SAPIENZA
UNIVERSITÀ DI ROMA

Mortality neural forecasting

School of Statistical Sciences

Ph.D in Actuarial Sciences – XXXIII Cycle

Candidate

Mario Marino
ID number 1405678

Thesis Advisor

Prof. Susanna Levantesi

March 2021

Thesis defended on 12 July 2021
in front of a Board of Examiners composed by:

Prof. Roy Cerqueti (chairman)

Prof. Pietro Millossovich

Prof. Emilio Russo

Mortality neural forecasting
Ph.D. thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Version: 12/07/2021

Author's email: m.marino@uniroma1.it

To my grandmothers, to my mentors and, finally, to me.

Abstract

Predicting mortality is a major challenge for both demographers and actuaries. The latter need to anticipate various future mortality scenarios with the greatest possible accuracy, as in the case of annuities pricing and longevity risk assessments. However, the current wide range of stochastic mortality models highlights some deficiencies in predicting future mortality realizations, particularly when accelerations or decelerations of longevity occur. The aim of this research thesis is to investigate the adequacy of a new mortality forecasting approach based on artificial Neural Networks. To this end, after an examination of the theoretical Neural Networks fundamentals, the present work shows the Neural Networks competitiveness in predicting the future dynamics of human mortality, also allowing the efficacy of already existing predictive models, such as the canonical Lee-Carter model. Therefore, our data-driven proposal contributes to the mortality literature as new advance in mortality forecasting, that is the neural forecasting approach.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays, the continuous increase in longevity has become a proven fact worldwide. Population death dynamics have shown the beneficial effects of the lack of worst living conditions, as in the extreme case of wars or shortages in the medical and social sciences. Indeed, developments in medicine, public health, and socio-economic attitudes over the past century have led to unprecedented extensions of life expectancy around the world. Considering also reductions in fertility, the resulting global population ageing presents new opportunities, as well as challenges. In particular, financial challenges emerge for life insurers, pension plans and social security schemes involving benefits related to illness, death or survival of people. For all these entities, is crucial understand the demographic nature of longevity and contextualize it into business processes.

Currently, longevity improvements are moved by multiple, heterogeneous factors, such as nutrition, education, lifestyles, pollution levels, technological advances diseases treatment, and so on (Riley (2001)). How each of these factors impact on mortality is difficult to analyse, also requiring a biometric parameters granularity that is hardly available. Thus, actuarial modelling typically involves models to describe mortality by an aggregate perspective, looking at the insured population or the national one. By a stochastic approach to projections, a widely used basis to anticipate mortality is the class of extrapolative models, starting from the pioneering Lee–Carter model (Lee and Carter (1992)). In their original paper, Lee and Carter (1992) present a log-bilinear relation both to explain and to predict the age-period death rates. Consequently, the Lee-Carter model establishes life expectancy forecast that first increases at the historical trend and then decelerates over time. Such a lifespan demeanour seems to be reasonable. However, as was introduced by Oeppen and Vaupel (2002), the life expectancy has increased fairly linearly for more than 150 years, breaking predictions and limits figured out by actuaries. Several model variants and extensions has been proposed in literature in order to boost the Lee-Carter proposal (see, for instance, Brouhns et al. (2002), Renshaw and Haberman (2006) and Cairns et al. (2009)), reaching more actual mortality scenarios. Undeniably, Lee-Carter model improvements has been implied the addition of parameters within the mortality model, leading with parameter identification problems. Therefore, practitioners and researchers often apply models with a simpler parametric form, being able to guarantee both robustness and biological reasonableness upon projected

life tables (Cairns et al. (2011)). However, while longevity continues to increase over time, it is changing its rate of growth (Debonneuil et al. (2018)) and some future non-linear mortality behaviour may occur. In some extent, the need of accurate mortality forecasts remains nowadays a compelling task.

The present research work is born to accept this predictive challenge. The main objective of the research carried out is to increase the explanatory capacity of predictive analysis on mortality, by resorting a new modelling approach: the artificial Neural Networks. The latter was born as computational system in order to reproduce how the biological brain works. The artificial Neural Networks models are based on the cognitive paradigm to calculus: given a set of examples, the network learns their relations and try to generalize them to predict new occurrences. The Neural Network learning process stems from well posed theoretical studies (see for example Vapnik (1999)), formalizing mathematically how the network is shaped according to the data. Basically, the Neural Networks are data-driven tools, as well as the overall class of machine and deep learning techniques, whose way of working meets mathematical and statistical postulates. In light of this, Neural Network are flexible models capable to catch hidden features within data, representing the fundamental functional relations describing the inspected phenomenon. In other words, as new advance in mortality forecasting applications we will investigate the suitability of Neural Networks models to discover patterns within mortality data and reproduce them on designed forecasting horizon.

The reminder of the present work is the following.

Chapter 2. We will briefly recall the analytical foundations necessary to probabilistically study the mortality dynamics. In particular, we will focus both on the indicators and on the mortality models that will be the subject of further elaborations in the course of this thesis;

Chapter 3. With our goal in mind, a necessary examination of the Neural Networks foundations is presented. Specifically, we aim to offer to the reader a harmonious guide within the Neural Networks, reaching the explanation of the model to which we will refer to investigate the future mortality: the Recurrent Neural Networks with a Long Short Term Memory architecture. At the state of the art, the latter identifies the best performing non-hybrid tool in the field of deep learning forecast;

Chapter 4. We will illustrate our first research analysis on the Neural Network suitability in predicting mortality. Such a study will concern the prediction of life expectancy and lifespan disparity, both independently and simultaneously. The research carried out makes a new contribution to the relevant literature, both in terms of a new predictive approach and for the joint prediction of lifespan indicators at birth and age 65. We emphasize that Chapter 4 is an excerpt from the following peer-reviewed research work:

- Nigri, A., Levantesi, S. and Marino M. (2020). Life expectancy and lifespan disparity forecasting: a long short-term memory approach.

Scandinavian Actuarial Journal.
DOI:10.1080/03461238.2020.1814855.

Chapter 5. As a further investigative analysis, we plan to study the predictive capacity gain over canonical mortality models by mixing them with Neural Network. In doing so, we propose the concept of model integration within the Lee-Carter model framework, both about the accuracy of expected mortality trend and its uncertainty. Our contribution to the mortality literature is twofold: on the one hand, we introduce the use of Neural Networks in projection phase, replacing the use of the classic time series models, such as the ARIMA; on the other hand, we propose a methodology for estimating the uncertainty of future mortality realizations stemming from a data-driven model. The Chapter 5 is extrapolated from the following peer-reviewed research works:

- Nigri, A., Levantesi, S., Marino, M., Scognamiglio, S., Perla, F. (2019). A Deep Learning Integrated Lee–Carter Model, *Risks* 7(1): 33. DOI:10.3390/risks7010033
- Marino M., Levantesi S. (2020). Forecasting Neural Network Lee-Carter model with parameter uncertainty: the case of Italy. *Appearing to Mathematical and Statistical Methods for Actuarial Sciences and Finance*.

Finally, the present work considers a final discussion in **Chapter 6** about the results achieved, emphasizing further suitable research projects for mortality forecasting through Neural Networks.

Chapter 2

Mortality modeling

We provide a dutiful recall of mortality modelling characterization. In details, we summon the fundamental framework about demographic variables and prominent discrete time stochastic mortality models employed to fashion and predict mortality by ages and over time.

2.1 Demographic variables

Whatsoever mortality analysis founds on an essential mathematical framework to formalize the main demographic variables to model mortality. As stressed in Pitacco (2004) and Pitacco et al. (2010), an age-period mortality representation is necessary in a dynamic context. Hence, considered biometric variables shall be expressed as a function of both age $x \in \mathcal{X} = \{0, 1, \ldots, \omega\}$ and the calendar year $t \in \mathcal{T} = \{t_0, t_1, \ldots, t_n\}$, where $0 \leq t_0 < t_1 < \cdots < t_n < \infty$. We remind that ω is the so called *extreme age*, i.e. the maximum attainable age in life by an individual.
Let us denote as $T_0(t-x)$ the random lifetime of a newborn in the calendar year $(t-x)$ and let $S_t(x)$ the survival function, i.e. the probability that $T_0(t-x)$ is longer than x attained in t:

$$S_t(x) = \mathbb{P}\left(T_0(t-x) > x\right). \tag{2.1}$$

Furthermore, let us denote as $\mu_{x,t}$ the instantaneous death rate, or force of mortality, for an individual aged x at time t. Assuming $S_t(x)$ differentiable with respect to x, the following holds:

$$\mu_{x,t} = -\frac{\partial}{\partial x} \ln S_t(x), \tag{2.2}$$

so that, solving Eq.(2.2) with initial condition $S_t(0) = 1$, we have:

$$S_t(x) = \exp\left(-\int_0^x \mu_{s,t} ds\right). \tag{2.3}$$

In doing so, death and survival probabilities can be determined recognizing the force of mortality. Indeed, the probability that an individual aged x at time t dies

before age $x+h$, with $h>0$, is:

$$\begin{aligned} {}_hq_{x,t} &= \mathbb{P}\left(x < T_0(t-x) \leq x+h | T_0(t-x) > x\right) = \\ &= \frac{S_t(x) - S_t(x+h)}{S_t(x)} = \\ &= 1 - \exp\left(-\int_x^{x+h} \mu_{s,t} ds\right), \end{aligned} \tag{2.4}$$

and the probability to alive at age $x+h$, for the same individual, is:

$${}_hp_{x,t} = \exp\left(-\int_x^{x+h} \mu_{s,t} ds\right). \tag{2.5}$$

Often, especially in actuarial assessments, the force of mortality is assumed to be a piece-wise constant function, that is:

$$\mu_{x+k,t+h} = \mu_{x,t}, \quad k,h \in [0,1) \tag{2.6}$$

and, consequently, death and survival probabilities become, respectively:

$${}_hq_{x,t} = 1 - \exp\left(-h\mu_{x,t}\right), \quad {}_hp_{x,t} = \exp\left(-h\mu_{x,t}\right). \tag{2.7}$$

Additionally, we can describe the force of mortality behaviour over the interval $(x, x+h)$ exploiting the so called mortality coefficient $m_{(x,x+h),t}$. It is defined as the age-continuous weighted mean of $\mu_{x,t}$ in $(x, x+h)$, where the survival function acts as weighting function, that is:

$$m_{(x,x+h),t} = \frac{\int_x^{x+h} \mu_{s,t} S_t(s) ds}{\int_x^{x+h} S_t(s) ds}. \tag{2.8}$$

From Eq.(2.8), posing $h=1$, stems the central death rate $m_{x,t}$, so outlined:

$$m_{x,t} = \frac{S_t(x) - S_t(x+1)}{\int_x^{x+1} S_t(s) ds}. \tag{2.9}$$

Under assumption (2.6), it is straightforward notes that $\mu_{x,t} = m_{x,t}$.
To appraise the central death rate, awareness of survival function is crucial. However, it is demonstrated that (see, for example, Pitacco et al. (2010)) the maximum likelihood estimate for $m_{x,t}$, and for $\mu_{x,t}$ under assumption (2.6), is the following:

$$\hat{m}_{x,t} = \frac{D_{x,t}}{E^c_{x,t}} \tag{2.10}$$

where $D_{x,t}$ is the number of deaths occurred in year t among the people aged x and $E^c_{x,t}$ is the central exposed to risk, namely the average number of the living people aged x in t. The estimate in Eq.(2.10) plays an important role within discrete time stochastic mortality models framework, as we will recall in Section 2.2.

2.1.1 Some keys to summarize lifetime distribution

The random lifetime distribution investigation allows to summarize key information about mortality. In particular, location measures, such as expectation, offer a first insight.
Let us denote as $T_x(t)$ the random lifetime of an individual aged x in the calendar year t. We indicate with $e_{x,t} = \mathbb{E}(T_x(t))$ the life expectancy at age x in t and it is defined as follows:

$$e_{x,t} = \frac{\int_x^\omega S_t(s)ds}{S_t(x)}. \tag{2.11}$$

Life expectancy is generally wielded to compare mortality of various populations. However, populations characterized by the same level of life expectancy could experience substantial differences in the time of death (Aburto et al. (2020)), with different age-at-death distributions. Being a location measure, life expectancy is not likely to detect variations in lifespan, which are instead captured by lifespan disparity allowing to describe variations in lifespan distribution (Bohk-Ewald et al. (2017))[1]. While life expectancy has been proved to hide heterogeneity in individual mortality paths, lifespan disparity measures the dispersion of observations around the time of death, evaluating from, respectively, a probabilistic and a descriptive point of view, uncertainty in age-at-death distribution and heterogeneity (van Raalte et al. (2018), Kaakai et al. (2019)). When mortality is highly variable, some individuals will die at a much younger age than the expected age-at-death, contributing many lost years to life disparities; conversely, when mortality is highly concentrated around older ages or the modal age, life disparity decreases (Aburto and van Raalte (2018)).
Therefore, is useful formalize the lifespan disparity measure, $e^\dagger_{x,t}$, representing the life expectancy lost due to death by an individual aged x at time t

$$e^\dagger_{x,t} = \frac{\int_x^\omega e_{s,t} \cdot D_t(s)ds}{S_t(x)}, \tag{2.12}$$

where $D_t(x)$ is the law governing death occurrences at age x and at time t.

A in depth life expectancy and lifespan disparity examination, mostly looking at their forecasting issues, will be explored in Chapter 4.

2.2 Discrete time stochastic mortality models

Different methods has been proposed in actuarial and demographic literature to fashion and project mortality. For forecasting reasons, predictive models based on an extrapolation procedures are commonly used. Generally speaking, a family of discrete time stochastic mortality models, namely generalized age-period-cohort (GAPC, from now on) models, can be mentioned as in Currie (2017), Hunt, Blake (2015) and Villegas et al. (2015). By a statistical point of view, the GAPC family

[1] In addition to life disparity, other inequality measures have been proposed in literature, e.g. the Gini coefficient and the Keyfitz's entropy (Wilmoth and Horiuchi (1999), Shkolnikov et al. (2003), van Raalte and Caswell (2013)) that appear to be linearly related and negatively correlated to life expectancy at birth (Colchero et al. (2016), Nemeth (2017) and Aburto et al. (2020)).

mirrors the wider class of generalized non-linear models, and several stochastic mortality models in literature can be embedded within this family.
A GAPC stochastic mortality models assumes:

- the death counts, $D_{x,t}$, as a random component described by a probability distribution falling in the overdispersed exponential class. In practice, the death counts, $D_{x,t}$, follow a Poisson distribution (see, for example, Brouhns et al. (2002)):
$$D_{x,t} \sim Poi(E^c_{x,t} m_{x,t}); \tag{2.13}$$

- a predictor, $\eta_{x,t}$, viz. a systematic component catching effects of ages, calendar years and years of birth on mortality dynamics:
$$\eta_{x,t} = \alpha_x + \sum_{i=1}^{N} \beta_x^{(i)} k_t^{(i)} + \beta_x^{(0)} \gamma_{t-x} \tag{2.14}$$
where:
  - α_x is an age-dependent parameter depicting the age-mortality shape;
  - $\beta_x^{(i)}$ is the i^{th} age-dependent parameter portraying age-specific sensitivity to mortality behaviour over time;
  - $k_t^{(i)}$ is the i^{th} time-dependent parameter describing a peculiar mortality trend over time;
  - γ_{t-x} reports the cohort effect, with $\beta_x^{(0)}$ inflecting its influence across ages.

  The GAPC family supposes that time indexes parameters, $k_t^{(i)}$ and γ_{t-x}, are stochastic processes. Hence, their modeling provides probabilistic forecasts achievements of future mortality rates;

- a link function g associating the expectation of the random component to the predictor:
$$g\left(\mathbb{E}\left(\frac{D_{x,t}}{E^c_{x,t}}\right)\right) = \eta_{x,t}. \tag{2.15}$$
Under Poisson distribution assumption for the death counts, the link function employed is the canonical one, that is the log link function $g(\cdot) = \ln(\cdot)$;

- a set of parameter constraints ensuring a unique parameters estimate.

Furthermore, models in the GAPC framework can be divided into following subclasses:

- the Lee and Carter (LC, from now on) class. It relies on the pioneering study presented in Lee and Carter (1992) and linked variants and extensions after submitted in literature, such as Booth et al. (2002), Brouhns et al. (2002), Hyndman and Ullah (2007), Li and Miller (2005), and Renshaw and Haberman (2006).

- the Cairns, Blake and Dowd (CBD, from now on) class. Unlike the LC class, the CBD one exploits a combination of multi-periods and cohort effects to design the predictor, without an age-dependent parameter marking the mortality surface. Cornerstones are the CBD models in Cairns et al. (2006) and Cairns et al. (2009).

Finally, a merge between LC and CBD classes exists and it is set out in Plat (2009).

2.2.1 The Poisson LC model

We briefly delve into LC model as in Brouhns et al. (2002), i.e the LC model with a Poisson distribution assumption for the death counts as in Eq.(2.13). Therefore, according to Eq.(2.14) and Eq.(2.15), the linear predictor, $\eta_{x,t}$, contemplates the age-parameter term, $N = 1$ time-dependent parameter, $k_t = k_t^{(1)}$, and absence of cohort term. Formally, $\eta_{x,t}$ is detailed as a Poisson log-bilinear equation as below:

$$\eta_{x,t} = \ln\left(\mathbb{E}\left(\frac{D_{x,t}}{E^c_{x,t}}\right)\right) = \ln m_{x,t} = \alpha_x + \beta_x k_t, \tag{2.16}$$

We stress again that the link function under assumption (2.13) is the canonical one, i.e. the natural logarithm of the central death rate.
To calibrate the predictor (2.16), formulation (2.10) is considered. Therefore, selecting a time horizon $\mathcal{T}$ where a dataset of observations $\left\{\left(D_{x,t}, E^c_{x,t}\right), t \in \mathcal{T}, x \in \mathcal{X}\right\}$ is available, the predictor is fitted finding the optimal estimates $\hat{\alpha}_x$, $\hat{\beta}_x$ and $\hat{k}_t$. Consequently, in order to produce predictions about the future mortality, let t_n the forecasting year and let $\mathcal{T}' = \{t_n + h, h = 1, \ldots, s,\ s \in \mathbb{N}\}$ the forecast horizon. The estimates $\hat{\alpha}_x$ and $\hat{\beta}_x$ are time-invariant, so that the future mortality matures according to the time-index pattern suggesting the linear predictor equation over the forecast horizon, for each age x:

$$\ln m_{x,t_n+h} = \hat{\alpha}_x + \hat{\beta}_x k_{t_n+h}. \tag{2.17}$$

The future time-index values stems from an ARIMA(0,1,0) process, like the following equation states:

$$k_{t_n+h} = k_{t_n} + h\delta + \sum_{k=1}^{h} \epsilon_{t_n+k}, \tag{2.18}$$

where δ is the drift parameter and the ϵ_{t_n+k}'s are indipendent and normally distributed innovations with zero mean and variance σ^2_ϵ, i.e. $\sum_{k=1}^{h} \epsilon_{t_n+k} \sim \mathcal{N}\left(0, h^2\sigma^2_\epsilon\right)$. To develop LC model forecasting ability, a general ARIMA(p,d,q) process could be acknowledged, and the random walk in Eq.(2.18) would be a special case. Properly, the k_{t_n+h} value is extracted from its d^{th}-differences equation:

$$\bigtriangledown^d k_{t_n+h} = h\delta + \sum_{i=1}^{p} \phi_i \bigtriangledown^d k_{(t_n+h)-i} + \sum_{j=1}^{q} \theta_j \epsilon_{(t_n+h)-j} + \sum_{k=1}^{h} \epsilon_{t_n+k}, \tag{2.19}$$

with p the autoregressive order, d the degree of differencing, q the moving-average order and ϕ, θ the autoregressive and moving-average coefficients, respectively, and $\bigtriangledown$ is the difference operator.
Profit from Eq.(2.19), the following forecasting equations are derived:

- The point prediction equation for the log-death rates:

$$\begin{aligned} \ln \hat{m}_{x,t_n+h} &= \mathbb{E}\left(\ln m_{x,t_n+h}\right) = \\ &= \hat{\alpha}_x + \hat{\beta}_x \left(h\delta + \sum_{i=1}^{p} \phi_i \bigtriangledown^d k_{(t_n+h)-i} + \sum_{j=1}^{q} \theta_j \epsilon_{(t_n+h)-j} \right). \end{aligned} \tag{2.20}$$

- The prediction interval equation for the log-death rates:

$$\ln \hat{m}_{x,t_n+h} \pm \hat{\beta}_x \sqrt{h} \sigma_\epsilon z_{\frac{\alpha}{2}}. \tag{2.21}$$

  with z_α the α-quantile of a Standard Normal distribution.

