure for constructing mortality models. *North American Actuarial Journal* 18 (1): 116–38. doi:10.1080/10920277.2013.852963

Hyndman, R. J., and Y. Khandakar. 2008. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software* 26 (3):1–22. doi:10.18637/jss.v027.i03

Kasiviswanathan, K. S., and K. P. Sudheer. 2013. Quantification of the predictive uncertainty of artificial neural network based river flow forecast models. *Stochastic Environmental Research and Risk Assessment* 27:137–46. doi:10.1007/s00477-012-0600-2

Khosravi, A., S. Nahavandi, D. Creighton, and A. F. Atiya. 2011. Comprehensive review of neural network–based prediction intervals and new advances. *IEEE Transactions on Neural Networks* 22 (9):1341–56. doi:10.1109/TNN.2011.2162110

Khosravi, A., S. Nahavandi, D. Srinivasan, and R. Khosravi. 2015. Constructing optimal prediction intervals by using neural networks and bootstrap method. *IEEE Transactions on Neural Networks and Learning Systems* 26 (8):1810–15. doi:10.1109/TNNLS.2014.2354418

Koissi, M., A. Shapiro, and G. Hognas. 2006. Evaluating and extending the Lee-Carter model for mortality forecasting confidence interval. *Insurance: Mathematics and Economics* 1:1–20.

Lee, R. D., and L. R. Carter. 1992. Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association* 87 (419):659–71. doi:10.1080/01621459.1992.10475265

Levantesi, S., and A. Nigri. 2020. A random forest algorithm to improve the Lee-Carter mortality forecasting: Impact on q-forward. *Soft Computing* 24: 8553–67. doi:10.1007/s00500-019-04427-z

Levantesi, S., A. Nigri, and G. Piscopo. 2021. Clustering-based simultaneous forecasting of life expectancy time series through long-short term memory neural networks. *International Journal of Approximate Reasoning* 140:282–97. doi:10.1016/j.ijar.2021.10.008

Levantesi, S., and V. Pizzorusso. 2019. Application of machine learning to mortality modeling and forecasting. *Risks* 7 (1):26.

Li, J., M. Hardy, and K. Tan. 2009. Uncertainty in mortality forecasting: An extension to the classical Lee-Carter approach. *ASTIN Bulletin* 39 (1):137–64. doi:10.2143/AST.39.1.2038060

Li, K., R. Wang, H. Lei, T. Zhang, Y. Liu, and X. Zheng. 2018. Interval prediction of solar power using an improved bootstrap method. *Solar Energy* 159: 97–112. doi:10.1016/j.solener.2017.10.051

MacKay, D. J. C. 1992. A practical Bayesian framework for backpropagation networks. *Neural Computation* 4 (3):448–72. doi:10.1162/neco.1992.4.3.448

Makridakis, S., E. Spiliotis, and V. Assimakopoulos. 2020. The M4 Competition: 100,000 Time series and 61 forecasting methods. *International Journal of Forecasting* 36 (1):54–74. doi:10.1016/j.ijforecast.2019.04.014

Mazloumi, E., G. Rose, G. Currie, and S. Moridpour. 2011. Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction. *Engineering Applications of Artificial Intelligence* 24 (3):534–42. doi:10.1016/j.engappai.2010.11.004

Mitchell, D., P. Brockett, R. Mendoza-Arriaga, and K. Muthuraman. 2013. Modeling and forecasting mortality rates. *Insurance: Mathematics and Economics* 52 (2):275–85. doi:10.1016/j.insmatheco.2013.01.002

Nair, V., and G. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 807–14. Madison, WI: Omnipress.

Nigri, A., E. Barbi, and S. Levantesi. 2021. The relationship between longevity and lifespan variation. *Statistical Methods and Applications.* doi:10.1007/s10260-021-00584-4

Nigri, A., S. Levantesi, and M. Marino. 2020. Life expectancy and lifespan disparity forecasting: A long short-term memory approach. *Scandinavian Actuarial Journal* 2:110–33.

Nigri, A., S. Levantesi, M. Marino, S. Scognamiglio, and F. Perla. 2019. A deep learning integrated Lee-Carter model. *Risks* 7 (1):33.

Nix, D. A., and A. S. Weigend. 1994. Estimating the mean and variance of the target probability distribution. *In Proceeding of IEEE International Conference on Neural Networks*, vol. 1, 55–60. Washington DC, U. S.: IEEE Computer Society Press.