Chapter 3

Neural Networks foundamentals

The artificial Neural Networks (NNs, from now on) appeared for the first time in literature thanks to the pioneering works of McCulloch and Pitts (1943) and Turing (1948). NNs models are adaptive systems based on the cognitive paradigm to calculus. The building blocks of a NN model are the so called neurons or units. They represent the network nodes apt to receive information, elaborate them and produce an output communicated to the next units. Since neurons are linked through weighted connections, each network unit receives a pondered information, namely activation. This one are transformed applying a differentiable function, namely activation function. In doing so, each neuron provides an own output useful for further computations. Overall, a NN model figures out a function such that a given output is provided by propagating different transformations of the input, using the weights as intermediate parameters.

We can also interpret a NN model in terms of computational graph constitutes by different layers wherein one or more neurons are present. The first and the last layers are namely the input and the output layer, respectively, while the intermediate layers are the so called hidden layers. The NN processing flow typically moves from the input layer to the output one, passing through hidden states: this is how work feed-forward NNs. Moreover, exist several types of NN architectures with peculiar ways to elaborate data. For example, in the feed-forward NN structure it is possible adds recurrent connections among neurons, getting the so called Recurrent Neural Networks.
For the sake of simplicity, we briefly recall the simplest feed-forward NN, namely perceptron (see Rosenblatt (1958)), whose structure is composed by a single layer of input neurons, as depicted in Figure 3.1. Therefore, given an input vector $\boldsymbol{x} \in \mathbb{R}^n$ we can associate its input layer in the NN, later computing the activation $A = \sum_{i=1}^{n} w_i x_i$, with $\boldsymbol{w} \in \mathbb{R}^n$ the vector of weights. The NN output $y \in \mathbb{R}$ is obtained applying the activation function, $f : \mathbb{R} \to \mathbb{R}$, to the activation A.

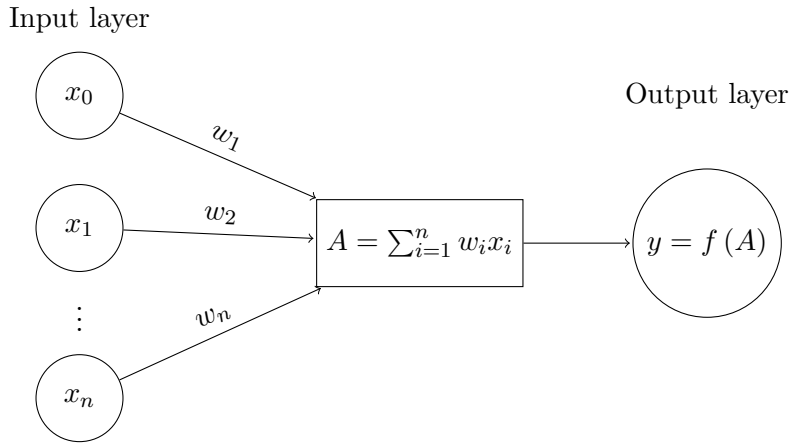


Figure 3.1. Graph for a perceptron, with n input units and one output unit.

NNs with multiple layers can be built adding hidden layers. Such a NN structure is called Multilayer Perceptron (MLP, from now on), allowing for very in depth input transformations. In particular, MLP architecture manifests the NNs power to discover hidden features in a set of examples and generalize them. In order to define the MLPs structure, let $N_0, N_p, N_{P+1} \in \mathbb{N}$ be, respectively, the number of neurons within the input layer, the p^{th} hidden layer and the output layer, for $p \in \{1, \dots, P\}$, $P \in \mathbb{N}$. Let $A^{(p)} : \mathbb{R}^{N_{p-1}} \to \mathbb{R}^{N_p}$ an affine map defining the p^{th} hidden layer activation, given the output produced by the $(p-1)^{th}$ hidden layer.

Definition 1. *Let $\phi : \mathbb{R}^{N_p} \to \mathbb{R}^{N_p}$ be a differentiable activation function. The p^{th} hidden layer output is:*

$$H^{(p)} = \left(\phi \circ A^{(p)}\right)\left(H^{(p-1)}\right) = \phi\left(\langle \boldsymbol{W}^{(p)}, H^{(p-1)} \rangle + \boldsymbol{b}^{(p)}\right), \tag{3.1}$$

where $\boldsymbol{W}^{(p)} \in \mathbb{R}^{N_p \times N_{p-1}}$ is the weight matrix for feed-forward hidden layer connections and $\boldsymbol{b}^{(p)} \in \mathbb{R}^{N_p}$ is the bias term. The latter is an additional parameter necessary to govern the triggering value of the activation function ϕ. Therefore, each bias component acts as a neuron activation threshold.

Definition 2. *Let $\mathcal{D} = \left\{(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{x} \in \mathbb{R}^{N_0}, \boldsymbol{y} \in \mathbb{R}^{N_{P+1}}\right\}$ be a dataset wherein $\boldsymbol{x}$ is the explicative variable and $\boldsymbol{y}$ is the associated response. Denoting with $\boldsymbol{\mathcal{W}}$ the set of all parameters employed along the NN graph, a feed-forward NN model is a function $f_{NN} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{P+1}}$ such that:*

$$\boldsymbol{y} = f_{NN}(\boldsymbol{x}; \boldsymbol{\mathcal{W}}) + \gamma = \psi \circ \left(H^{(P)} \circ H^{(P-1)} \circ \cdots \circ H^{(1)}\right)(\boldsymbol{x}; \boldsymbol{\mathcal{W}}) + \gamma. \tag{3.2}$$

where $\psi : \mathbb{R}^{N_P} \to \mathbb{R}^{N_{P+1}}$ is the output layer activation function and γ is a data noise, having zero mean and constant variance.

Figure 3.2 offers a graphical visualization of the NN structure according to Eq.(3.2).

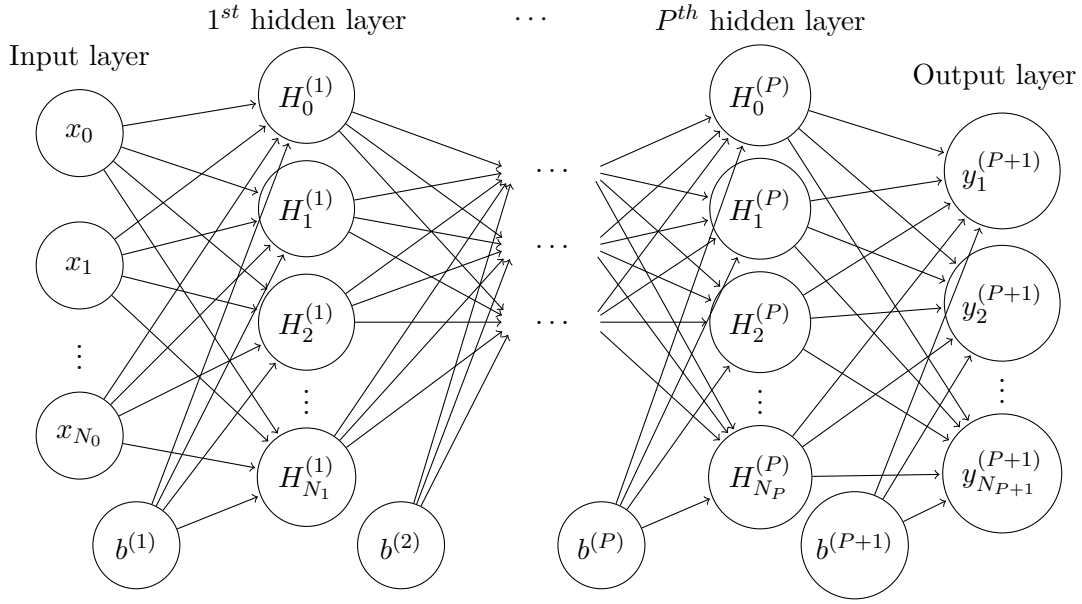


Figure 3.2. Graph for a P-hidden layered NN, with N_0 input units and N_{P+1} output units.

Eq.(3.2) shows the strength of neural models with respect to the classical statistical ones. Indeed, combining multiple neurons in different layers the learning ability of the model increases, discovering complex relations in the input-output data. Mathematically speaking, this NN feature is nowadays well-known as the universal functional approximation property.

3.1 The universal approximation property

According to Eq.(3.2), a NN model is described by composition and superposition of differentiable functions, and the resulting outputs stem from the conjunction of several non-linearities. In doing so, the NN function can be viewed as a map with functional approximation capability. In 1989 Cybenko (1989), Hornik et al. (1989) and Funahashi (1989) proved, for the first time in literature, the approximation property of feed-forward NNs. In particular, Cybenko (1989) showed that NNs with one hidden layer and an arbitrary continuous sigmoid activation function can approximate continuous functions, with arbitrary accuracy and without constraints on the number of hidden neurons. Therefore, considering a space of continuous functions $C\left(\mathbb{R}^{N_0}\right)$ and a continuous sigmoid function $\sigma(\boldsymbol{x}) = (1+e^{-\boldsymbol{x}})^{-1} \in C\left(\mathbb{R}^{N_0}\right),\ \boldsymbol{x} \in \mathbb{R}^{N_0}$, for any $g \in C\left(\mathbb{R}^{N_0}\right)$ and $\varepsilon > 0$ exists a finite sum:

$$G(\boldsymbol{x}) = \sum_{i=1}^{N_1} w_i \sigma\left(\langle \boldsymbol{w}_i^{(1)}, \boldsymbol{x}\rangle + b_i\right) \tag{3.3}$$

such that:

$$\left|G(\boldsymbol{x}) - g(\boldsymbol{x})\right| < \varepsilon \quad \forall \boldsymbol{x} \in \mathbb{R}^{N_0}, \tag{3.4}$$

i.e. $G(\boldsymbol{x})$ is dense in $C\left(\mathbb{R}^{N_0}\right)$. This result is presented in Cybenko (1989), and therein extended for any function $g \in C\left(\mathbb{R}^{N_0}\right)$ over any Lesbegue-measurable partition of $\mathbb{R}^{N_0}$, as well as for discontinuous sigmoid functions. Similarly, Funahashi (1989) proved the approximation property for both one and two hidden layered feed-forward NNs. Finally, Hornik et al. (1989) demonstrated that MLPs constitute a class of universal functional approximators for any Borel-measurable function on finite dimensional spaces, achieving high accuracy with an arbitrary number of hidden neurons. Thus, since '90 a line of research on the NNs functional approximation capabilities was born. We refer the reader to Siegel and Xu (2019), and references therein, for a formal and well-posed literature review.
We stress that the approximation property implies the following considerations:

- The approximation property refers to high precision in reproducing maps, involving the trade-off between model accuracy and its interpretability. NNs present high accuracy with a low interpretability, earning the ephitet of "black box" models. Properly, the NNs function are not merely "black boxes", but for sure their interpretability depends on the dimensionality of the functional composition;
- NNs models can approximate any continuous (and several discontinuous) functions with an arbitrary number of hidden neurons. Therefore, the learning ability of a NN model derives from the tuning procedures in order to define the optimal architecture of the graph, i.e. the dimensionality of both function superposition and composition;
- NN outputs approximate a realization of a map, exploiting an adequate dataset of examples. From a statistical perspective, a NN model could be view as an estimator, whose functional form is shaped according to the data.

The approximation property has opened up a wide area of NNs applications for solving various theoretical and practical tasks, testing how effectively the approximation property itself responds to the observed features within data.

3.1.1 Neural Networks and the Statistical Learning Theory

Because of the universal approximation property, NNs have been systematically included in the field of Statistical Learning Theory. The latter concerns the study of conditions for a consistent approach to function estimation problems. Essentially, the goal is to define a learning model apt to express a solid, universal approximator given a set of observations. As stated in Vapnik (1999), learning models like NNs allow to minimize estimation errors in the main statistical learning problems, as in the case of regression estimation. Recalling $\mathcal{D} = \left\{(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{x} \in \mathbb{R}^{N_0}, \boldsymbol{y} \in \mathbb{R}^{N_{P+1}}\right\}$ the available set of observations, the variable $\boldsymbol{x}$ is a realization from an unknown distribution $\mathbb{P}(\boldsymbol{x})$, while the response variable stems from the probability distribution $\mathbb{P}(\boldsymbol{y}|\boldsymbol{x})$. However, such distributions are unknown and a possible set of theoretical maps, $\Sigma = \{\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{w}), \boldsymbol{w} \in \boldsymbol{\mathcal{W}}\}$, should be addressed to represent the input-output relation. Thus, the goal of the learning process is to identify a function $f(\boldsymbol{x}, \boldsymbol{w}^{\star}) \in \Sigma$,

given the data $\mathcal{D}$, in order to minimize the discrepancy with respect to $\boldsymbol{y}$. Letting $\mathcal{L}(\boldsymbol{y}, f(\boldsymbol{x}, \boldsymbol{w}))$ the measure of discrepancy, namely risk o loss function, then the learning process aims to solve the following problem:

$$\min_{\boldsymbol{w}\in\mathcal{W}} \mathbb{E}^{\mathbb{P}(\boldsymbol{x},\boldsymbol{y})}\left(\mathcal{L}(\boldsymbol{y}, f(\boldsymbol{x}, \boldsymbol{w}))\right) = \int \mathcal{L}(\boldsymbol{y}, f(\boldsymbol{x}, \boldsymbol{w}))\ d\mathbb{P}(\boldsymbol{x}, \boldsymbol{y}).$$

where the joint probability distribution $\mathbb{P}(\boldsymbol{x}, \boldsymbol{y})$ is also unknown and the only available knowledge is the dataset $\mathcal{D}$. For example, in regression estimation problems the scope is to approximate the regression function

$$\mathbb{E}^{\mathbb{P}(\boldsymbol{y}|\boldsymbol{x})}(\boldsymbol{y}|\boldsymbol{x}) = \int \boldsymbol{y}\ d\mathbb{P}(\boldsymbol{y}|\boldsymbol{x})$$

finding the function $f(\boldsymbol{x}, \boldsymbol{w}^\star) \in \Sigma$ such that

$$\mathcal{L}(\boldsymbol{y}, f(\boldsymbol{x}, \boldsymbol{w})) = (\boldsymbol{y} - f(\boldsymbol{x}, \boldsymbol{w}))^2$$

reaches its minimum value. Since the probability distributions of the considered variables are unknown, the datata $\mathcal{D}$ is exploited to replace the expected discrepancy measure with its empirical version

$$\hat{\mathbb{E}}\left(\mathcal{L}(\boldsymbol{y}, f(\boldsymbol{x}, \boldsymbol{w}))\right) = \frac{1}{N_0 N_{P+1}} \sum_{i=1}^{N_0 N_{P+1}} f(\boldsymbol{x}, \boldsymbol{w}_i).$$

As stated in Vapnik (1999), this is the Empirical Risk Minimization induction principle: lacking any information about distributions, we elect as the best functional approximator the one allowing for the minimum empirical mean error.
In this framework, the NN guarantees consistency in the learning process, a non-asymptotic behaviour of the rate of convergence towards the minimum error and a performing generalization capacity. However, the NN generalization ability depends on how well they minimize the empirical risk. In particular, some problems may emerge regarding the algorithms speed of convergence, as well as multiple local minima occurrences for the risk function. Therefore, the empirical risk minimization heavily depends on the learning process.

3.2 Neural Network learning

The NNs learning process is focused on the search of both the best NN graph architecture, $\hat{f}_{NN}(\cdot)$, and the optimal value for the weights, $\hat{\boldsymbol{\mathcal{W}}}$, starting from available data. Fixing the NN architecture, the network parameters are calibrated in order to provide the minimum empirical error. According to Eq.(3.2), the output estimated by the NN is:

$$\hat{\boldsymbol{y}} = \hat{f}_{NN}\left(\boldsymbol{x}; \hat{\boldsymbol{\mathcal{W}}}\right). \tag{3.5}$$

To optimize the weights value, gradient-based methods are employed. In particular, considering a differentiable loss function, $\mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}})$, the weights are iteratively adjusted up to minimize the empirical error following the rules below:

$$\Delta w_{ij}^{(p)} = -\eta \frac{\partial \mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{y}})}{\partial w_{ij}^{(p)}}, \tag{3.6}$$

where $\Delta w_{ij}^{(p)}$ is the variation of the weight related to the output elaborated by the neuron j within the $(p-1)^{th}$ layer, and received by the units i in the p^{th} layer, for $j = 1, \dots, N_{p-1}$ and $i = 1, \dots, N_p$. The parameter $\eta \in \mathbb{R}_+$ is the learning rate governing the speed of convergence to the minimum. To calculate the loss variations to changes in parameters, the procedure proposed in Rumelhart et al. (1986), namely backpropagation (BP, from now on), represents a milestone. The BP algorithm allows to backpropagate the error along the NN graph, determining what is the necessary weights adjustment to attain a smaller loss function value, up to a minimum.
BP algorithm implementation concerns two specific phases, namely the forward and backward phases. In the former, each hidden layer values are computed up to the output layer, getting the NN result $\hat{\boldsymbol{y}}$. Consequently, the loss function magnitude can be figure out. The backward phase spreads the error from the output layer to the input one, revising the parameters value. Such parameters adjustment requires the computation of the partial derivative of the loss function with respect to the NN weights, that is:

$$\begin{aligned}
\frac{\partial \mathcal{L}\left(\hat{\boldsymbol{y}}, \boldsymbol{y}\right)}{\partial \boldsymbol{\mathcal{W}}} &= \sum_{p=1}^{P+1} \frac{\partial \mathcal{L}_{(p)}\left(\hat{\boldsymbol{y}}, \boldsymbol{y}\right)}{\partial \boldsymbol{W}^{(p)}} = \\
&= \sum_{p=1}^{P+1} \frac{\partial \mathcal{L}_{(p)}\left(\partial \hat{\boldsymbol{y}}, \boldsymbol{y}\right)}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial H^{(p)}} \frac{\partial H^{(p)}}{\partial \boldsymbol{W}^{(p)}} = \\
&= \sum_{p=1}^{P+1} \frac{\partial \mathcal{L}_{(p)}\left(\partial \hat{\boldsymbol{y}}, \boldsymbol{y}\right)}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial H^{(p)}} \prod_{j=p}^{P} \frac{\partial H^{(j+1)}}{\partial H^{(j)}} \frac{\partial H^{(p)}}{\partial \boldsymbol{W}^{(p)}} = \\
&= \sum_{p=1}^{P+1} \frac{\partial \mathcal{L}_{(p)}\left(\partial \hat{\boldsymbol{y}}, \boldsymbol{y}\right)}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial H^{(p)}} \prod_{j=p}^{P} \phi'\left(\langle \boldsymbol{W}^{(p)}, H^{p-1} \rangle + \boldsymbol{b}^{(p)}\right) \boldsymbol{W}^{(p)} \frac{\partial H^{(p)}}{\partial \boldsymbol{W}^{(p)}}.
\end{aligned} \tag{3.7}$$

Within Eq.(3.7), the BP algorithm engages the counting of loss function gradient for each layer, implementing the following procedure:

1. Consider a random parameters initialization to make the forward phase, producing a NN output so that the loss function is computed;