Oeppen, J., and J. W. Vaupel. 2006. The linear rise in the number of our days. In *Perspectives on mortality forecasting. III: The linear rise in life expectancy: history and prospects*, vol. 3, ed. T. Bengtsson, 9–18. Stockholm: Swedish Social Insurance Agency.

Pascanu, R., T. Mikolov, and Y. Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of Machine Learning Research* 28 (3): vol. 28, 1310–18.

Perla, F., R. Richman, S. Scognamiglio, and M. Wüthrich. 2021. Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal* 7:572–98.

Plat, R. 2009. On stochastic mortality modeling. *Insurance: Mathematics and Economics* 45:393–404. doi:10.1016/j.insmatheco.2009.08.006

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rau, R., E. Soroko, D. Jasilionis, and J. W. Vaupel. 2008. Continued reductions in mortality at advanced ages. *Population and Development Review* 34: 747–68. doi:10.1111/j.1728-4457.2008.00249.x

Renshaw, A. E., and S. Haberman. 2006. A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics* 38 (3):556–70. doi:10.1016/j.insmatheco.2005.12.001

Richman, R., and M. V. Wüthrich.. 2021. A neural network extension of the Lee-Carter model to multiple populations. *Annals of Actuarial Science* 15 (2): 346–66.

Rumelhart, D., G. Hinton, and R. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323:533–36. doi:10.1038/323533a0

Schäfer, A. M., and H. G. Zimmermann. 2007. Recurrent neural networks are universal approximators. *International Journal of Neural Systems* 17 (4): 253–63. doi:10.1142/S0129065707001111

Schnürch, S., and R. Korn. 2021. Point and interval forecasts of death rates using neural networks. *ASTIN Bulletin* 8 (52):333–60.

Tibshirani, R. 1996. A comparison of some error estimates for neural network models. *Neural Computation* 8:152–63. doi:10.1162/neco.1996.8.1.152

Villegas, A. M., V. K. Kaishev, and P. Millossovich. 2018. StMoMo: An R Package for Stochastic Mortality Modeling. *Journal of Statistical Software* 84 (3):1–38. doi:10.18637/jss.v084.i03

Wild, C. J., and G. A. F. Seber. 1989. *Nonlinear regression*. New York: Wiley.

Zhou, Z. H., J. Wu, and W. Tang. 2002. Ensembling neural networks: Many could be better than all. *Artificial Intelligence* 137 (1–2):239–63. doi:10.1016/S0004-3702(02)00190-X

APPENDIX A

To prove Equation (11), we recall the following set of assumptions:

$$\begin{cases} k_t = \varphi(\boldsymbol{\kappa}_{\mathcal{T}}) + \gamma_t \\ \mathbb{E}(\gamma_t) = 0 \\ \sigma_\gamma^2 = \mathbb{E}(\gamma_t^2) \\ \hat{k}_t = \hat{f}\left(\boldsymbol{\kappa}_{\mathcal{T}}; \hat{\mathcal{W}}\right) \\ \gamma_t \perp \mathrm{k_t},\ \gamma_\mathrm{t} \perp \hat{\mathrm{k}}_\mathrm{t} \end{cases} \tag{A.1}$$

It is straightforward to define the bias–variance decomposition for the mean squared error as:

$$\begin{aligned} \mathbb{E}\left[\left(k_t - \hat{k}_t\right)^2\right] &= \mathbb{E}\left[\left(\varphi(\boldsymbol{\kappa}_{\mathcal{T}}) + \gamma_t \pm \mathbb{E}\left(\hat{k}_t\right)\right)^2\right] = \\ &= \mathbb{E}\left[\left(\varphi(\boldsymbol{\kappa}_{\mathcal{T}}) - \mathbb{E}\left(\hat{k}_t\right)\right)^2\right] + \mathbb{E}\left(\gamma_t^2\right) + \mathbb{E}\left[\left(\mathbb{E}\left(\hat{k}_t\right) - \hat{k}_t\right)^2\right] \\ &\quad + 2\mathbb{E}\left[\gamma_t\left(\varphi(\boldsymbol{\kappa}_{\mathcal{T}}) - \mathbb{E}\left(\hat{k}_t\right)\right)\right] + 2\mathbb{E}\left[\gamma_t\left(\mathbb{E}\left(\hat{k}_t\right) - \hat{k}_t\right)\right] \\ &\quad + 2\mathbb{E}\left[\left(\mathbb{E}\left(\hat{k}_t\right) - \hat{k}_t\right)\left(\varphi(\boldsymbol{\kappa}_{\mathcal{T}}) - \mathbb{E}\left(\hat{k}_t\right)\right)\right] = \\ &= \mathbb{E}\left[\left(\mathbb{E}\left(\hat{k}_t\right) - \hat{k}_t\right)^2\right] + \mathbb{E}\left[\left(\varphi(\boldsymbol{\kappa}_{\mathcal{T}}) - \mathbb{E}\left(\hat{k}_t\right)\right)^2\right] + \sigma_{\gamma_t}^2, \end{aligned} \tag{A.2}$$