2. Start the backward phase computing the error to propagate backwards along the NN graph. To this end, the error attributable to each layer is:

$$\delta^{(p)} := \frac{\partial \mathcal{L}_{(p)}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right)}{\partial A^{(p)}}, \tag{3.8}$$

and it must computed from the output layer up to the input one, that is:

$$\delta^{(P+1)} = \frac{\partial \mathcal{L}_{(P+1)}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right)}{\partial A^{(P+1)}} = \frac{\partial \mathcal{L}_{(P+1)}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right)}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial A^{(P+1)}} = \nabla_{\hat{\boldsymbol{y}}} \mathcal{L}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) \odot \psi'\left(A^{(P+1)}\right)$$

$$\delta^{(P)} = \frac{\partial \mathcal{L}_{(P)}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right)}{\partial A^{(P)}} = \frac{\partial \mathcal{L}_{(P)}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right)}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial A^{(P+1)}} \frac{\partial A^{(P+1)}}{\partial H^{(P)}} \frac{\partial H^{(P)}}{\partial A^{(P)}} =$$
$$= \delta^{(P+1)} \frac{\partial A^{(P+1)}}{\partial H^{(P)}} \frac{\partial H^{(P)}}{\partial A^{(P)}} = \left[\left(\boldsymbol{W}^{(P+1)}\right)^{\mathrm{T}} \delta^{(P+1)}\right] \odot \phi'\left(A^{(P)}\right)$$
$$\vdots$$
$$\delta^{(1)} = \frac{\partial \mathcal{L}_{(1)}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right)}{\partial A^{(1)}} = \frac{\partial \mathcal{L}_{(1)}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right)}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial A^{(P+1)}} \prod_{k=1}^{P} \frac{\partial A^{(k+1)}}{\partial H^{(k)}} \frac{\partial H^{(1)}}{\partial A^{(1)}} =$$
$$= \left[\left(\boldsymbol{W}^{(2)}\right)^{\mathrm{T}} \delta^{(2)}\right] \odot \phi'\left(A^{(1)}\right),$$

where the operator $\odot$ is the Hadamard product. We can note that the BP phase is based on the recursive calculus of the error for each layer, i.e. holds that:

$$\delta^{(p)} = \left[\left(\boldsymbol{W}^{(p+1)}\right)^{\mathrm{T}} \delta^{(p+1)}\right] \odot \phi'\left(A^{(p)}\right),$$

and other quantity involved in are all defined in the above forward phase;

3. Since the error is computed and backpropagated, the weights changes can be computed for each layer according to Eq.(3.6), i.e.:

$$\Delta \boldsymbol{W}^{(p)} = -\eta \frac{\partial \mathcal{L}_{(p)}\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right)}{\partial \boldsymbol{W}^{(p)}} = -\eta H^{(p-1)} \delta^{(p)}. \tag{3.9}$$

After step 3., new weights are available and then a new forward phase can begin, so that steps from 1. to 3. are repeated until the empirical error is small as desired. The number of iterations of steps 1.-3. are technically called epochs, and we can speed up the algorithm implementing the early stopping technique. This simply consists in stopping the iterations after a certain number of epochs if the loss function improvements are not significant.
Finally, we stress that the BP method is executed fixing the NN architecture, i.e. considering a certain dimensionality in functions composition. So, given the input-output examples, the BP allows to operationally determine the best weights value to optimize the loss function: this is the concept of NN training. However, the learning process includes also procedures to define what is the best composition of the NN architecture, $\hat{f}(\cdot)$. This is the concept of NN tuning.
The first step for tuning and training a NN model is to split the available dataset in different partitions. Commonly, the dataset is divided into three parts: the training set, the validation set and the testing set. The training data are used to build the model architecture, which is technically defined by the so called model hyper-parameters and learning hyper-parameters. Both are parameters different from the weights, but the former allowing to define the NN structure, while the latter serves to technically support the optimization problem. For example, the number of hidden layers, the number of neurons within each hidden layer and the types of activation functions to employ are model hyper-parameters; while, the learning rate magnitude, the epochs level and the parameters governing the early stopping rules are learning hyper-parameters. The search of both the best model and learning

hyper-parameters value constitutes the tuning process. Moreover, during the tuning process also the weights are optimized through the aforementioned BP method. Hence, for a NN architecture we assign its optimal set of weights.

The actual evaluation of the best NN model is not done on the training data, but on the validation set. Indeed, we consider an hyper-parameters domain containing several values attributable to the hyper-parameters, trying each architecture combination on the training data and, finally, validate it on the validation set. The favourite alternative of hyper-parameters is selected through accuracy measures on the validation set. In doing so, every model architecture, equipped with its optimized weights, are tested for the first time on the validation data to understand the NN ability to generalize. Finally, the testing set is utilized to prove the accuracy of the final tuned NN model. It is essential that the testing data are not employed during the tuning process, in order to express a consistent judgment on both model efficiency and robustness. To inspect the wide range of algorithmic techniques for tuning purposes, we refer the reader to Aggarwal (2018) and Goodfellow et al. (2016), and references therein.

3.3 Neural forecasting

In the field of forecasting, NNs are suitable tools for practitioners and researchers in many predictive tasks. As highlighted in Makridakis et al. (2020), deep learning based models represents the state of the art in time series forecasting. Deep learning models are NN models capable to reach high level of abstraction thanks to the presence of several hidden layers o recurrent connections among neurons. In doing so, very in depth data features are discovered, supercharging classical forecasting models or figuring out new predictive approaches. In the area of deep learning models, MLPs structure identifies a vanilla design. However, by construction, MLPs are not adequate to explore temporal relations among data, since phenomena like serial correlations or time invariant patterns are not caught. Therefore, *ad hoc* NN architecture for handle sequential data, such as time series, has been created: the Recurrent Neural Networks (RNN, from now on).

3.3.1 Recurrent Neural Networks

RNNs are network models characterized by neurons self-connected or connected to units of the previous layers, in addition to the feed-forward connections. The presence of recurrences donates to the RNN the nature of dynamic system, since each data of a sequence is analysed jointly with information stored from the past inputs. Hence, a dynamic memory is naturally formed, allowing for a proper inspection of temporal correlations between events that may be far from each other (see, for example, Rumelhart et al. (1986), Werbos (1988) and Elman (1990)). This latter feature is vital for time series learning and their forecasting. In particular, let us consider a temporal sequence of input-output pairs $\{(\boldsymbol{x}_t, \boldsymbol{y}_t)\}$, where $\boldsymbol{x}_t$ is the input received by the network at time t and $\boldsymbol{y}_t$ is the related target. The role of a RNN is to approximate the temporal map $\varphi : \boldsymbol{x}_t \mapsto \boldsymbol{y}_t$. To this end, RNNs working is based

on two essentials concepts: the unfolding operation and the weights sharing. Indeed, by construction, a RNN is a cyclical graph to which it is possible to associate its acyclic version, called unfolded RNN. Unrolling the RNN implies the transformation of recurrent connections in feed-forward connections following the data sequential order. On the unfolded RNN operates the weights sharing: for each sequence of unrolled nodes, the elaborated data are pondered by the same weights.

An illustration of the RNN architecture and its unfolded version is provided in Figure 3.3.

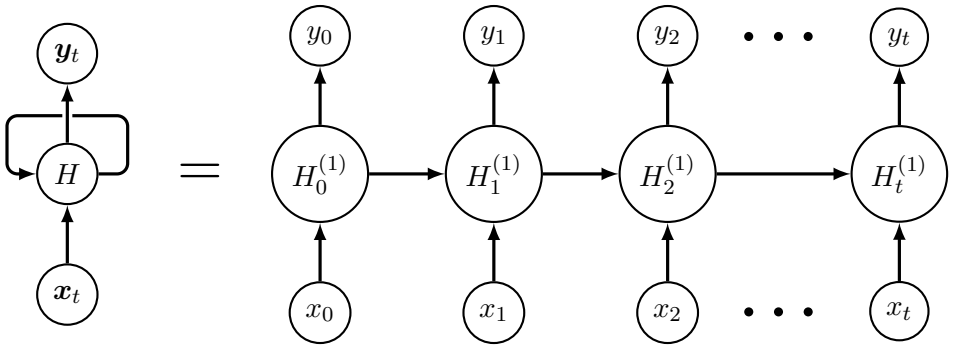


Figure 3.3. A one-hidden layered vanilla RNN and its unfolded version.

Therefore, in a RNN the data manipulation stems from a one-to-one correspondence between the layers within the acyclic graph and the time-stamp of a single data in the sequence. Given the weights sharing, for each time-stamp we apply the same set of parameters, ensuring a coherent modelling over the time.
Referring to a vanilla RNN with a single, recurrent, hidden layer as in Figure 3.3, let $H_t^{(1)} \in \mathbb{R}^{N_1}$ be the hidden layer output at time t and let f_H and f_y be the differentiable activation functions of the hidden layer and the output one, respectively. $H_t^{(1)}$ is defined as function of the input at the same time step and the hidden layer of the previous time step, $H_{t-1}^{(1)}$, while the output at time t is a function of the hidden layer at the same time step. Therefore, the target of this dynamic system is defined by the following equation:

$$\boldsymbol{y}_t = f_y\left(H_t^{(1)}\right) = f_y\left(f_H\left(\boldsymbol{x}_t, H_{t-1}^{(1)}\right)\right). \tag{3.10}$$

Since NNs are parameter-dependent maps (see Eq.(3.2)), let define

$$\boldsymbol{\mathcal{W}} = \left\{\boldsymbol{W}^{(1)}, \boldsymbol{U}^{(1)}, \boldsymbol{W}^{(2)}, \boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}\right\}$$

the set of parameters characterizing the RNN, where:

- $\boldsymbol{W}^{(1)} \in \mathbb{R}^{N_1 \times N_0}$ is the weight matrix between the input and the hidden layer;
- $\boldsymbol{U}^{(1)} \in \mathbb{R}^{N_1 \times N_1}$ is the recurrences weight matrix within hidden layers, i.e. the weights shared along the unfolded graph;
- $\boldsymbol{W}^{(2)} \in \mathbb{R}^{N_2 \times N_1}$ is the weight matrix between the hidden layer and the output.

- $\boldsymbol{b}^{(1)} \in \mathbb{R}^{N_1}$ and $\boldsymbol{b}^{(2)} \in \mathbb{R}^{N_2}$ are the bias vectors of the hidden layer and the output layer, respectively.

Therefore, Eq.(3.10) is rewritten as follow:

$$\begin{aligned} \boldsymbol{y}_t &= f_y\left(\langle \boldsymbol{W}^{(2)}, H_t^{(1)}\rangle + \boldsymbol{b}^{(2)}\right) = \\ &= f_y\left(\langle \boldsymbol{W}^{(2)}, \left(f_h\left(\langle \boldsymbol{W}^{(1)}, \boldsymbol{x}_t\rangle + \langle \boldsymbol{U}^{(1)}, H_{t-1}^{(1)}\rangle + \boldsymbol{b}^{(1)}\right)\right)\rangle + \boldsymbol{b}^{(2)}\right) \end{aligned} \tag{3.11}$$

As emerges in Eq.(3.11), the RNN is a function obtained by composition and superposition of differentiable activation functions, as well as feed-forward NNs. The only difference is related to the model complexity: composition and superposition are execute over one or more layers and over different time steps. In light of this, RNNs fall within the set of deep learning models even without involving many hidden layers. Obviously, a general multi-layer representation is useful and it is expressed in the following Definition 3 and 4.

Definition 3. *Let $A_t^{(p)} : \mathbb{R}^{N_{p-1}} \to \mathbb{R}^{N_p}$ an affine map defining the p^{th} hidden layer activation at time t. Let $\phi : \mathbb{R}^{N_p} \to \mathbb{R}^{N_p}$ a differentiable activation function. The output at time t from the p^{th} hidden layer is:*

$$H_t^{(p)} = \left(\phi \circ A_t^{(p)}\right)\left(H_t^{(p-1)}, H_{t-1}^{(p)}\right) = \phi\left(\langle \boldsymbol{W}^{(p)}, H_t^{(p-1)}\rangle + \langle \boldsymbol{U}^{(p)}, H_{t-1}^{(p)}\rangle + \boldsymbol{b}^{(p)}\right), \tag{3.12}$$

where $\boldsymbol{W}^{(p)} \in \mathbb{R}^{N_p \times N_{p-1}}$ is the weight matrix for feed-forward hidden layer connections, $\boldsymbol{U}^{(p)} \in \mathbb{R}^{N_p \times N_p}$ is the weight matrix for recurrent hidden layer connections and $\boldsymbol{b}^{(p)} \in \mathbb{R}^{N_p}$ is the bias term.

Definition 4. *Let us consider a dataset $\mathcal{D} = \left\{(\boldsymbol{x}_t, \boldsymbol{y}_t), \boldsymbol{x}_t \in \mathbb{R}^{N_0}, \boldsymbol{y}_t \in \mathbb{R}^{N_{P+1}}\right\}$, where $\boldsymbol{x}_t$ is the input received by the network at time t and $\boldsymbol{y}_t$ is the related target. An RNN model is a function $f_{RNN} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{P+1}}$ such that:*

$$\boldsymbol{y}_t = f_{RNN}\left(\boldsymbol{x}_t; \boldsymbol{\mathcal{W}}\right) + \gamma_t = \psi \circ \left(H_t^{(P)} \circ H_t^{(P-1)} \circ \cdots \circ H_t^{(1)}\right)\left(\boldsymbol{x}_t; \boldsymbol{\mathcal{W}}\right) + \gamma_t, \tag{3.13}$$

with $\psi : \mathbb{R}^{N_P} \to \mathbb{R}^{N_{P+1}}$ is the output layer activation function, and γ_t is the zero mean data noise at time t.

Finally, as for feed-forward NNs, also for RNNs holds the universal functional approximation property (Schäfer and Zimmermann (2007)). Hence, RNNs are used as data-driven predictors approximating temporal maps and being able to produce realizations over a designed forecast horizon.

3.3.2 Learning over the time

The RNN learning process follows the same philosophy explained in Section 3.2. As stated in Werbos (1988), to train a RNN the steps provided by the BP method are employed, with some necessary adjustment to take into account the peculiarities of the RNN structure: the resulting optimization procedures are named backpropagation through the time (BPTT, from now on). The latter requires to run the input

elaboration in the forward direction, producing a time-indexed output and computing the loss function for each time step. Consequently, the backpropagation begins considering the gradients of the loss with respect to the weights over the unrolled graph, starting from the last time step in the last layer and continuing to scale on times and layers. Because of the weights are shared among different time stamp, the BPTT method is a BP applied recursively over the various temporal nodes, within each layer.
Formally, given the overall loss function $\mathcal{L}(\hat{\boldsymbol{y}}, \boldsymbol{y})$, the following derivative must be computed:

$$\begin{aligned}
&\frac{\partial \mathcal{L}(\hat{\boldsymbol{y}}, \boldsymbol{y})}{\partial \boldsymbol{\mathcal{W}}} = \sum_{p=1}^{P+1} \sum_{t=1}^{T} \frac{\partial \mathcal{L}_t^{(p)}(\hat{\boldsymbol{y}}, \boldsymbol{y})}{\partial \boldsymbol{\mathcal{W}}} = \\
&= \sum_{p=1}^{P+1} \sum_{t=1}^{T} \sum_{i=1}^{t} \frac{\partial \mathcal{L}_t^{(p)}(\hat{\boldsymbol{y}}, \boldsymbol{y})}{\partial \hat{\boldsymbol{y}}_t} \frac{\hat{\boldsymbol{y}}_t}{\partial H_t^{(p)}} \prod_{j=i}^{t-1} \frac{\partial H_{j+1}^{(p)}}{\partial H_j^{(p)}} \frac{\partial H_i^{(p)}}{\partial \boldsymbol{W}^{(p)}} = \\
&= \sum_{p=1}^{P+1} \sum_{t=1}^{T} \sum_{i=1}^{t} \frac{\partial \mathcal{L}_t^{(p)}(\hat{\boldsymbol{y}}, \boldsymbol{y})}{\partial \hat{\boldsymbol{y}}_t} \frac{\hat{\boldsymbol{y}}_t}{\partial H_t^{(p)}} \cdot \\
&\qquad \cdot \prod_{j=i}^{t-1} diag\left(\phi'\left(\langle \boldsymbol{W}^{(p)}, H_{j+1}^{(p)} \rangle + \langle \boldsymbol{U}^{(p)}, H_j^{(p)} \rangle + \boldsymbol{b}^{(p)}\right)\right) \boldsymbol{U}^{(p)} \frac{\partial H_i^{(p)}}{\partial \boldsymbol{W}^{(p)}},
\end{aligned} \tag{3.14}$$

where each addendum of the sum is calculated through classical BP method defined in Section 3.2. However, dealing with RNNs an important training problem emerges, namely vanishing or exploding gradient problem (Bengio et al. (1994) and Pascanu et al. (2013)). In particular, when the length of the sequence increases, the derivative in Eq.(3.14) is affected by short-term dependencies. Indeed, any change in the hidden state has a multiplicative effect so that, if the number of recurrences increases, the multiplication in Eq.(3.14) rapidly converges to 0 when the Jacobian eigenvalues are less than 1. This means that, if the largest eigenvalue is less than 1, the gradient will vanish, otherwise it explodes. Denoting with $\lambda_{max}^{\langle H \rangle}$ and $\lambda_{max}^{\langle \boldsymbol{U} \rangle}$ the largest eigenvalues associated to $||diag\left(\phi'\left(\langle \boldsymbol{W}^{(p)}, H_{j+1}^{(p)} \rangle + \langle \boldsymbol{U}^{(p)}, H_j^{(p)} \rangle + \boldsymbol{b}^{(p)}\right)\right)||$ and $||\boldsymbol{U}^{(p)}||$, respectively, holds that:

$$\left|\left|\frac{\partial H_{j+1}^{(p)}}{\partial H_j^{(p)}}\right|\right| \leq \left|\left|diag\left(\phi'\left(\langle \boldsymbol{W}^{(p)}, H_{j+1}^{(p)} \rangle + \langle \boldsymbol{U}^{(p)}, H_j^{(p)} \rangle + \boldsymbol{b}^{(p)}\right)\right)\right|\right| \cdot \left|\left|\boldsymbol{U}^{(p)}\right|\right| \leq \lambda_{max}^{\langle H \rangle} \lambda_{max}^{\langle \boldsymbol{U} \rangle}, \tag{3.15}$$

$$\left|\left|\frac{\partial H_t^{(p)}}{\partial H_i^{(p)}}\right|\right| = \left|\left|\prod_{j=i}^{t-1} \frac{\partial H_{j+1}^{(p)}}{\partial H_j^{(p)}}\right|\right| \leq \left(\lambda_{max}^{\langle H \rangle} \lambda_{max}^{\langle \boldsymbol{U} \rangle}\right)^{t-i}. \tag{3.16}$$

As the sequence becomes longer, i.e the distance between t and i increases, the eigenvalues will determine if the gradient either becomes exceptionally large (explodes) or very small (vanishes). The vanishing problem has been overcome by managing the recurrent hidden units through special neuronal engineering, namely gates. In light of this, in the recent years different gated RNN has been created, and at the state of the art the most popular is the Long Short-Term Memory (LSTM, from now on) thanks to its forecasting performances.

3.3.3 Learning to forget: the LSTM block

The LSTM has been introduced by Hochreitern and Schmidhuber (1997) as solution to the vanishing (exploding) gradient problem. In particular, the LSTM is not properly a NN, but a particular neurons engineering oriented to both control the information flow and to maintain temporal characteristics learned. Hence, the LSTM neuron, or LSTM block or LSTM cell, depicts an innovative units structure grafted into a RNN, resulting the so called RNN model with LSTM architecture. Initially employed for natural processing language tasks, the LSTM has showed its ability to handle sequential data, avoiding gradients problems and being able to learn what time dependencies to preserve or to forget. Developments and variants have been presented in past literature (see, for example, Gers et al. (1989), Gers and Schmidhuber (2000), Graves et al. (2009)), involving forecasting intents of several fields, from natural sciences to economics.
The vanilla LSTM neuron is made up of two fundamental parts. The first is the memory or cell unit, $\boldsymbol{c}_t$, which incorporates significant information over time, allowing long-term dependencies to be maintained by integrating them from time to time with the inputs of the current time step. The second one refers to the gates, that is perceptrons inserted into the the LSTM neuron. At each time t, the LSTM block receives as input three type of data: the memory state at time $(t-1)$, $\boldsymbol{c}_{t-1}$, the elaboration of the previous node, H_{t-1}, and the current input, x_t. All of them are processed through the gates in order to generate a two-fold output to transmit to the next node: the updated long-term memory, $\boldsymbol{c}_t$, and the short-term outcome, H_t. In particular, the LSTM gates are the following:

- *the forget gate*, $\boldsymbol{f}_t$, whose role is to establish if the inputs received should be take into account to modify the memory state acquired. To this purpose, the previous hidden state result and the current input are transformed by the sigmoid function. Given the codomain of the latter, the forget gate output is in $[0, 1]$: the closer to 0 implies forgetting and the closer to 1 means keeping;

- *the input gate*, $\boldsymbol{i}_t$, that aims to update the current memory state given the previous hidden state and current input. This information are treated thanks to a sigmoid function, so that an output close to 0 does not affect the current memory state, vice versa in the case of the input gate output tends to 1. To fulfil its role, the input gate is helped by an auxiliary perceptron, that receives the hidden state and the current input as arguments of the hyperbolic tangent function. In doing so, the input gate working is regulated;

- *the output gate*, $\boldsymbol{o}_t$, necessary to compute the current hidden state result. Also the output gate collects both the previous hidden state result and the current input, passing them into a sigmoidal function. The sigmoidal output is multiplied with an hyperbolic tangent output in order to produce the current hidden state, embedding also the current memory influences.