where $\sigma_{\hat{k}_t}^2 := \mathbb{E}\left[\left(\mathbb{E}\left(\hat{k}_t\right) - \hat{k}_t\right)^2\right]$ and $\mathbb{E}\left[BIAS^2\left(\hat{k}_t\right)\right] := \mathbb{E}\left[\left(\varphi(\boldsymbol{\kappa}_{\mathcal{T}}) - \mathbb{E}\left(\hat{k}_t\right)\right)^2\right]$, completing the proof.

APPENDIX B

We summarize the residual bootstrap procedure proposed by Koissi, Shapiro, and Hognas (2006), to which we refer in our bagging scheme. For the sake of clarity, in the following we indicate by $\hat{k}_t$ the maximum likelihood estimate for the time index. Given such an estimate and a lag j, we built the training data $\boldsymbol{\kappa}_{\mathcal{T}}$ for the LSTM model. Thus, to obtain the bootstrap samples of the training data—that is, $\boldsymbol{\kappa}_{\mathcal{T}}^{(b)}$—we need to sample the $\hat{k}_t$ estimates. To this end, the following steps are performed:

i. Compute the matrix of Poisson deviance residuals:

$$r_D = \mathrm{sign}\left(D_{x,t} - \hat{D}_{x,t}\right)\sqrt{D_{x,t}\ln\left(\frac{D_{x,t}}{\hat{D}_{x,t}} - \left(D_{x,t} - \hat{D}_{x,t}\right)\right)}, \quad x \in \mathcal{X},\ t \in \mathcal{T}, \tag{B.1}$$

where $\hat{D}_{x,t} = E_{x,t}^c \exp\left(\hat{\alpha}_x + \hat{\beta}_x\hat{k}_t\right)$ are the fitted number of deaths.

ii. Sample with replacement the elements of r_D to generate B replications; that is, $\left(r_D^{(b)}, b = 1, \ldots, B\right)$.

iii. Invert each bth residual matrix to define the corresponding matrix of death counts. To this end, it is necessary to find the matrix $\hat{D}_{x,t}^{(b)}$ solving the following equation:

$$\hat{D}_{x,t}^{(b)} - D_{x,t}\ln\left(\hat{D}_{x,t}^{(b)}\right) - \left(r_D^{(b)}\right)^2 - D_{x,t} + D_{x,t}\ln D_{x,t} = 0, \tag{B.2}$$

whose solutions are numerically obtained through the Newton-Raphson method.

iv. Given the matrices of death counts, $\left(\hat{D}^{(b)}_{x,t}, b = 1, \ldots, B\right)$, the maximum likelihood procedure is executed B times to re-estimate the LC model parameters, so that is possible to get the bootstrap samples of the time index; that is, $\left(\hat{k}^{(b)}_{t}, b = 1, \ldots, B\right)$.

v. Considering the lag j, from each bootstrap sample $\hat{k}^{(b)}_{t}$ it is straightforward to derive the bootstrapped training data employed in the bagging scheme; that is, $\boldsymbol{\kappa}^{(b)}_{\mathcal{T}}$.

APPENDIX C

For the sake of comparison, we briefly recall the fundamental forecasting equations concerning the LC model. According to the LC structure in Equation (5), the time index is usually projected through a random walk with drift. Generalizing, in our experiment we consider an ARIMA(p,d,q) process, so that the realizations of k_t over $\mathcal{T}'$ originate from the following equation:

$$\nabla^{d} k_{t_n+h} = h\delta + \sum_{i=1}^{p} \phi_i \nabla^{d} k_{(t_n+h)-i} + \sum_{j=1}^{q} \theta_j \epsilon_{(t_n+h)-j} + \sum_{k=1}^{h} \epsilon_{t_n+k}, \quad h = 1, \ldots, s, \tag{C.1}$$

where δ is the drift parameter and ϕ and θ are the coefficients for the autoregressive terms and for the moving average terms, respectively. In addition, the sum of errors are normally distributed; that is, $\sum_{k=1}^{h} \epsilon_{t_{n+k}} \sim \mathcal{N}\left(0, h^2\sigma^2_{\epsilon}\right)$. Under this modeling framework, the LC model point predictions are

$$\log \hat{m}_{x,t_n+h} = \hat{\alpha}_x + \hat{\beta}_x \quad h\hat{\delta} + \sum_{i=1}^{p} \hat{\phi}_i \nabla^{d} k_{(t_n+h)-i} + \sum_{j=1}^{q} \hat{\theta}_j \epsilon_{(t_n+h)-j} \Bigg). \tag{C.2}$$