The Figure 3.4 displays the LSTM block as mentioned above.

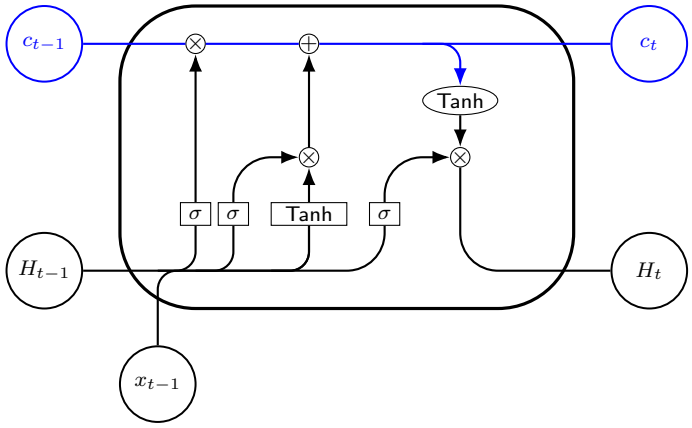


Figure 3.4. Representation of a single LSTM neuron and its internal forward flow.

Formally, the LSTM neuron working stems from the system of equations posed in the following Definition 5.

Definition 5. *Let $H_{t-1}^{(p)}$ and $\boldsymbol{c}_{t-1}^{(p)}$ the hidden state and the memory state, respectively, resulting at the time $(t-1)$ within the p^{th} hidden layer, and let $H_t^{(p-1)}$ the current input. Given the LSTM gates equations below:*

$$\begin{aligned} \textit{Forget gate}: \quad & \boldsymbol{f}_t^{(p)} = \sigma\left(\langle \boldsymbol{W}_f^{(p)}, H_t^{(p-1)}\rangle + \langle \boldsymbol{U}_f^{(p)}, H_{t-1}^{(p)}\rangle + \boldsymbol{b}_f^{(p)}\right), \\ \textit{Input gate}: \quad & \boldsymbol{i}_t^{(p)} = \sigma\left(\langle \boldsymbol{W}_i^{(p)}, H_t^{(p-1)}\rangle + \langle \boldsymbol{U}_i^{(p)}, H_{t-1}^{(p)}\rangle + \boldsymbol{b}_i^{(p)}\right), \\ \textit{Output gate}: \quad & \boldsymbol{o}_t^{(p)} = \sigma\left(\langle \boldsymbol{W}_o^{(p)}, H_t^{(p-1)}\rangle + \langle \boldsymbol{U}_o^{(p)}, H_{t-1}^{(p)}\rangle + \boldsymbol{b}_o^{(p)}\right), \end{aligned} \tag{3.17}$$

the LSTM outputs at time t within the p^{th} hidden layer stems from the following equations:

$$\begin{aligned} \textit{Memory state}: \quad & \boldsymbol{c}_t^{(p)} = \boldsymbol{f}_t^{(p)} \odot \boldsymbol{c}_{t-1}^{(p)} + \boldsymbol{i}_t^{(p)} \odot tanh\left(\langle \boldsymbol{W}_c^{(p)}, H_t^{(p-1)}\rangle + \langle \boldsymbol{U}_c^{(p)}, H_{t-1}^{(p)}\rangle + \boldsymbol{b}_c^{(p)}\right), \\ \textit{Short-term output}: \quad & H_t^{(p)} = \boldsymbol{o}_t^{(p)} \odot tanh\left(\boldsymbol{c}_t^{(p)}\right), \end{aligned} \tag{3.18}$$

where $\sigma(x)$ 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 weights matrices for gates feed-forward connections and $\left\{\boldsymbol{U}_l^{(p)},\ l = f, i, o, c\right\}$ are the weights matrices for gates recurrent connections.

Definition 6. *Let us consider a dataset $\mathcal{D} = \left\{(\boldsymbol{x}_t, \boldsymbol{y}_t), \boldsymbol{x}_t \in \mathbb{R}^{N_0}, \boldsymbol{y}_t \in \mathbb{R}^{N_{P+1}}\right\}$, where $\boldsymbol{x}_t$ is the input received by the network at time t and $\boldsymbol{y}_t$ is the related target. A RNN model with a LSTM architecture is a function $f_{LSTM} : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{P+1}}$ wherein each composition arguments is defined according to Eq.(3.18) and such that:*

$$\boldsymbol{y}_t = f_{LSTM}(\boldsymbol{x}_t; \boldsymbol{\mathcal{W}}) + \gamma_t = \psi \circ \left(H_t^{(P)} \circ H_t^{(P-1)} \circ \cdots \circ H_t^{(1)}\right)(\boldsymbol{x}_t; \boldsymbol{\mathcal{W}}) + \gamma_t. \tag{3.19}$$

The Figure 3.5 exposes a graphical example for an unrolled RNN with a LSTM architecture.

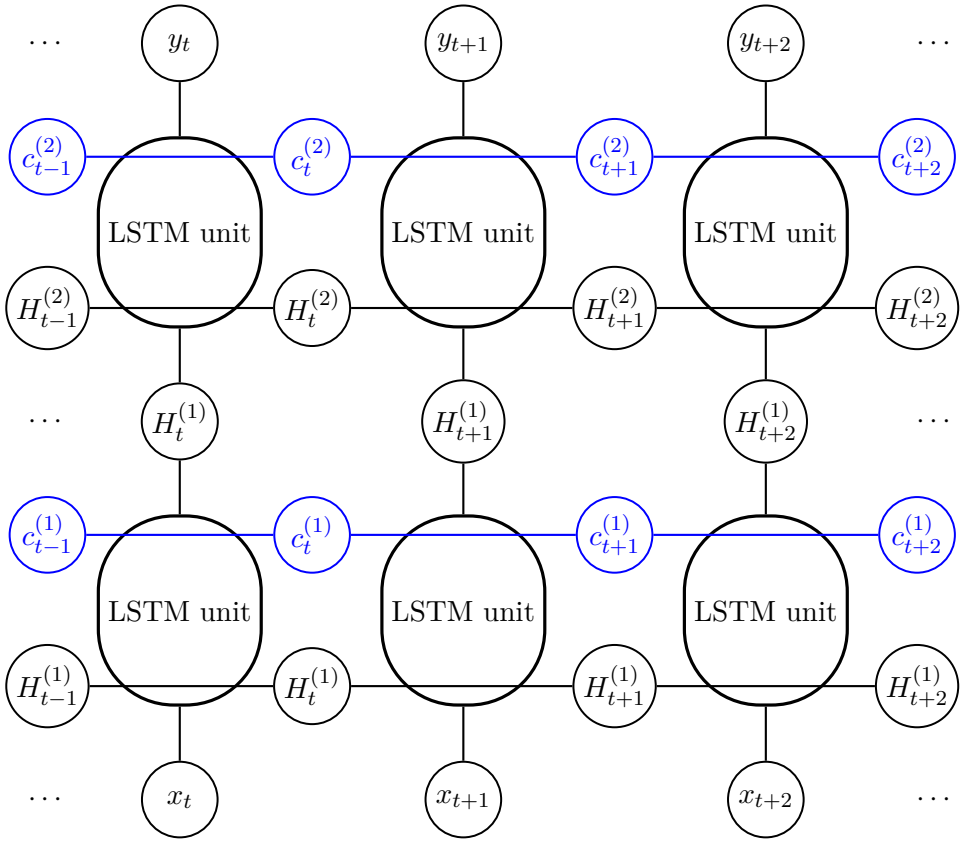


Figure 3.5. An explicative unrolled graph for a 2-hidden layered RNN with a LSTM architecture.

Remark. *Considering the learning process for a RNN with a LSTM architecture, the backward flow involved in the BPTT, for each LSTM block in each hidden layer, concerns also the memory cell changes over the time:*

$$\frac{\partial \boldsymbol{c}_t^{(p)}}{\partial \boldsymbol{c}_{t-1}^{(p)}} = \frac{\partial \boldsymbol{c}_t^{(p)}}{\partial \boldsymbol{f}_t^{(p)}} \frac{\partial \boldsymbol{f}_t^{(p)}}{\partial H_{t-1}^{(p)}} \frac{\partial H_{t-1}^{(p)}}{\partial \boldsymbol{c}_{t-1}^{(p)}} + \frac{\partial \boldsymbol{c}_t^{(p)}}{\partial \boldsymbol{i}_t^{(p)}} \frac{\partial \boldsymbol{i}_t^{(p)}}{\partial H_{t-1}^{(p)}} \frac{\partial H_{t-1}^{(p)}}{\partial \boldsymbol{c}_{t-1}^{(p)}} + \frac{\partial \boldsymbol{c}_t^{(p)}}{\partial \boldsymbol{z}_t^{(p)}} \frac{\partial \boldsymbol{z}_t^{(p)}}{\partial H_{t-1}^{(p)}} \frac{\partial H_{t-1}^{(p)}}{\partial \boldsymbol{c}_{t-1}^{(p)}}, \quad (3.20)$$

where $\boldsymbol{z}_t = tanh\left(\langle \boldsymbol{W}_c^{(p)}, H_t^{(p-1)} \rangle + \langle \boldsymbol{U}_c^{(p)}, H_{t-1}^{(p)} \rangle + \boldsymbol{b}_c^{(p)}\right)$. *We notice that at each time step the algorithm backpropagates the error through both loss function and memory unit of the next time step. Hence, we observe that if the terms* $\frac{\partial \boldsymbol{c}_t^{(p)}}{\partial \boldsymbol{c}_{t-1}^{(p)}}$ *start to converge towards zero, higher gates values could be set to reach the value close to 1, thus preventing the gradients from vanishing.*

Nowadays, predictive models based on the LSTM are the state of the art in neural forecasting. Hence, we will investigate the LSTM suitability to achieve mortality forecasts for demographic and actuarial purposes, as explained in the following Chapter 4 and Chapter 5.

Chapter 4

Life expectancy and lifespan disparity forecasting

Since nineteenth century, developed countries have been experiencing a steady improvement in mortality level, and the impact of human longevity on population dynamics has become crucial in defining social and financial policies. The achievement of longer lives has been driven by a decline in infant mortality, and by reductions in mortality at older ages after the WWII (Rau et al. (2008); Vaupel (1997)). The investigation on human lifespan boundaries leads to new approaches focused on life expectancy, bringing new perspectives into mortality forecasting. A breakthrough has been posed by Oeppen and Vaupel (2002) who introduced the concept of best-practice life expectancy (BPLE, from now on), i.e. the maximum life expectancy observed among national populations in a given calendar year. They underlined the absence of an impending limit in human life expectancy, disproving the historical estimates of the human life boundary. The constant improvement of BPLE suggests that the mortality reductions should not be viewed as a disconnected sequence of unrepeatable revolutions, but rather as a regular flow of continuous progress (Oeppen and Vaupel (2006)). Indeed, mortality developments are linked to social progress in terms of health, nutrition, education, hygiene, and medicine (Riley (2001)). However, could be substantial differences in the time of death among countries having same life excpectancy level. Hence, lifespan disparity measure, such as in Eq.(2.12), becomes crucial to a suitable inspection of lifetime evolution. Moreover, since the same information is involved in the calculation of both life expectancy and lifespan disparity, the relationship between these two indicators has been discussed by several researchers. For example, Bohk-Ewald et al. (2017) proposed to evaluate the performance of extrapolative mortality models by analysing both the average lifespan and lifespan disparity, while Rabbi and Mazzuco (2020) to adjust the time component of the LC model with the observed lifespan disparity. Aburto and van Raalte (2018) explored trends in lifespan disparity under periods of life expectancy decline by focusing on Central and Eastern Europe. They measured the relationship between life expectancy and lifespan disparity by their absolute and relative changes. Aburto et al. (2020) developed a mathematical framework to jointly explore the evolution over time of life expectancy at birth and life span equality analysing three different indicators of life span equality: life table entropy, Gini

coefficient, and coefficient of variation of the age-at-death distribution. They found a strong link between life expectancy and each lifespan inequality indicator, especially when life expectancy is less than 70 years. These studies generally investigate life expectancy and lifespan variation since birth, without considering the dispersion in the time of death conditioned on survival at a specific age, as well as the forecasting. Both life expectancy and lifespan disparity might be understood as latent variables encompassing many factors that, directly or indirectly, affect mortality dynamics. This latent behaviour should be emphasized in forecasting by incorporating both short term history and contribution from long term improvements in more recent periods. Bearing in mind the latter, we need models able to catch more in-depth the unobservable features in the historical observations. Therefore, we rely on the LSTM network to meet these needs. The goal is to provide more accurate forecasts of life expectancy and lifespan disparity with respect to other well-established models, overcoming the above limitations. In particular, our investigations contribute to the present literature by proposing a new method for forecasting life expectancy and life disparity, at birth and at age 65. In fact, LSTM allows to predict future values maintaining the noteworthy influence of the past trend and adequately reproducing it into forecasting. Therefore, the resulting future values of life expectancy and lifespan disparity should be more consistent with the historical dynamics and meet biological reasonableness criteria, first the non-linearity. Our approach mainly consists of forecasting life expectancy and life disparity independently using an univariate network. The analysis of lifespan disparity may allow us to acquire further knowledge on the life expectancy future evolution. However, these indicators may be linked by a long-term relationship (Bohk-Ewald et al. (2017), Aburto and van Raalte (2018), Aburto et al. (2020)), therefore the forecasting accuracy might take advantage by simultaneous modelling, exploiting the potential link between the dynamics of the two series. Within the recurrent neural network setting, the simultaneous forecasting of two time series requires the construction of a bivariate network. Thus, we also propose a bivariate LSTM framework aimed at forecasting life expectancy and lifespan disparity simultaneously. We provide a numerical application carried out on five countries of the world, Australia, Italy, Japan, Sweden, and the USA, to demonstrate the strong predictive power of univariate LSTM networks. We refer to other life expectancy forecasting models as comparison terms, such as the ARIMA model and the Double Gap model (DG) proposed by Pascariu et al. (2018), which applies to life expectancy but not to lifespan disparity. The ARIMA model can be considered as a benchmark for time series forecasting, while the DG model represents a prominent approach which might be seen as an improvement of ARIMA, allowing to consider the gender gap in life expectancy trend. In addition, we provide a further comparison with two extrapolative models: the LC Poisson model (Brouhns et al. (2002)) and the CoDa (Oeppen (2008)) model, which is based on the principal component analysis. The bivariate LSTM is compared to the first order Vector Autoregression model (VAR) that is often used as a benchmark for multivariate series forecasting.

4.1 Life expectancy and lifespan disparity modeling

In this section, we will describe the model used to forecast country-specific life expectancy and lifespan disparity, both independently and simultaneously, considering two ages: $x = 0$ and $x = 65$.
Let $\{e_{x,t}\}_{t=t_0}^{t_s}$ and $\left\{e_{x,t}^{\dagger}\right\}_{t=t_0}^{t_s}$, for $t_0 < t_s$, be the country-specific observed time series of life expectancy and lifespan disparity, respectively. Let $\left\{e_{x,t}, e_{x,t}^{\dagger}\right\}_{t=t_0}^{t_s}$ be the country-specific bivariate series we would like to model simultaneously. Following an appropriate rule, each series is split into a training set and a testing set, where the first one is used for fitting the model parameters, while the second one to test the model prediction and calculate the error. Let t_τ be the calendar year corresponding to the last realization on the training set. The training and testing sets for the life expectancy series are defined as follows:

$$\text{TRAINING SET} : \mathcal{TR}^{(e)} = \{e_{x,t}\}_{t=t_0}^{t_\tau}$$

$$\text{TESTING SET} : \mathcal{TS}^{(e)} = \{e_{x,t}\}_{t=t_\tau+1}^{t_s}$$

Similarly, we can define training and testing sets for lifespan disparity, $\mathcal{TR}^{(e^\dagger)}$ and $\mathcal{TS}^{(e^\dagger)}$, and for the bivariate series, $\mathcal{TR}^{(e,e^\dagger)}$ and $\mathcal{TS}^{(e,e^\dagger)}$.

4.1.1 LSTM model

In the LSTM network, aimed at forecasting life expectancy and lifespan disparity, we adopt a first-order autoregressive approach. Therefore, according to Eq.(3.19), the model is described by:

$$\begin{aligned} e_{x,t} &= f_{LSTM}^{(e)}\left(e_{x,t-1}; \mathcal{W}\right) + \gamma_t^{(e)} \qquad \text{or} \\ e_{x,t}^{\dagger} &= f_{LSTM}^{(e^\dagger)}\left(e_{x,t-1}^{\dagger}; \mathcal{W}\right) + \gamma_t^{(e^\dagger)} \qquad \text{or} \\ \left[e_{x,t}, e_{x,t}^{\dagger}\right] &= f_{LSTM}^{(e,e^\dagger)}\left\{\left[e_{x,t-1}, e_{x,t-1}^{\dagger};\right]; \mathcal{W}\right\} + \gamma_t^{(e,e^\dagger)}. \end{aligned} \tag{4.1}$$

where $\gamma_t^{(\cdot)}$ is a zero mean error. The set of functions $f_{LSTM}^{(\cdot)}$ is the map linking life expectancy or lifespan disparity or both at an annual pace. In a first-order autoregressive approach, the network learns at each time step the relationship between consecutive values on the training set and, according to the same logic, predicts the future values on the testing set. This process is optimized using a L2 loss function:

$$\begin{aligned} &\min_{\mathcal{W}} \frac{1}{2} \sum_{t=t_0}^{t_\tau} \left(e_{x,t} - \hat{e}_{x,t}\right)^2 \qquad \text{or} \\ &\min_{\mathcal{W}} \frac{1}{2} \sum_{t=t_0}^{t_\tau} \left(e_{x,t}^{\dagger} - \hat{e}_{x,t}^{\dagger}\right)^2 \qquad \text{or} \\ &\min_{\mathcal{W}} \frac{1}{2} \sum_{t=t_0}^{t_\tau} \left\{\left[e_{x,t}, e_{x,t}^{\dagger}\right] - \left[\hat{e}_{x,t}, \hat{e}_{x,t}^{\dagger}\right]\right\}^2, \end{aligned} \tag{4.2}$$

where $\boldsymbol{\mathcal{W}} = \left\{\left\{\boldsymbol{W}_l^{(p)}\right\} \cup \left\{\boldsymbol{U}_l^{(p)}\right\} \cup \left\{\boldsymbol{b}_l^{(p)}\right\},\ l = f, i, o, c,\ p = 1, \dots P\right\}$ is the LSTM parameters set.

LSTM in a demographical framework

We are now going to connect the concepts from RNN, exposed in Chapter 3, to the input data used in the application, aiming at creating a bridge between RNN and demography. In the following, we will only refer to life expectancy (the extension to life disparity and to the bivariate case is straightforward). In our model, the input received by the network at state t is life expectancy at a given age x, i.e. $\boldsymbol{x}_t \equiv \boldsymbol{e}_{x,t}$. The output of the network at state t is the life expectancy at time $t+1$, consistently with the first-order autoregressive pattern, that is $\boldsymbol{y}_t \equiv \boldsymbol{e}_{x,t+1}$. Therefore, following the Eq.(3.19), $\boldsymbol{e}_{x,t+1} \equiv \psi\left(\langle \boldsymbol{W}_e^{(P+1)}, H_t^{(P)}\rangle + \boldsymbol{b}_e^{(P+1)}\right) + \gamma_{t+1}$ is the theoretical relationship defining the life expectancy at year $t+1$, given the life expectancy at year t, and the LSTM block processing. The final output of LSTM, after the estimation procedure, which implies to estimate the weights, becomes:

$$\hat{\boldsymbol{e}}_{x,t+1} = \psi\left(\langle \hat{\boldsymbol{W}}_e^{(P+1)}, H_t^{(P)}\rangle + \hat{\boldsymbol{b}}_e^{(P+1)}\right) \tag{4.3}$$

Where $\hat{\boldsymbol{e}}_{x,t+1}$ is the life expectancy estimation resulting from the application of the estimated parameters through the BPTT described in Section 3.3.2.

4.1.2 Other models

The actuarial and demographic literature provides a wide variety of mortality models. In our analysis, the performance of the univariate LSTM network is compared to the ARIMA, DG, LC and CoDa models. LSTM, ARIMA and DG models allow to directly work with the life expectancy and life disparity time series, without passing through an extrapolative stochastic models which provides the mortality rates used to calculate such demographical indicators. However, we also consider two extrapolative models: LC, which is probably the most used by practitioners and CoDa, which forecasts the life table distributions of deaths using principal component analysis in a compositional data pattern. While the performance of the bivariate LSTM is compared to the VAR model. A brief description of these models is reported in the follow.

ARIMA model is a well-established approach that can be considered as the reference model for the forecast of mortality. This model has three parameters p, d, q, representing respectively the auto-regressive, the differencing and the moving average order. The generic ARIMA(p, d, q) for life expectancy takes the following form:

$$\bigtriangledown^d e_{x,t} = \delta + \sum_{i=1}^{p} \phi_i \bigtriangledown^d e_{x,t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \tag{4.4}$$

where δ is the drift process, ϕ_i are the autoregressive parameters, ϵ_t the error terms normally distributed with zero mean and variance σ_ϵ^2 and θ_j are the moving average parameters.

Double Gap model is one of the most recent and most prominent approaches in forecasting life expectations. It provides the life expectancy forecasts for both the genders by modeling first the gap between country-specific female life expectancy, e^f,

and female BPLE (the female world record level), e^{bp}, and then the gap between male life expectancy, e^m, and female life expectancy, e^f, in a given country. Therefore, the future female life expectancy at age x and time t for a given country is calculated as the difference between the future $e^{bp}_{x,t}$ and the predicted values of the gap, $D_{x,t}$, between the country-specific female life expectancy and the female best-practice trend: $e^f_{x,t} = e^{bp}_{x,t} - D_{x,t}$. While, the future male life expectancy is calculated as the difference between the future female life expectancy and the predicted values of the gap, $G_{x,t}$, between the country-specific female and male life expectancy: $e^m_{x,t} = e^f_{x,t} - G_{x,t}$. The first gap, $D_{x,t}$, is modeled according to a traditional ARIMA(p, d, q):

$$\bigtriangledown^d D_{x,t} = \delta^{(1)} + \sum_{i=1}^{p} \phi_i^{(1)} \bigtriangledown^d D_{x,t-i} + \epsilon_t^{(1)} + \sum_{j=1}^{q} \theta_j^{(1)} \epsilon_{t-j}^{(1)} \tag{4.5}$$

where $\delta^{(1)}$ is the drift process, $\phi_i^{(1)}$ are the autoregressive parameters, $\epsilon_t^{(1)}$ the error terms normally distributed with zero mean and variance $\sigma^2_{\epsilon^{(1)}}$ and $\theta_j^{(1)}$ are the moving average parameters. The second gap, $G_{x,t}$, is modeled by a linear model and a random walk without drift:

$$G^*_{x,t} = \begin{cases} \beta_0 + \beta_1 \cdot G_{x,t-1} + \beta_2 \cdot G_{x,t-2} + \beta_3 \cdot \left(e^f_{x,t} - \tau\right)^+ + \epsilon_t^{(2)} & if\ e^f_{x,t} < A, \\ G_{x,t-1} + \epsilon_t^{(3)} & otherwise \end{cases}$$

Where τ and A are fixed levels calculated on historical data by maximizing the resulting maximum likelihoods of the linear model over integer values of τ and A (see Pascariu et al. (2018) for further details on the estimation procedure). The algorithm is implemented by the function available in the R package *MortalityGap*. The DG model is not applied in the case of lifespan disparity due to the non-existence of a best practice for disparity measures.

LC model works with the linear extrapolations of age-specific mortality rates on the logarithmic scale. Its first formulation (Lee and Carter (1992)) based on the latent approach using SVD has been widely improved across time. We use the extension proposed by Brouhns et al. (2002), widely described in Section 2.2.1.

CoDa model was proposed by Oeppen (2008) and suggests forecasting $d_{x,t}$ using principal component analysis in a compositional data pattern, following the Lee and Carter's original approach:

$$clr(d_{x,t} \ominus \alpha_x) = \kappa_t \beta_t + \epsilon_{x,t} \tag{4.6}$$

Where clr is one of the log-ratio representations of compositional data. According to Bergeron-Boucher et al. (2017) it is defined as the logarithm of the composition divided by its geometric mean: $clr(d_{x,t} = ln(\frac{d_{x,t}}{g_t}))$, where g_t is the geometric mean of the age-composition at time t. The $\ominus$ operator represents the standard operation in compositional data analysis consisting in perturbing a composition by the inverse element of another composition. It is used to center the matrix while retaining the constant sum. The parameter, obtained by SVD, are $\kappa_t = u_t s$ and $\beta_t = v_x$, where s is the leading singular value, u_t and v_x refer to period and age components

that are respectively the first left and the first right-singular vectors, and the α_x is the age-specific geometric mean of $d_{x,t}$ over time. Then, the model provides the age at death distribution through the closing procedure $C[\cdot]$ used to transform the estimates into compositional data summing up to the initial constant:

$$d_{x,t} = \alpha_x \otimes C[e^{\kappa_t \beta_x + \epsilon_{x,t}}] \tag{4.7}$$

Vector autoregression model, also known as VAR, is one of the most applied models in empirical economics and finance for the analysis of multivariate time series. It is a multivariate stochastic process that can be used to model the joint evolution of two or more series over time. We refer to the first-order VAR model which consists in jointly modeling life expectancy and life disparity as follows:

$$e_{x,t} = \phi_0 + \phi_1 e_{x,t-1} + \phi_2 e^{\dagger}_{x,t-1} + \epsilon^{e}_{x,t} \tag{4.8}$$

$$e^{\dagger}_{x,t} = \theta_0 + \theta_1 e_{x,t-1} + \theta_2 e^{\dagger}_{x,t-1} + \epsilon^{e^{\dagger}}_{x,t} \tag{4.9}$$

Where ϕ_i and θ_i (for $i = 0, 1, 2$) are the model parameters, and the errors $\epsilon^{e}_{x,t}$ and $\epsilon^{e^{\dagger}}_{x,t}$ follow a bivariate normal distribution with a zero mean vector and a constant covariance matrix.

4.2 Empirical investigation and results

In the numerical application, we consider historical mortality data collected by gender from the HMD (www.mortality.org) for Australia, Italy, Japan, Sweden, and USA.

It is well known that mortality modelling is a process that should fulfil some qualitative criteria, robustness, among others. Thus, the forecast should not be too sensitive towards the selected period's choice, but it should be consistent with historical data. Therefore, in our analysis, we will carry out an out-of-sample test considering the same forecast horizon for two different overlapping estimation periods: 1938-1999 and 1947-1999. The time frame 2000-2014 is then used as evaluation chunk. In this way, we obtain a sufficient size for training and testing sets in both the time frames, according to the common splitting rule: 80% and 20%. Finally, to assess the models' accuracy, we calculate the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Before training the LSTM for all countries and both genders, we will implement a preliminary tuning process to identify the optimal hyper-parameters, such as mini-batch size, epochs, and neurons number for each hidden layer. For this purpose, we will select a finite set for each hyper-parameter, exploring the specification minimizing the loss function. The best combination obtained in the training phase is used to calibrate LSTM in the forecasting one. The mini-batch size is equal to the number of training samples in one forward/backward pass before updating the model weights. In our case, the mini-batch size is equal to 1, as our input data have been arranged into a column vector, where each row represents the life expectancy at a generic time t. Therefore, we need to compute the weight's update for each one-time step. It is worth noting that a batch size greater than 1 is not

consistent with our autoregressive framework based on one-order of differentiation. Not least, the literature suggests that the use of small batch sizes improves the out-of-sample performance and the optimization convergence (LeCun et al. (2012), Keskar et al. (2016)) requiring small memory (then gaining efficiency) by exploiting memory locality. The architectures with a single hidden layer work better than others, and the number of neurons and epochs depends on the specific-country data. In our model, the loss function is minimized over the neural network weights using the Adadelta (Zeiler (2012)), a variant of the Stochastic Gradient Descent (SGD) method. We use the Rectified Linear Unit (ReLU) (Glorot et al. (2011)) as output layer activation function, ψ, that outperformed the other tested functions.

The LSTM performances are compared to the models presented in Section 4.1.2. Therefore, we will compare the univariate LSTM to the best ARIMA(p, d, q), DG, LC and CoDa models, while the bivariate LSTM is compared to the VAR model. All these models are trained aiming at generating life expectancy and lifespan disparity projections on the testing set. The following goodness of fit measures are used to evaluate the forecasting quality:

- *Mean Absolute Error*

$$MAE = \frac{\sum_{t=t_\tau+1}^{t_s} \mid e_{x,t} - \hat{e}_{x,t} \mid}{(t_s - t_\tau - 1)}, \tag{4.10}$$

- *Root Mean Square Error*

$$RMSE = \sqrt{\frac{\sum_{t=t_\tau+1}^{t_s} (e_{x,t} - \hat{e}_{x,t})^2}{(t_s - t_\tau - 1)}}. \tag{4.11}$$

where $\hat{e}_{x,t}$ represents the future estimation of life expectancy produced by the models. These measures are also used to evaluate the forecasting of the lifespan disparity $e^\dagger_{x,t}$ and the bivariate series $\left[e_{x,t}, e^\dagger_{x,t}\right]$.
All the experiments were performed using the R packages: *keras* and *tensorflow* (version 1.13.1) for LSTM, *forecast* for ARIMA, *MortalityGap* for DG model, *MortalityForecast* for CoDA model, *StMoMo* for LC model and *vars* for VAR model.

4.2.1 Results of the out-of-sample test: independent modeling

This section will provide the estimated future life expectancy and lifespan disparity at birth and age 65 from separate modeling. The results are provided for five countries, Australia, Italy, Japan, Sweden, and USA, and both genders over the testing period. As already pointed out, mortality models should also satisfy biological reasonableness criteria, with respect to both the short and the long term dynamics must be biologically consistent. Hence, we will perform the considered models on two different time windows, carrying out a sensitivity analysis based on two periods according to the historical demographic changes. The longest period (1938-1999) covers the WWII mortality shocks, which is excluded in the shortest one (1947-1999). Japan is not considered in the period starting from 1938 as data were made available starting from 1947. Finally, to illustrate our findings, the results are displayed in a tabular form, posing the associated graphical depiction in Appendix A.

Results for life expectancy: $e_{0,t}$ and $e_{65,t}$

Table 4.1 shows MAE and RMSE values of life expectancy at birth for both the estimation periods and each country by gender. Overall, the univariate LSTM provides remarkably high accuracy compared to the other models, overperforming in 72% of cases. Our model is only beaten in case of Japan females, Italy, and US females for both periods, however reaching the second-best performance. By a graphical perspective (see Figure A.1 in Appendix A) we generally observe that when life expectancy does not experience any trend changes, the reduction of mortality compression does not provide any evidence of imminent interruption (Bohk-Ewald et al. (2017)) as detected by lifespan disparity (see Fig. Figure A.3 in Appendix A).

Table 4.1. Out-of-sample test for $e_{0,t}$: MAE and RMSE for LSTM, ARIMA, DG, LC and CoDa model by country and gender. Years 2000-2014. Fitting periods: 1938-1999 (columns 3-6) and 1947-1999 (columns 7-10).

Country	Model	Fitting period: 1938-1999				Fitting period: 1947-1999			
		Female		Male		Female		Male	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
	ARIMA	0.3118	0.4149	0.8111	0.8504	0.2167	0.2844	0.2901	0.3216
	DG	0.3139	0.4175	0.2693	0.2896	0.1945	0.2277	0.3219	0.3468
Australia	LSTM	**0.1139**	**0.1412**	**0.1485**	**0.1895**	**0.1110**	**0.1362**	**0.1407**	**0.1804**
	LC	0.2525	0.2806	1.0740	1.2133	0.2869	0.3204	1.0368	1.1640
	CoDa	0.1347	0.1655	1.1936	1.2763	0.1304	0.1639	1.0022	1.0629
	ARIMA	1.5759	1.8872	0.9157	1.0819	0.3434	0.4455	0.1768	0.2155
	DG	0.2986	0.3836	0.2355	0.2697	0.2314	0.2722	0.2209	0.2444
Italy	LSTM	0.1914	0.2304	**0.1396**	**0.1767**	0.2104	0.2587	**0.1758**	**0.2124**
	LC	**0.1663**	**0.2068**	1.8194	1.9259	**0.1518**	**0.1969**	1.5136	1.6463
	CoDa	0.4275	0.5507	0.9763	1.0531	0.4156	0.5356	1.0985	1.1880
	ARIMA	0.4305	0.4659	0.4760	0.5484	0.4467	0.4672	0.2696	0.3058
	DG	0.4305	0.4659	0.1659	0.1888	0.4467	0.4671	0.3983	0.4232
Sweden	LSTM	**0.0773**	**0.0964**	**0.0574**	**0.0703**	**0.0752**	**0.1000**	**0.0598**	**0.0718**
	LC	0.1761	0.1973	0.9698	1.0815	0.0823	0.1149	1.0199	1.1245
	CoDa	0.4079	0.4574	0.9496	1.0627	0.6612	0.7449	0.8578	0.9571
	ARIMA	0.7358	0.8898	0.1892	0.2449	0.6822	0.8165	0.1455	0.1845
	DG	0.7358	0.8898	1.3553	1.5444	0.6821	0.8164	1.0669	1.2119
USA	LSTM	0.2466	0.2939	**0.1140**	**0.1381**	0.3522	0.4279	**0.1137**	**0.1375**
	LC	**0.1173**	**0.1451**	0.5017	0.5950	0.3847	0.4096	0.7549	0.8336
	CoDa	0.2390	0.2688	0.4505	0.5432	**0.1038**	**0.1266**	0.4529	0.5451
	ARIMA	-	-	-	-	**0.1712**	**0.2291**	1.2220	1.4085
	DG	-	-	-	-	0.5569	0.5894	0.3721	0.4210
Japan	LSTM	-	-	-	-	0.3342	0.3694	**0.2252**	**0.2662**
	LC	-	-	-	-	0.6543	0.7650	0.4330	0.5068
	CoDa	-	-	-	-	1.2086	1.5032	0.9961	1.2106

The results of the backtesting exercise for life expectancy at age 65 are reported in Table 4.2 for both the estimation periods and each country by gender. Also, by graphical analysis, the univariate LSTM seems to well catch the nonlinearity of the future mortality trend, showing its aptitude to better represent the decreasing dynamics of mortality at age 65. In this case, our model overperforms all the other models in 69% of cases. Overall, e_{65} shows a nonlinear behavior and irregular patterns, especially for males, and the gain provided by LSTM is more evident if compared to the other models (see Figure A.2 in Appendix A). Indeed, one of the main features of LSTM is to reproduce in the projections the irregular patterns of a phenomenon observed in the past. In particular, in the case of US females, we

speculate that the historical periods 1973-1979 and 1989-1992 seem to strongly affect the LSTM weights, by reproducing in the forecasts the sudden longevity growth after the stagnation following the WWII.

Table 4.2. Out-of-sample test for $e_{65,t}$: MAE and RMSE for LSTM, ARIMA, DG, LC and CoDa model by country and gender. Years 2000-2014. Fitting periods: 1938-1999 (columns 3-6) and 1947-1999 (columns 7-10).

Country	Model	Fitting period: 1938-1999				Fitting period: 1947-1999			
		Female		Male		Female		Male	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
	ARIMA	0.2928	0.3271	0.1501	0.1811	0.2277	0.2587	0.2846	0.3664
	DG	0.3205	0.3564	0.8552	0.9366	0.2817	0.3151	0.8163	0.8928
Australia	LSTM	**0.0804**	**0.0999**	**0.0764**	**0.0998**	**0.0782**	**0.0975**	**0.0764**	**0.0996**
	LC	0.4688	0.4962	1.2422	1.3336	0.3583	0.3878	1.0972	1.189
	CoDa	0.1842	0.2143	0.9851	1.0834	0.1295	0.1569	0.8639	0.9467
	ARIMA	0.2604	0.2954	1.0379	1.1296	0.2732	0.3059	1.0918	1.2212
	DG	0.2604	0.2954	0.5539	0.5946	0.2732	0.3059	0.5669	0.6100
Italy	LSTM	**0.1578**	**0.1972**	**0.1529**	**0.1803**	**0.1591**	**0.2022**	**0.1576**	**0.1893**
	LC	0.4672	0.4936	1.4899	1.5551	0.3479	0.3798	1.2268	1.2999
	CoDa	0.2347	0.2765	0.7775	0.8372	0.2437	0.2878	0.8047	0.8681
	ARIMA	0.1361	0.1705	0.8900	0.9902	0.2786	0.3042	0.7178	0.8177
	DG	**0.1007**	0.1384	0.4020	0.4703	0.2786	0.3042	0.2381	0.2836
Sweden	LSTM	0.1058	0.1357	**0.0828**	**0.1015**	0.1147	0.1455	**0.0861**	**0.1032**
	LC	0.1095	**0.1248**	1.15	1.2278	**0.0541**	**0.0637**	0.9145	1.0011
	CoDa	0.2121	0.2575	0.8872	0.9903	0.1015	0.1181	0.7832	0.8718
	ARIMA	0.2529	0.2923	0.9051	1.0138	0.3112	0.3753	0.6753	0.7572
	DG	0.2616	0.2734	0.3081	0.3449	**0.1775**	**0.2047**	0.7755	0.8431
USA	LSTM	0.6146	0.7095	**0.2773**	**0.3109**	0.5283	0.6094	**0.2485**	**0.2963**
	LC	**0.2212**	**0.2512**	0.9987	1.093	0.2601	0.2979	0.9337	1.0345
	CoDa	0.1915	0.2226	0.9193	1.0245	0.2732	0.3267	0.8893	0.9908
	ARIMA	-	-	-	-	0.2804	0.3762	**0.1815**	**0.2073**
	DG	-	-	-	-	**0.2590**	0.3287	0.3436	0.4494
Japan	LSTM	-	-	-	-	0.2928	**0.3189**	0.2173	0.2392
	LC	-	-	-	-	0.5048	0.5775	0.3906	0.4240
	CoDa	-	-	-	-	0.5747	0.7193	0.4275	0.5262

Results for lifespan disparity: $e^{\dagger}_{0,t}$ and $e^{\dagger}_{65,t}$

The results of the out-of-sample test for $e^{\dagger}_0$ are shown in Table 4.3 for both the estimation periods and each country by gender. We can observe that for lifespan disparity at birth, the univariate LSTM outperforms the other models in 83% of the cases. Our model does not reach the best performance only for Australia females for both periods, where however, the prediction errors are incredibly low. The most remarkable out-of-sample result for $e^{\dagger}_0$ is provided by US females. Such a result shows a decreasing trend periodically interrupted by stagnation periods. In this case, the LSTM weights are probably influenced by the two short periods of stagnation, 1960-1970 and 1985-1990, that are reproduced in the projections, allowing to reach a high level of accuracy (see Figure A.3). The same speculation holds for US males where the stagnation periods are more evident. We assume that a similar forecast behavior is challenging to be achieved by a canonical model that could ignore the long-short term dynamics.