Because errors are Gaussian, the estimated prediction interval boundaries at a fixed confidence level $\alpha \in (0, 1)$ are given by

$$\log \hat{m}^{U,L}_{x,t_n+h} = \log \hat{m}_{x,t_n+h} \pm \hat{\beta}_x \sqrt{h}\hat{\sigma}_{\epsilon} z_{\frac{\alpha}{2}}, \tag{C.3}$$

where $\hat{\sigma}_{\epsilon}$ is the error's variance estimate, and z_{α} is the α-quantile of a standard normal distribution. In Table C.1 we show the best ARIMA(p,d,q) models applied in our emprical investigation, for each country and for both genders, where such models are calibrated according to the Hyndman-Khandakar algorithm (Hyndman and Khandakar 2008).

TABLE C.1
Best Selected ARIMA(p,d,q) Models for k_t.

Country	Gender	Training period 1950–2000	Training period 1960–2000
Australia	Male	ARIMA (0,2,2)	ARIMA (1,2,1)
	Female	ARIMA (1,1,0)	ARIMA (1,1,0)
Japan	Male	ARIMA (2,1,0)	ARIMA (0,1,1)
	Female	ARIMA (0,1,1)	ARIMA (0,1,1)
Spain	Male	ARIMA (0,1,1)	ARIMA (1,1,0)
	Female	ARIMA (0,1,3)	ARIMA (1,1,0)

APPENDIX D

TABLE D.1
Statistical Tests for Noise in the Training Set: Males

Country	Test	Training period 1950–2000		Training period 1960–2000	
		Statistics value	p value	Statistics value	p value
Australia	Shapiro-Wilk	.96352	.12489***	0.98379	.82539***
	D’Agostino-Pearson	1.62692	.44332***	0.85534	.65203***
	Jarque-Bera	1.55177	.46030***	0.64381	.72477***
	ADF	–3.05447	.15132***	–2.58739	.34294***
Japan	Shapiro-Wilk	0.96193	.10710***	0.97511	.51356***
	D’Agostino-Pearson	8.05556	.01781*	1.45996	.48192***
	Jarque-Bera	7.35771	.02525*	1.20406	.54770***
	ADF	–3.49574	.05128**	–2.73088	.28662***
Spain	Shapiro-Wilk	0.97654	.41696***	0.95790	.14191***
	D’Agostino-Pearson	1.83229	.40006***	2.82652	.24335***
	Jarque-Bera	1.05350	.59052***	2.31446	.31436***
	ADF	–7.55942	.01000	–4.11879	.01516*

Note: Significant at *$p > .01$, **$p > .05$, ***$p > .1$.

TABLE D.2
Statistical Tests for Noise in the Training Set. Females

Country	Test	Training period 1950–2000		Training period 1960–2000	
		Statistics value	p value	Statistics value	p value
Australia	Shapiro-Wilk	0.96907	.21209***	0.96724	.29319***
	D’Agostino-Pearson	2.52531	.28290***	0.78319	.67598***
	Jarque-Bera	1.78204	.41024***	0.60740	.73808***
	ADF	–3.07190	.14432***	–2.50033	.37711***
Japan	Shapiro-Wilk	0.97452	.34985***	0.98888	.95815***
	D’Agostino-Pearson	3.12195	.20993 ***	0.79814	.67094***
	Jarque-Bera	2.09605	.35063***	0.62112	.73303***
	ADF	–5.14239	.01000	–3.89596	.02383*
Spain	Shapiro-Wilk	0.93640	.02619*	0.97970	.67844***
	D’Agostino-Pearson	8.69754	.01292*	1.74855	.41716***
	Jarque-Bera	7.56206	.02280*	1.20753	.54675***
	ADF	–5.80177	.01000	–3.46488	.06172***

Note: Significant at *$p > .01$, **$p > .05$, ***$p > .1$.