Table 4.3. Out-of-sample test for $e^{\dagger}_{0,t}$: MAE and RMSE for LSTM, ARIMA, LC and CoDa model by country and gender. Years 2000-2014. Fitting periods: 1938-1999 (columns 3-6) and 1947-1999 (columns 7-10).

Country	Model	Fitting period: 1938-1999				Fitting period: 1947-1999			
		Female		Male		Female		Male	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	ARIMA	0.0916	0.0985	0.1058	0.1348	**0.0718**	0.0994	0.1426	0.1618
	LSTM	0.0906	0.1018	**0.0631**	**0.0756**	0.0794	0.0929	0.0880	0.1103
	LC	**0.0729**	0.0949	0.0844	0.1150	0.1931	0.2041	**0.0764**	**0.0981**
	CoDa	0.0757	**0.0864**	0.2147	0.2270	0.0845	**0.0924**	0.1601	0.1785
Italy	ARIMA	0.3209	0.3709	0.9104	1.0565	0.5013	0.5810	0.3444	0.3955
	LSTM	**0.0545**	**0.0643**	**0.0702**	**0.0866**	**0.1362**	**0.1528**	**0.0646**	**0.0827**
	LC	0.2100	0.2222	0.3984	0.4641	0.1649	0.1792	0.3860	0.4513
	CoDa	0.2451	0.2833	0.2807	0.3247	0.2331	0.2702	0.3533	0.4073
Sweden	ARIMA	0.2438	0.2666	0.3020	0.3390	0.2944	0.3262	0.2166	0.2442
	LSTM	**0.0598**	**0.0736**	**0.0468**	**0.0550**	**0.0572**	**0.0669**	**0.0439**	**0.0565**
	LC	0.1204	0.1291	0.0559	0.0734	0.1025	0.1126	0.1798	0.1955
	CoDa	0.2195	0.2398	0.0612	0.0729	0.2379	0.2634	0.0771	0.0920
USA	ARIMA	0.5569	0.6499	0.8677	0.9733	0.4795	0.5508	0.5935	0.6626
	LSTM	**0.0457**	**0.0547**	**0.0497**	**0.0603**	**0.0517**	**0.0561**	**0.0529**	**0.0628**
	LC	0.1670	0.2006	0.3277	0.3742	0.3281	0.3514	0.1659	0.1931
	CoDa	0.3269	0.3885	0.4529	0.5027	0.1970	0.2433	0.4246	0.4715
Japan	ARIMA	-	-	-	-	0.0635	0.0868	0.2265	0.2863
	LSTM	-	-	-	-	**0.0573**	**0.0760**	**0.0726**	**0.0799**
	LC	-	-	-	-	1.1321	1.1351	0.7256	0.7276
	CoDa	-	-	-	-	0.9129	0.9958	0.5617	0.6017

The results for $e^{\dagger}_{65}$ are shown in Table 4.4 for both the estimation periods and each country by gender. The MAE and RMSE values highlight the LSTM ability to detect the hidden patterns of noisy time series, outperforming the other models in 89% of the cases. Indeed, $e^{\dagger}_{65}$ is characterized by a high variability level, since it summarizes disparity across individuals who have already survived to age 65. In case of US male (Figure A.4 in Appendix A), the LSTM prediction is not consistent with the historical values and might be influenced by short-term stagnation dynamics.

4.2.2 Results of the out-of-sample test: simultaneous modeling

The estimates of future life expectancy and lifespan disparity at birth and age 65 given by the out-of-sample test, resulting from the simultaneous modeling (namely LSTM-2D) are shown in the following tables, compared with the first-order VAR model that is used as a benchmark for multivariate series forecasting. The results for e_0 and e_{65} are respectively reported in Table 4.5 and Table 4.6 for both the estimation periods and each country by gender. We note that the LSTM-2D outperforms the VAR model for life expectancy at birth in 86% of the cases, while this percentage drops to 47% at age 65. Similar behavior can be observed for lifespan disparity (Tables 4.7 and 4.8), where LSTM-2D obtains the best performance in 78% of the cases at birth and 58% at age 65. In some few cases, the bivariate network provides lower errors if compared to the other models (univariate and bivariate), especially for life expectancy at birth: for example, Italy females in the fitting period 1947-1999, Sweden males in the fitting period 1938-1999 and for e_{65} Japan females in the fitting period 1947-1999.

Table 4.4. Out-of-sample test for $e^{\dagger}_{65,t}$: MAE and RMSE for LSTM, ARIMA, LC and CoDa model by country and gender. Years 2000-2014. Fitting periods: 1938-1999 (columns 3-6) and 1947-1999 (columns 7-10).

Country	Model	Fitting period: 1938-1999				Fitting period: 1947-1999			
		Female		Male		Female		Male	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	ARIMA	0.2077	0.2265	0.1743	0.1883	0.2119	0.2304	0.0905	0.1048
	LSTM	**0.0466**	**0.0525**	**0.0399**	**0.0525**	**0.0435**	**0.0539**	**0.0469**	**0.0596**
	LC	0.3093	0.3176	0.1887	0.1962	0.2895	0.2976	0.1354	0.1450
	CoDa	0.1148	0.1285	0.0771	0.0932	0.1128	0.1263	0.0646	0.0789
Italy	ARIMA	0.1284	0.1499	0.2735	0.3149	0.1405	0.1636	0.2755	0.3111
	LSTM	**0.0505**	**0.0605**	**0.0441**	**0.0583**	**0.0498**	**0.0595**	**0.0425**	**0.0532**
	LC	0.0988	0.1154	0.4352	0.4586	0.0727	0.0856	0.3720	0.3943
	CoDa	0.0566	0.0743	0.2217	0.2633	0.0659	0.0841	0.2174	0.2592
Sweden	ARIMA	0.0733	0.0836	0.1412	0.1548	0.1379	0.1494	0.1408	0.1554
	LSTM	**0.0286**	**0.0342**	**0.0307**	**0.0388**	**0.0290**	**0.0353**	**0.0308**	**0.0388**
	LC	0.0705	0.0788	0.2767	0.2831	0.0541	0.0637	0.2316	0.2377
	CoDa	0.0814	0.0937	0.0668	0.0821	0.1015	0.1182	0.0701	0.0860
USA	ARIMA	0.0733	0.0826	**0.1033**	**0.1341**	0.0599	0.0693	**0.1074**	**0.1411**
	LSTM	**0.0439**	**0.0539**	0.1221	0.1613	**0.0446**	**0.054**	0.1153	0.1532
	LC	0.1872	0.1948	0.1673	0.1879	0.2361	0.2407	0.1295	0.1510
	CoDa	0.0634	0.0860	0.1092	0.1462	0.0470	0.0585	0.1158	0.1543
Japan	ARIMA	-	-	-	-	0.0896	0.1050	0.1923	0.2363
	LSTM	-	-	-	-	**0.0647**	**0.0765**	**0.0773**	**0.0912**
	LC	-	-	-	-	0.2086	0.2174	0.1232	0.1519
	CoDa	-	-	-	-	0.3542	0.3695	0.1402	0.1461

Our empirical analysis shows that the simultaneous modeling of life expectancy and lifespan disparity may be not suitable, however, it leads us to speculate that only life expectancy at birth projections may take advantage of a simultaneous forecasting with life disparity.

Table 4.5. Out-of-sample test for $e_{0,t}$: MAE and RMSE for LSTM-2D and VAR, by country and gender. Years 2000-2014. Fitting periods: 1938-1999 (columns 3-6) and 1947-1999 (columns 7-10).

Country	Model	Fitting period: 1938-1999				Fitting period: 1947-1999			
		Female		Male		Female		Male	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	LSTM-2D	**0.1331**	**0.1674**	**0.4488**	**0.5152**	**0.1497**	**0.1916**	0.5602	0.6343
	VAR	0.6999	0.7269	0.8535	0.8658	0.2134	0.2520	**0.4687**	**0.5386**
Italy	LSTM-2D	**0.2442**	**0.3019**	**0.4417**	**0.4991**	**0.1235**	**0.1646**	**0.1392**	**0.1654**
	VAR	0.2957	0.3409	2.4970	2.6610	0.2957	0.3409	2.3200	2.5000
Sweden	LSTM-2D	**0.2437**	**0.2851**	**0.0488**	**0.0605**	**0.0909**	**0.1180**	**0.1001**	**0.1176**
	VAR	0.8205	0.9675	1.9560	2.1530	0.4442	0.5284	0.8585	0.9164
USA	LSTM-2D	**0.1786**	**0.2257**	0.2413	0.2754	**0.1216**	**0.1488**	**0.1960**	**0.2262**
	VAR	0.5654	0.6964	**0.1471**	**0.1824**	0.6170	0.7516	0.2121	0.2993
Japan	LSTM-2D	-	-	-	-	**0.3225**	0.3734	**0.5602**	**0.6343**
	VAR	-	-	-	-	0.3320	**0.3685**	1.2200	1.3300

Table 4.6. Out-of-sample test for $e_{65,t}$: MAE and RMSE for LSTM-2D and VAR, by country and gender. Years 2000-2014. Fitting periods: 1938-1999 (columns 3-6) and 1947-1999 (columns 7-10).

Country	Model	Fitting period: 1938-1999				Fitting period: 1947-1999			
		Female		Male		Female		Male	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	LSTM-2D	**0.0986**	**0.1144**	**0.3142**	**0.3618**	**0.0826**	**0.1127**	**0.2934**	**0.3339**
	VAR	1.1765	1.3661	1.2694	1.6620	1.1546	1.3594	1.2624	1.6553
Italy	LSTM-2D	0.4061	0.4615	**0.3784**	**0.4227**	0.6179	0.6510	**0.4195**	**0.4622**
	VAR	**0.2985**	**0.3436**	1.3782	1.4669	**0.2219**	**0.2586**	1.2280	1.2955
Sweden	LSTM-2D	0.6780	0.7144	0.4292	0.4767	0.5264	**0.5452**	0.6009	0.6173
	VAR	**0.6617**	**0.7070**	**0.2310**	**0.2425**	**0.5137**	0.5524	**0.3113**	**0.4375**
USA	LSTM-2D	0.3981	0.4398	0.6644	0.7180	0.8393	0.9477	0.8614	0.9118
	VAR	**0.2985**	**0.3436**	**0.1725**	**0.2435**	**0.3174**	**0.3729**	**0.3627**	**0.3983**
Japan	LSTM-2D	-	-	-	-	**0.1317**	**0.1365**	**0.2139**	**0.2490**
	VAR	-	-	-	-	0.4882	0.5486	0.6607	0.6680

Table 4.7. Out-of-sample test for $e^{\dagger}_{0,t}$: MAE and RMSE for LSTM-2D and VAR, by country and gender. Years 2000-2014. Fitting periods: 1938-1999 (columns 3-6) and 1947-1999 (columns 7-10).

Country	Model	Fitting period: 1938-1999				Fitting period: 1947-1999			
		Female		Male		Female		Male	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	LSTM-2D	0.1247	0.1363	**0.1406**	**0.1620**	0.1450	0.1569	**0.1846**	**0.2054**
	VAR	**0.0929**	**0.1027**	0.1627	0.2000	**0.1084**	**0.1150**	0.2076	0.2670
Italy	LSTM-2D	**0.2178**	**0.2416**	0.5282	0.5353	**0.2840**	**0.3237**	**0.0630**	**0.0782**
	VAR	0.7315	0.7936	**0.0830**	**0.1004**	0.7315	0.7936	0.2318	0.2737
Sweden	LSTM-2D	**0.1868**	**0.2190**	**0.0393**	**0.0445**	**0.0950**	**0.1194**	**0.0584**	**0.0701**
	VAR	0.3118	0.3414	0.1111	0.1371	0.2304	0.2496	0.2177	0.2375
USA	LSTM-2D	**0.1192**	**0.1436**	**0.2004**	**0.2306**	**0.1199**	**0.1406**	**0.2228**	**0.2615**
	VAR	0.1478	0.1557	0.5467	0.6241	0.2365	0.2424	0.5103	0.5956
Japan	LSTM-2D	-	-	-	-	0.2758	0.2902	**0.0875**	**0.0995**
	VAR	-	-	-	-	**0.1818**	**0.2353**	0.0897	0.1107

Table 4.8. Out-of-sample test for $e^{\dagger}_{65,t}$: MAE and RMSE for LSTM-2D and VAR, by country and gender. Years 2000-2014. Fitting periods: 1938-1999 (columns 3-6) and 1947-1999 (columns 7-10).

Country	Model	Fitting period: 1938-1999				Fitting period: 1947-1999			
		Female		Male		Female		Male	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Australia	LSTM-2D	**0.1781**	**0.1875**	**0.3494**	**0.3727**	**0.1628**	**0.1713**	**0.2981**	**0.3175**
	VAR	0.5819	0.6067	0.5209	0.5642	0.5427	0.5669	0.4879	0.5262
Italy	LSTM-2D	**0.2813**	0.3151	0.5826	0.6568	**0.2225**	**0.2483**	0.4985	0.5704
	VAR	0.2895	**0.2940**	**0.1536**	**0.1842**	0.2695	0.2990	**0.1417**	**0.1694**
Sweden	LSTM-2D	**0.1042**	**0.1118**	0.2810	0.3086	**0.1074**	**0.1151**	0.2179	0.2393
	VAR	0.2735	0.2841	**0.2687**	**0.2943**	0.2519	0.2624	**0.1438**	**0.1645**
USA	LSTM-2D	**0.1459**	**0.1582**	0.2301	0.2328	**0.1476**	**0.1574**	0.1277	0.1346
	VAR	0.2895	0.2940	**0.0667**	**0.0788**	0.2735	0.2783	**0.1048**	**0.1310**
Japan	LSTM-2D	-	-	-	-	**0.1174**	**0.1131**	0.1348	0.1573
	VAR	-	-	-	-	0.1442	0.1281	**0.1210**	**0.1384**

Chapter 5

A Neural Network integration of stochastic mortality models

Since the second half of the 20th century, mortality has exhibited notable improvements 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 made efforts to improve the socio-economic development, health system, and lifestyle of their populations, impacting on 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 of accurate forecasting to address longevity risk and adequately pricing the annuities products has led actuaries towards more sophisticated extrapolative methods, in a stochastic environment, see for instance Lee and Carter (1992), Brouhns et al. (2002), Renshaw and Haberman (2006), Cairns et al. (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 the model functional form and its parametrization 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 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 et al. (2004), where forthcoming values are interpreted as missing-value findable by smoothing procedures. A development of this model is presented in Camarda (2019), overcoming robustness forecasting problems. An innovative proposal has been introduced in Mitchell et al. (2013), wherein the LC time-index is 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 the machine and deep learning models, have recently appeared in the literature. The first insight based on machine learning tools is offered in Deprez et al. (2017), where regression trees algorithms are adopted to improve the estimation of death rates from canonical models, such as the LC and the Renshaw-Haberman. These findings are extended in Levantesi and Pizzorusso (2019) and Levantesi and Nigri (2020) for predictive purposes. A Neural Network design for mortality analysis

is initially scrutinized by Hainaut (2018), profitably aiming to extrapolate suitable non-linearities in the observed force of mortality. A NN vision within the LC framework is presented in Perla et al. (2021) and in Richman and Wüthrich (2019a). The latter proposes a NN representation for the multi-population LC model, overcoming parameters optimization problems and achieving reliable forecasting performances. Following this wake, Perla et al. (2021) takes the moves showing the remarkable accuracy achieved in a large-scale prediction of mortality. In particular, different NN structures are tested, such as the LSTM and the convolutional NN, engaging each of them to produce point forecasts of mortality rates simultaneously for many countries.

Deep learning models, especially 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 et al. (2020), our approach pursues a model integrating deep learning techniques, 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 a deep NN model.

It is worth to recall 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, at the same time, 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, since learning models such as NNs furnish only point predictions. To this purposes, Khosravi et al. (2011) provided an extensive methodological review of the main approaches for calculating confidence and prediction intervals, concluding that no method beats the other ones in each considered comparison metric. Anyhow, procedures based on structural assumptions, such as the Delta method (Wild and Seber, 1989), the Mean-Variance Estimation (Nix and Weigend, 1994) and the Bayesian approach (MacKay, 1992), are relevant solutions but suffering computational troubles 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, favouring the estimation of a theoretical probability distribution through an empirical one, see for instance Tibshirani (1996), Heskes (1997), Khosravi et al. (2011), Mazloumi et al. (2011), Kasiviswanathan and Sudheer (2014), Khosravi et al. (2015) and Li et al. (2018). In particular, bootstrap procedures seem to represent the more tempting alternative since 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 et al. (2005), Koissi et al. (2006), Li et al. (2009), D'Amato et al. (2011, 2012a,b). However, to the best of our knowledge, machine and deep learning literature in mortality forecasting lacks studies about uncertainty estimation.

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 a RNN with LSTM architecture to forecast the LC time-index. The resulting integrated model, namely LC-LSTM, and mortality boundaries it provides, fills the gap between the deep learning integrated mortality models and the uncertainty estimation, getting suitable ranges of variability. This allows at reaching 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 are 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 predictions robustness w.r.t. the historical mortality trend. The latter, like performance metrics, are used to assess the resulting mortality forecasts with a backtesting approach. Our findings confirm the LC-LSTM 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.

5.1 The LC-LSTM model

Let us consider the LC Poisson model proposed in Brouhns et al. (2002) as the reference model describing the behaviour of the age-period mortality rates. For the sake of clarity, we recall that, for ages $x \in \mathcal{X} = \{0, 1, \dots, \omega\}$ and calendar years $t \in \mathcal{T} = \{t_0, t_1, \dots, t_n\}$, the observed number of deaths, $D_{x,t}$, follows a Poisson distribution:

$$D_{x,t} \sim Poi(E^c_{x,t} m_{x,t}), \tag{5.1}$$

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 equation defining the LC model structure associated to the assumption (2.13) (Currie (2017)) is:

$$\ln m_{x,t} = \alpha_x + \beta_x k_t, \tag{5.2}$$

where α_x and β_x are age-dependent parameters illustrating the mortality age pattern and k_t is a time-index parameter representing the mortality behaviour over time. As is well-known, parameters constraints must be satisfied to ensure model identification, i.e. $\sum_{t=t_0}^{t_n} k_t = 0$ and $\sum_{x=0}^{\omega} b_x = 1$.

Let $\boldsymbol{\kappa}_{\mathcal{T}} = (k_{t-j})_{t \in \mathcal{T}}$ be the vector of the time lagged k_t, being $j \in \mathbb{N}$ the time lag. According to Eq.(3.19), we model the LC time-index as below:

$$k_t = f_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}; \boldsymbol{\mathcal{W}}\right) + \gamma_t = \psi \circ \left(H_t^{(P)} \circ H_t^{(P-1)} \circ \cdots \circ H_t^{(1)}\right)\left(\boldsymbol{\kappa}_{\mathcal{T}}; \boldsymbol{\mathcal{W}}\right) + \gamma_t. \tag{5.3}$$

Integrating Eq.(5.3) within the LC structure in Eq.(5.2), the LSTM will act as a predictor over the forecasting horizon $\mathcal{T}' = \{t_n + 1, t_n + 2, \dots, t_n + s\}$, and the LC-LSTM model expression is:

$$\ln m_{x,t} = \hat{\alpha}_x + \hat{\beta}_x \left(f_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}'}; \boldsymbol{\mathcal{W}}\right) + \gamma_t\right), \quad \forall t \in \mathcal{T}'. \tag{5.4}$$

with $\hat{\alpha}_x$ and $\hat{\beta}_x$ the estimates of age-dependent parameters.

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

$$k_t = \varphi\left(\boldsymbol{\kappa}_{\mathcal{T}}\right) + \gamma_t, \quad \forall t \in \mathcal{T}, \tag{5.5}$$

where the unknown function $\varphi : \mathbb{R}^j \to \mathbb{R}$ maps the vector $\boldsymbol{\kappa}_{\mathcal{T}}$ to k_t over the time horizon $\mathcal{T}$, unless the noise component. Referring to the RNNs universal functional approximation property (Schäfer and Zimmermann (2007)), the proposed model integration allows to resemble the unknown map $\varphi\left(\boldsymbol{\kappa}_{\mathcal{T}}\right)$ through a RNN with LSTM architecture, whose functional form is shaped according to the available time-index history. As the RNN model approximates the map $\varphi\left(\boldsymbol{\kappa}_{\mathcal{T}}\right)$, it also defines the mean of response variable conditioned to the explicative ones (Bishop (1995)), that is:

$$\hat{k}_t = \hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}; \hat{\boldsymbol{\mathcal{W}}}\right) = \mathbb{E}\left(k_t | \boldsymbol{\kappa}_{\mathcal{T}}\right), \tag{5.6}$$

where $\hat{f}_{LSTM}$ is the fitted function composition and $\hat{\boldsymbol{\mathcal{W}}}$ is the NN parameters estimate. Such a relation highlights that the LSTM model captures the LC time-index conditional expectation. Therefore, the LC-LSTM model provides the following point predictions:

$$\ln \hat{m}_{x,t} = \mathbb{E}\left(\ln m_{x,t}\right) = \hat{\alpha}_x + \hat{\beta}_x \hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}'}; \hat{\boldsymbol{\mathcal{W}}}\right), \quad \forall t \in \mathcal{T}'. \tag{5.7}$$

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

5.2 Prediction intervals for the LC-LSTM model

Prediction intervals (henceforth PI) outline a probabilistic range suitable to incorporate various forecasting scenarios, then probing uncertainty on the future mortality realizations. Stochastic mortality models forecast PIs, whose estimates act as uncertainty measure linked to the expected future mortality, see for instance Booth and Tickle (2008). Thus, in a proper forecasting process PIs are meaningful in supporting both risk evaluations and the model estimates reliability.

Referring to NNs, PIs construction is a challenging task because of different uncertainty sources impact on the learning process, then conditioning the NN generalization performances. By a broad perspective, NNs models are exposed to a learning uncertainty, depending both on the data and the NN functioning. Since the data employed in the learning process are a realization of an underlying stochastic process, a training data uncertainty looms. Indeed, varying input could involve in distinct function compositions, generating a distribution for the output values. In addition, a variability could arise due to the optimization procedures necessary to learn NN parameters value from data. As the loss 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 a model uncertainty could occur for a possible structural model misspecification.

Addressing the measurement of uncertainty sources separately is a complex problem, as 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 model integration.

Recalling that age-dependent parameters are time invariant, the uncertainty in death rates concerns the temporal dynamic described by Eq.(5.3). Thus, we focus on the construction of time-index PI, exploiting the k_t total variance, $\sigma^2_{k_t}$. To this end, the PI characterization is based on the following result.

Proposition. *Let $(k_t)_{t\in\mathcal{T}'}$ the time-index series over the forecast horizon $\mathcal{T}'$. The total variance associated to each time-index value is:*

$$\sigma^2_{k_t} = \sigma^2_{\hat{k}_t} + \sigma^2_\gamma + \mathbb{E}\left[BIAS\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}'}\right)^2\right] \tag{5.8}$$

where $BIAS\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}'}\right) = \mathbb{E}\left(\varphi\left(\boldsymbol{\kappa}_{\mathcal{T}'}\right) - \hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}'}\right)$ and $\sigma^2_{\hat{k}_t}$ is the NN output variance.

Proof. Recalling Eq.(5.6), over the forecasting horizon is straightforward noting that

$$\mathbb{E}\left(k_t\right) = \mathbb{E}\left[\mathbb{E}\left(k_t|\boldsymbol{\kappa}_{\mathcal{T}'}\right)\right] = \mathbb{E}\left(\hat{k}_t\right).$$

We proceed to define the time-index variance by direct calculation:

$$\begin{aligned}\sigma^2_{k_t} &= \mathbb{E}\left[\left(k_t - \mathbb{E}(k_t)\right)^2\right] = \mathbb{E}\left[\left(k_t - \mathbb{E}(k_t) + \hat{k}_t - \hat{k}_t\right)^2\right] = \\ &= \mathbb{E}\left[\left(k_t - \hat{k}_t\right)^2\right] + \mathbb{E}\left[\left(\hat{k}_t - \mathbb{E}\left(\hat{k}_t\right)\right)^2\right] + 2\mathbb{E}\left[\left(k_t - \hat{k}_t\right)\left(\hat{k}_t - \mathbb{E}(k_t)\right)\right]\end{aligned} \tag{5.9}$$

Assuming stochastic independence between $\left(k_t - \hat{k}_t\right)$ and $\left(\hat{k}_t - \mathbb{E}(k_t)\right)$, follows that:

$$\sigma^2_{k_t} = \mathbb{E}\left[\left(k_t - \hat{k}_t\right)^2\right] + \sigma^2_{\hat{k}_t} \tag{5.10}$$

The term $\mathbb{E}\left[\left(k_t - \hat{k}_t\right)^2\right]$ identifies the mean squared error of prediction associated to $\hat{k}_t$, whose expression can be developed as below:

$$\begin{aligned}\mathbb{E}\left[\left(k_t - \hat{k}_t\right)^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(k_t - \hat{k}_t\right)^2\middle|\boldsymbol{\kappa}_{\mathcal{T}'}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\left(\varphi\left(\boldsymbol{\kappa}_{\mathcal{T}'}\right) + \gamma_t - \hat{k}_t\right)^2\middle|\boldsymbol{\kappa}_{\mathcal{T}'}\right]\right] = \\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\varphi\left(\boldsymbol{\kappa}_{\mathcal{T}'}\right) - \hat{k}_t\right)^2\middle|\boldsymbol{\kappa}_{\mathcal{T}'}\right]\right] + \mathbb{E}\left[\mathbb{E}\left[\gamma_t^2\middle|\boldsymbol{\kappa}_{\mathcal{T}'}\right]\right] = \\ &= \mathbb{E}\left[BIAS\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}'}\right)^2\right] + \sigma^2_\gamma.\end{aligned} \tag{5.11}$$

Substituting Eq.(5.11) in Eq.(5.10), we have:

$$\sigma^2_{k_t} = \sigma^2_{\hat{k}_t} + \sigma^2_{\gamma} + \mathbb{E}\left[BIAS\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}'}\right)^2\right], \tag{5.12}$$

completing the proof. □

Following Eq.(5.8), uncertainty in the future mortality behaviour is linked to the NN model. The NN ability to approximate data depends on the function composition extent, which is intrinsically related to the learning process. Hence, $\sigma^2_{\hat{k}_t}$ includes fluctuations due to training data and learned weights, as well as from model misspecification occurrences. In compliance to the bias-variance principle, an expected bias component is present. In fact, both bias and variance contribute to the NN prediction error and the NN model suitability is based on the reduction of both. Finally, the variance σ^2_{γ} constitutes an irreducible term of uncertainty, since it refers to the random noise component.

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

To derive the NN output variance, the conditioned time-index distribution, $\mathbb{P}\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}'}\right)$, 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 time-index variance refers to the NN ensemble paradigm, based on the jointly use of multiple NNs (Zhou et al. (2002)). Utilizing a bootstrap technique, multiple training data samples are generated in order to develop an empirical distribution, $\hat{\mathbb{P}}\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}'}\right)$, constitutes by different NN point predictions. The final estimates are then obtained aggregating, by average, the various outputs. The latter procedure, namely bootstrap aggregating or bagging (Breiman (1996)), produces less unbiased estimation, favouring an adequate variance measurement. This means that the expected bias in Eq.(5.8) is seen as a negligible component affecting the time-index variance (Khosravi et al. (2015)).

The bagging scheme proposed in the present work is described in the following steps:

Step 1. Using the available time-index series $\boldsymbol{\kappa}_{\mathcal{T}}$, we train the LSTM model to obtain the point estimates in Eq.(5.6) over the forecast horizon $\mathcal{T}'$;

Step 2. We generate $B \in \mathbb{N}$ samples of $\boldsymbol{\kappa}_{\mathcal{T}}$ through a proper bootstrap procedure. In particular, we refer to the bootstrap strategy proposed in Koissi et al. (2006);

Step 3. For each b^{th} sample, with $b = 1, \ldots, B$, we re-optimize the weights of the function composition defined in *Step 1.* In doing so, only the NN weights will change given the new data and the created NNs ensemble will include uncertainty for both training data and parameters;

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

$$\hat{\mathbb{P}}\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}'}\right) = \left(\hat{k}_t^{(b)} = \hat{f}_{LSTM}\left(\boldsymbol{\kappa}_{\mathcal{T}}^{(b)}, \hat{\boldsymbol{\mathcal{W}}}^{(b)}\right), b = 1, \ldots, B\right); \tag{5.13}$$

Step 5. From the bootstrap distribution $\hat{\mathbb{P}}\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}'}\right)$, we find the estimates of interest by aggregation. Hence, the bagged estimate of the variance $\sigma^2_{\hat{k}_t}$ is:

$$\hat{\sigma}^2_{\hat{k}_t} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{f}_{LSTM}\left(\boldsymbol{\kappa}^{(b)}_{\mathcal{T}}, \hat{\boldsymbol{\mathcal{W}}}^{(b)}\right) - \overline{k}_t\right), \tag{5.14}$$

where $\overline{k}_t = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_{LSTM}\left(\boldsymbol{\kappa}^{(b)}_{\mathcal{T}}, \hat{\boldsymbol{\mathcal{W}}}^{(b)}\right)$ is the bagged estimate for the conditional expectation $\mathbb{E}\left(\hat{k}_t|\boldsymbol{\kappa}_{\mathcal{T}'}\right)$.

We emphasize that 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 trade-off states. Howbeit, if the employed bootstrap technique fits the density estimation problem and the trained NN model is robust, then the estimated variance does not induce an explosive prediction intervals behaviour over time.

5.2.2 Estimating σ^2_γ

Looking at Eq.(5.3), mortality dynamic incorporates an intrinsic randomness not explained by the network. A NN appropriately trained catches the key input-output data schemes, skimming noisy examples. Consequently, the NN model is suitable to produce forecast avoiding overfitting occurrences. For our purposes, such noise is analysed and predicted. Considering the training set interval $\mathcal{T}$, we deal with the series $\left(k_t - \hat{k}_t\right)_{t\in\mathcal{T}}$ as a proxy of the unwrapped noise by NN. It helps to evaluate the estimates $\hat{\sigma}^2_\gamma$ as the time-index residual uncertainty over $\mathcal{T}$, spreading the random error over the forecast horizon $\mathcal{T}'$ through a random walk representation.

5.3 Performance metrics of forecasting

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

$$RMSE_{(k)} = \sqrt{\frac{\sum_{t=t_n+1}^{t_n+s}\left(k_t - \hat{k}_t\right)^2}{s-1}}, \quad RMSE_{(m)} = \sqrt{\frac{\sum_{t=t_n+1}^{t_n+s}\left(\ln m_{x,t} - \ln \hat{m}_{x,t}\right)^2}{s-1}}. \tag{5.15}$$

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 rates values fall within the PI provided by the mortality model. Let $\hat{k}^L_t$ be the estimated time-index lower

bound and be $\hat{k}_t^U$ the estimated time-index upper bound. Then, the PICP for the k_t series is defined as follows:

$$PICP_{(k)} = \frac{1}{s-1} \sum_{t=t_n+1}^{t_n+s} \mathbf{1}_{\{\hat{k}_t \in [\hat{k}_t^L, \hat{k}_t^U]\}}, \tag{5.16}$$

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 forecasting horizon, that is:

$$MPIW_{(k)} = \frac{1}{s-1} \sum_{t=t_n+1}^{t_n+s} \hat{k}_t^U - \hat{k}_t^L. \tag{5.17}$$

We also calculate PICP and MPIW on the log-mortality rates by a given age x. Let $\ln \hat{m}_{x,t}^L$ be the estimated mortality rates lower bound and be $\ln \hat{m}_{x,t}^U$ the estimated mortality rates upper bound. Then, we specify the PICP and MPIW as follows:

$$PICP_{(m)} = \frac{1}{s-1} \sum_{t=t_n+1}^{t_n+s} \mathbf{1}_{\{\ln \hat{m}_{x,t} \in [\ln \hat{m}_{x,t}^L, \ln \hat{m}_{x,t}^U]\}}, \tag{5.18}$$

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

$$MPIW_{(m)} = \frac{1}{s-1} \sum_{t=t_n+1}^{t_n+s} \ln \hat{m}_{x,t}^U - \ln \hat{m}_{x,t}^L. \tag{5.19}$$

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. An explosive demeanor in variability is reflected by greater MPIW levels, jeopardizing the biological plausibility of mortality forecasts. The latter qualitative criterion is valuable since it concerns the predicted uncertainty levels consistency w.r.t. the historical volatility at different ages (Cairns et al. (2011)).

5.4 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 getting from the LC Poisson model (Brouhns et al. (2002)) as a term of comparison. The analysis has been achieved using the R software (version 3.6.3), exploting the packages *StMoMo* (version 0.4.1), *forecast* (version 8.13), *Keras* (version 2.2.5) and *Tensorflow* (version 1.13.1).

5.4.1 Data

Our numerical experiment concerns three countries worldwide, Australia, Japan and Spain, analyzed by gender. Data were downloaded from the Human Mortality Database (HMD, www.mortality.org) and refer to the age range $\mathcal{X} = \{0, 1, \ldots, 99\}$. 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, i.e. the network robustness to changes in the training set length.

5.4.2 Neural Network tuning, training and ensembling

To apply the LSTM model, firstly we fit the LC model in Eq.(5.2) to the observed age-period mortality data, estimating the age-dependent parameters and the time-index series $(k_t)_{t\in\mathcal{T}}$. We pose $j = 1$ to define the one-step lagged time series, i.e. $\boldsymbol{\kappa}_{\mathcal{T}} = (k_{t-1})_{t\in\mathcal{T}}$, imposing to the LSTM model to sift mortality data at annual paces, that is $k_t = f_{LSTM}\left(k_{t-1}; \boldsymbol{\mathcal{W}}\right) + \gamma_t$ according to Eq.(5.3).

To tune and train the NN model is necessary to split the time-index series into distinct datasets. To this end, we exploit a hierarchical procedure. Setting $T = 2000$ as forecasting year for all the countries investigated, we define the training set and the testing set as below:

$$\begin{aligned} \text{TRAINING SET:} \quad & \mathcal{TR} = (k_t|k_{t-1})_{t=t_0,\dots,T} \\ \text{TESTING SET:} \quad & \mathcal{TS} = (k_t|k_{t-1})_{t=T+1,\dots,t_n}\,, \end{aligned} \tag{5.20}$$

where $t_0 = \{1950, 1960\}$ and $t_n = 2018$. In addition, to validate the model we divide the training set into a sub-training set and in a validation set, considering the splitting rule $80\% - 20\%$. Hence, denoting with T^{sub} the last year in the sub-training set, we have:

$$\begin{aligned} \text{SUB-TRAINING SET:} \quad & \mathcal{TR}^{\text{sub}} = (k_t|k_{t-1})_{t=t_0,\dots,T^{\text{sub}}} \\ \text{VALIDATION SET:} \quad & \mathcal{VS} = (k_t|k_{t-1})_{t=T^{\text{sub}}+1,\dots,T} \end{aligned} \tag{5.21}$$

We use the sets $\mathcal{TR}^{\text{sub}}$ and $\mathcal{VS}$ to tune the NN structure through a grid search technique. Thus, a bounded discrete parametric space is a priori settled, whose possible values are arbitrarily chosen acting as network hyper-parameters. Fixing a hyper-parameters combination, the learning process begins minimizing the Mean Squared Error loss function over the set $\mathcal{TR}^{\text{sub}}$. We select as optimal NN structure the one identified by the hyper-parameters 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. For each countries and both genders, the LSTM model is characterized by $p = 1$ hidden layer, considering the ReLu function Glorot et al. (2011) as feed-forward activation function, the tangent hyperbolic function as 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 countries and genders. Finally, the best NN architecture is afterwards employed on the training set, $\mathcal{TR}$, to spawn 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 backtesting exercise.

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 1960s generate predictions sensitive to small variations in the data. Therefore, we need at least 40 observations to adequately tune the network model.

The tuned LSTM model acts as the reference model in *Step 1.* of the proposed bagging scheme in Section 5.2.1. Following the bootstrap strategy proposed in Koissi et al. (2006), we generate $B = 1000$ bootstrap samples of the training set $\mathcal{TR}$. Maintaining the tuned network function composition, $\hat{f}_{LSTM}$, we estimates its weights on the b^{th} training set producing the related forecasts over testing set horizon. Therefore, the bootstrap distribution $\hat{\mathbb{P}}\left(\hat{k}_t | k_{t-1}\right)$ is obtained, allowing for the bagged variance calculation as in Eq.(5.14).

5.4.3 Results

In the following we provide the results of our numerical application, recalling the performance metrics presented in Section 5.3. We firstly refer to RMSE to evaluate the point forecasts accuracy, considering also the error of the LC projections as benchmark. To appreciate the PIs quality by PICP and MPIW indicators, after the bagging scheme we need to assess the noise variance in order to estimate PI boundaries. We consider the sample variance of the series $\left(k_t - \hat{k}_t\right)_{t \in \mathcal{TR}}$ as the noise variance estimate over training set. To project the noise and its uncertainty over testing set 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 normality tests, such as the Shapiro-Wilk, the D'Agostino-Pearson and the Jarque-Bera. For all the investigated countries and both genders, the noise analysis confirms the ability of a random walk representation with Gaussian innovations for the noise component (see Appendix B). Therefore, the LC-LSTM time-index values are embedded within the following PI, for a confidence level α:

$$\left[\hat{k}_t^L, \hat{k}_t^U\right] = \left[\hat{k}_t - z_{\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2_{\hat{k}_t} + \hat{\sigma}^2_{\gamma}}\ ,\ \hat{k}_t + z_{\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2_{\hat{k}_t} + \hat{\sigma}^2_{\gamma}}\right] \tag{5.22}$$

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 appear in Table 5.1, comparing the LSTM performances in the LC-LSTM, with the ARIMA ones in the LC model.

Table 5.1. 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**

For all the 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 characterizes the interval 1960-2000. Such a behavior has been driven by a decline in infant mortality, as well as reductions in mortality at older ages after WWII (see for instance Rau et al. (2008)).

As a general statement about prediction accuracy, our analysis confirms the ARIMA ability to represent linear evolution in mortality. On the other side, the LSTM seems to be advisable for linear, noisy, or non-linear series. Scrutinizing the uncertainty results, the LSTM offers always a 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 the ARIMA one.

The majority of cases promote the LSTM model's usefulness in affording a more actual mortality trend, as well as for uncertainty estimation. The most virtuous example concerns the Australian males, presenting the lower RMSE on 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, i.e. 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 performances improve 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 side, 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, on average, to anticipate half of the future realizations. An analogous result holds for the 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%.

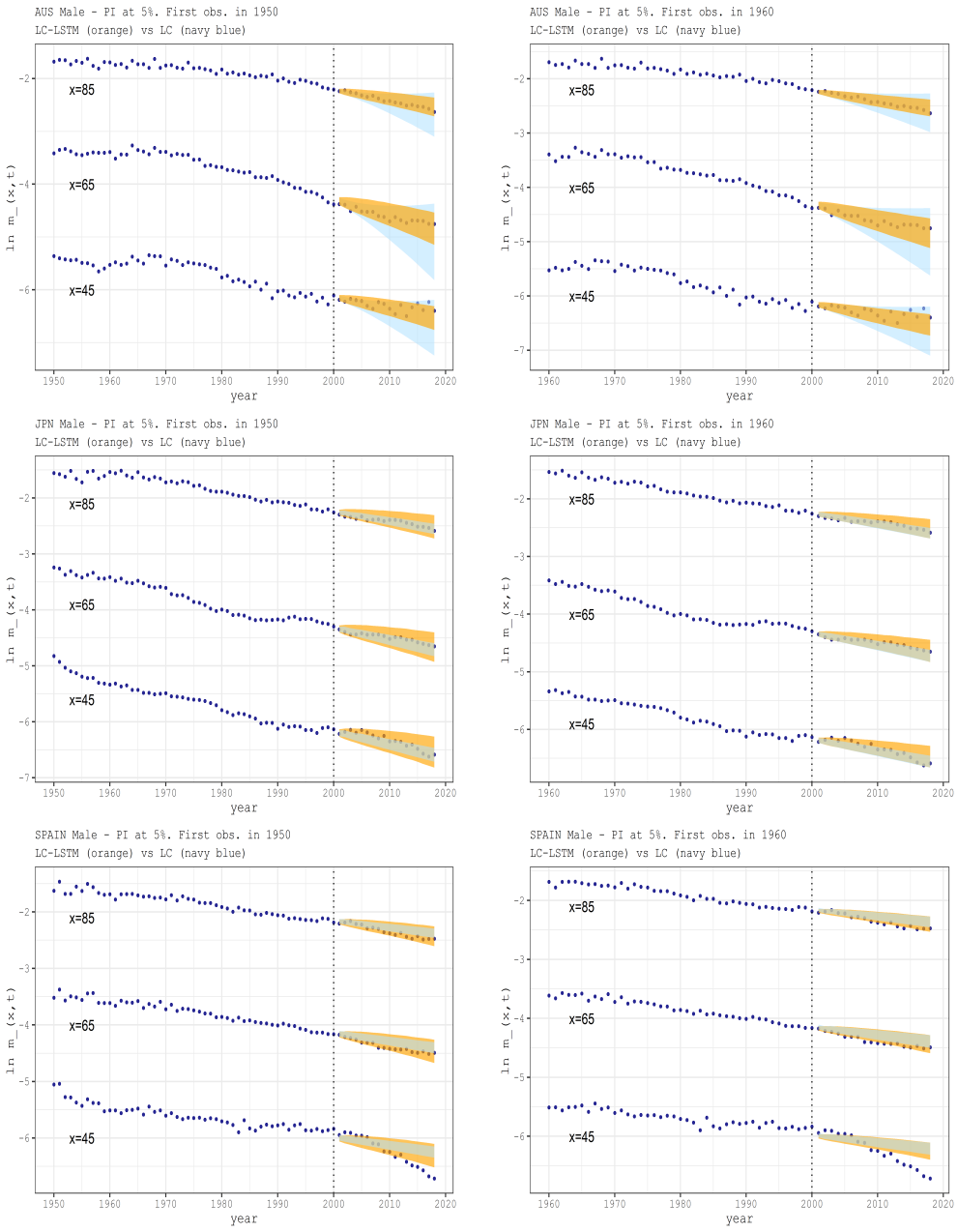
We also depict the mortality profile for both genders considering ages 45, 65 and 85. To explore these results, we display the performance metrics in Table 5.2, as well as the PIs graphs in Figure 5.1 and Figure 5.2.

We can highlight the estimated PIs for the LC-LSTM model both in terms of point and interval estimates. Looking at the Japanese population, we endorse the findings in Table 5.1 for ages 45 and 65. The LC-LSTM provides boundaries properly shaped according to death rates, while the LC model presents the narrowest ranges of variability lacking uncertainty information. For example, over the training period 1960-2000 for the Japanese females aged 65, the PIs for the LC model show a coverage probability around 33%, while the LC-LSTM provides $PICP_{(m)} = 1$ with a similar interval width. For age 85, where mortality reductions present slower linear changes over time, also the LC fits the future mortality profile.

For the Spanish population, the LC-LCTM seems to be the befitting model for predictive purposes. As reported in Table 5.1, 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, albeit PI be ever wider than the LC model one. We stress how both the LC and the LC-LSTM model fail to catch the non-linear mortality 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. Since we pose $T = 2000$ as the forecasting year, the extrapolation approach underlying both

the LC and the LC-LSTM induces misleading projections.

Figure 5.1. MALE PI ($\alpha = 5\%$). Forecasting period: 2001-2018. Training period: 1950-2000 (left), 1960-2000 (right).



Finally, we appreciate the LC model performances in uncertainty estimation for the Australian males. We highlight the LC model greatest probability coverage and interval width. Nevertheless, the latter hints at some questions about the LC model prediction suitability in the long-run. See, for instance, Figure 5.3 displaying a 50-year prediction for the Australian males aged 65, for both training periods.

Figure 5.2. FEMALE PI ($\alpha = 5\%$). Forecasting period: 2001-2018. Training period: 1950-2000 (left), 1960-2000 (right).

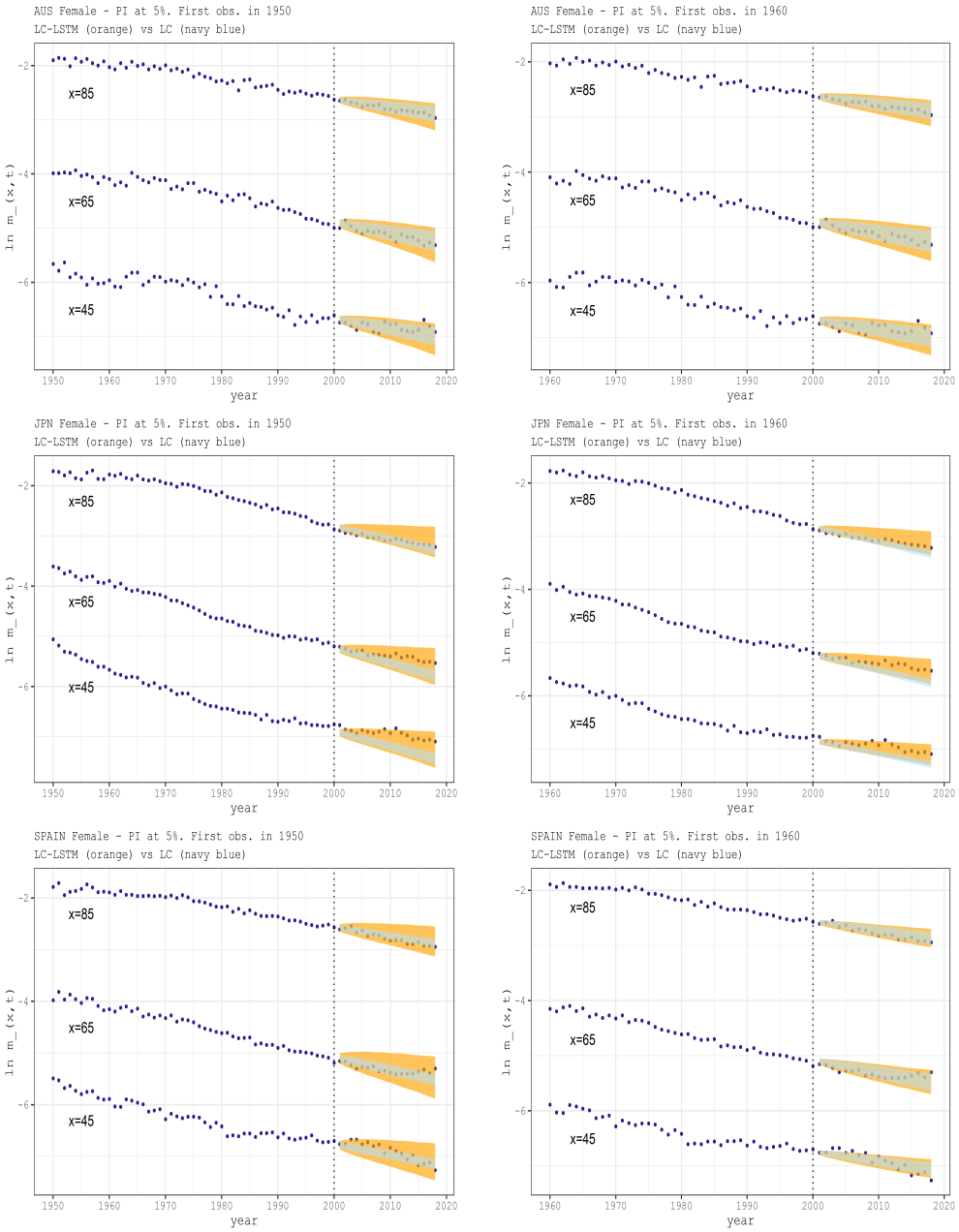


Table 5.2. $\ln m_{x,t}$ performance metrics values for each training period. Forecasting years: 2001-2018.

$x = 45$

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**

$x = 65$

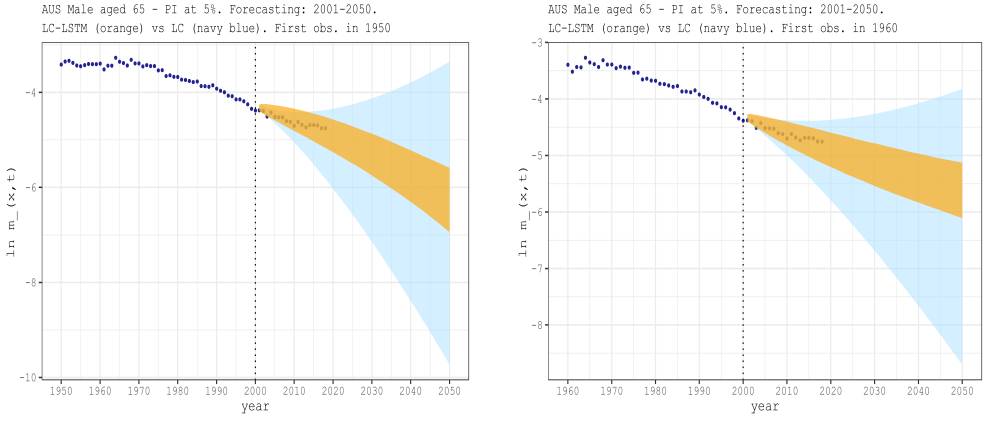
Country	Model	Training period 1950-2000						Training period 1960-2000					
		Male			Female			Male			Female		
		$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$	$RMSE_{(m)}$	$PICP_{(m)}$	$MPIW_{(m)}$
Australia	LC	0.157	1	**0.672**	0.061	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**	**0.66**	**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**

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 around 96%, in the case of the PI lower bound, or a possible increase of 68%, considering 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.

Figure 5.3. Australian Males. PI ($\alpha = 5\%$) for $x = 65$. Training period: 1950-2000 (left), 1960-2000 (right). Forecasting period: 2001-2050.



In particular, the LC-LSTM produces a 40-year decrease in mortality between 82%, considering the PI lower bound, and 46% according to the PI upper bound. Moreover, inspecting Figure 5.3 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 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. As opposite, the interval 1950-2000 points up a non-linear behavior due to the longevity accelerations in the period 1950-1960. In this case, the LSTM is able in extrapolating a coherent mortality range with the historical observation, allowing for biological plausibility but believing in 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 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 5.3, are indispensable to address forecasts in a backtesting 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 fair linear way over time and at different ages, properly reflecting the demographic trend observed since the 1960s.

Chapter 6

Conclusion

The present research work considers the RNN model with LSTM architetcure as model to improve mortality forecasting analysis. The studies executed confirm the suitability of our proposal to predict human lifetime measures, contributing to the mortality literature.

Looking at the investigations within Chapter 4, life expectancy and lifespan disparity indicators are coherently anticipated, allowing demographic reasonableness. Considering the nature of life expectancy, it is not a merely a time-trend index, but rather a "latent factor"incorporating different unobserved latent variables. It implicitly encompasses economic fluctuations, medical innovation and many other variables that directly (or indirectly) influenced the mortality trend. In this framework, our analysis proposes a new approach based on LSTM neural network to forecast longevity indexes both independently and simultaneously at birth and age 65, catching either short and long term factors on mortality improvements. As for the LSTM applied to life expectancy, we observe that without imposing model restrictions, we can obtain predictions coherent with historical trends and biological criteria. The univariate LSTM outperforms all the models analysed, especially for life expectancy at age 65, where e.g., the BPLE shows some weaknesses as the linear assumption.

The wide discussion in literature on the relationship between life expectancy and lifespan disparity suggests that projections of life expectancy and lifespan disparity may benefit from simultaneous forecasting. Accordingly, we introduce a bivariate LSTM, which represents a novelty in the demographic panorama, by simultaneously forecasting life expectancy and lifespan disparity in the RNN framework. Our simultaneous model obtains higher levels of accuracy compared to the first-order VAR model used as a benchmark for multivariate series forecasting. Our empirical analysis, based on five countries, two fitting periods and both genders, shows that the simultaneous forecasting of life expectancy and lifespan disparity is less adequate than independent modelling. Nevertheless, our results lead to speculating that only life expectancy at birth projections take advantage of simultaneous forecasting with life disparity. Extrapolative models, e.g., the Lee-Carter model, may also benefit from a parameter adjustment consistent not only with lifespan disparity as in Rabbi and Mazzuco (2020) but with both observed life expectancy and life disparity. We show that both independent and simultaneous forecasts of life expectancy and

lifespan disparity provide new insights for a comprehensive evaluation of the mortality forecasts, representing a useful tool to capture irregular mortality trajectories. Our findings support the decrease of lifespan disparity among developed countries, for which the evolution of age-at-death distribution assumes more compressed tails over time. Besides, our approach based on the long-short term enables to consider the entire time series, without excluding short-term shocks from the analysis. Using two different periods, we show that the LSTM provides robust forecasts to the unexpected mortality changes. This aspect sounds coherent with the *modus operandi* behind the LSTM architecture, where the neuron cell manages the time series noise, combining the long and short-term past information.

Looking at Chapter 5, the conceptualization of the deep learning integration allows to generalize the forecasting phase, achieving both accuracy in point predictions and reasonable prediction intervals. Such a model improvement relies on the LC age-period mortality representation, supporting the phenomenon interpretation. Indeed, among researchers and practitioners, the LC framework is widely employed as forecasting methodology, where the whole mortality surface is unfolded by two age-specific parameters and one time index. Its functional form is straightforward allowing a high degree of interpretability to mortality changes over time. Essentially, forecasting is greatly simplified, deriving from the projection of the single time-index. Furthermore, the LC is a probabilistic model, thereby allowing the derivation of prediction intervals of mortality rather than single deterministic point forecasts. These last points are crucial in a measure that LC gains the role of the benchmark model. Our proposal allows, at the same time, to represent the mortality surface through a canonical age-period model and to predict the future mortality realizations extrapolating the temporal mortality dynamic from data. The resulting LC-LSTM model poses 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-LSTM 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 annuities products or to appraise pension systems sustainability. Therefore, our proposal establishes a reliable improvement of the LC model in terms of predictive prowess, posing an innovative approach within mortality literature. The proposed framework might represent a prominent practice in the field of longevity forecasting, also for actuarial business tasks.

Finally, we stress that demographic and actuarial applications of NNs are quite recent. In fact, the first insight was from Hainaut in 2018 (Hainaut (2018)). Therefore, studies and extensions in the use of NNs in these fields are many, as are the types of NNs structures available nowadays. Referring to the demographic field, an intelligent use of NN models could materialize in mortality modelling by causes. In fact, through NNs is possible to capture the relationships existing between life expectancy, or death rates, with respect to the possible causes of death affecting different ages. This investigation would also be useful for risk management analyses implemented by pension funds and life insurers in general. Referring to the integration of extrapolative stochastic mortality models, a useful study could concern the extension of the model

integration concept to more structured mortality models than the LC, for example with a cohort effect. However, we believe that firstly is necessary to satisfy a primary need, that is to study the validation of deep learning models for business use by practitioners. This issue is by no means trivial, as several problems could emerge. For instance, a first problem concerns the observation of a fairly long historical period on the insured mortality, otherwise the insurer would to refer to the mortality results for the national population. In this sense, the NN could act not so much as a forecasting model, but as a backcasting model, expanding the insurer's mortality experience. Furthermore, the validation of an actuarial mortality model for risk analysis purposes requires the respect of various statistical properties, also empirically, first of all the robustness with respect to the data from which to extrapolate future mortality. Again, the uncertainty estimate must also be robust with respect to changes in the data, since the calculation of a solvency capital requirement derives from this estimate. Certainly, the road in the development of deep learning models to analyse the mortality/longevity risk in insurance business processes is uphill, but, we repeat, we are only at the beginning of an interesting line of research.

Appendix A

Graphical visualization of life expectancy and lifespan disparity forecasts

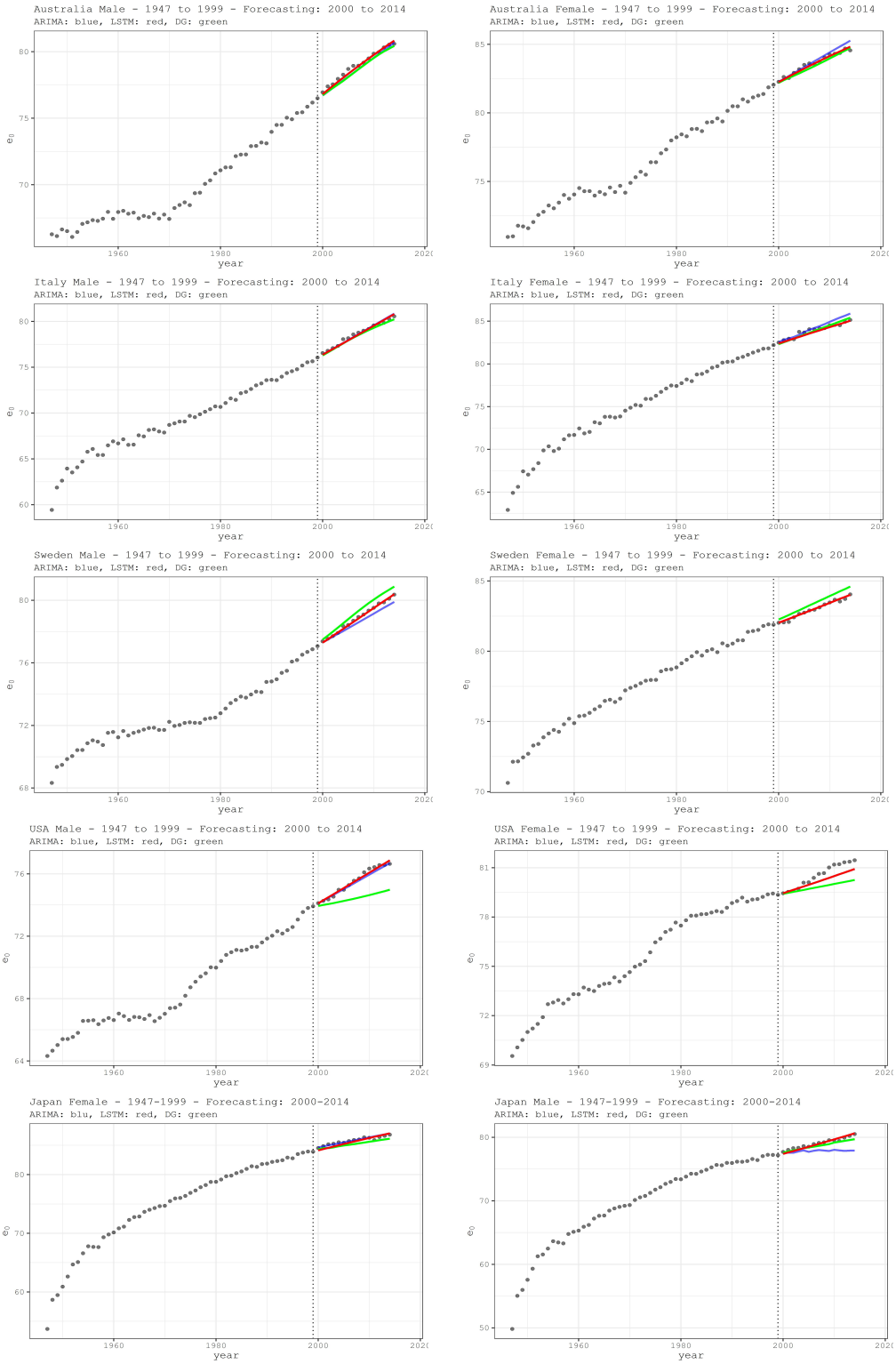


Figure A.1. Historical and forecasted values of $e_{0,t}$ by country and gender (females on the left, males on the right).

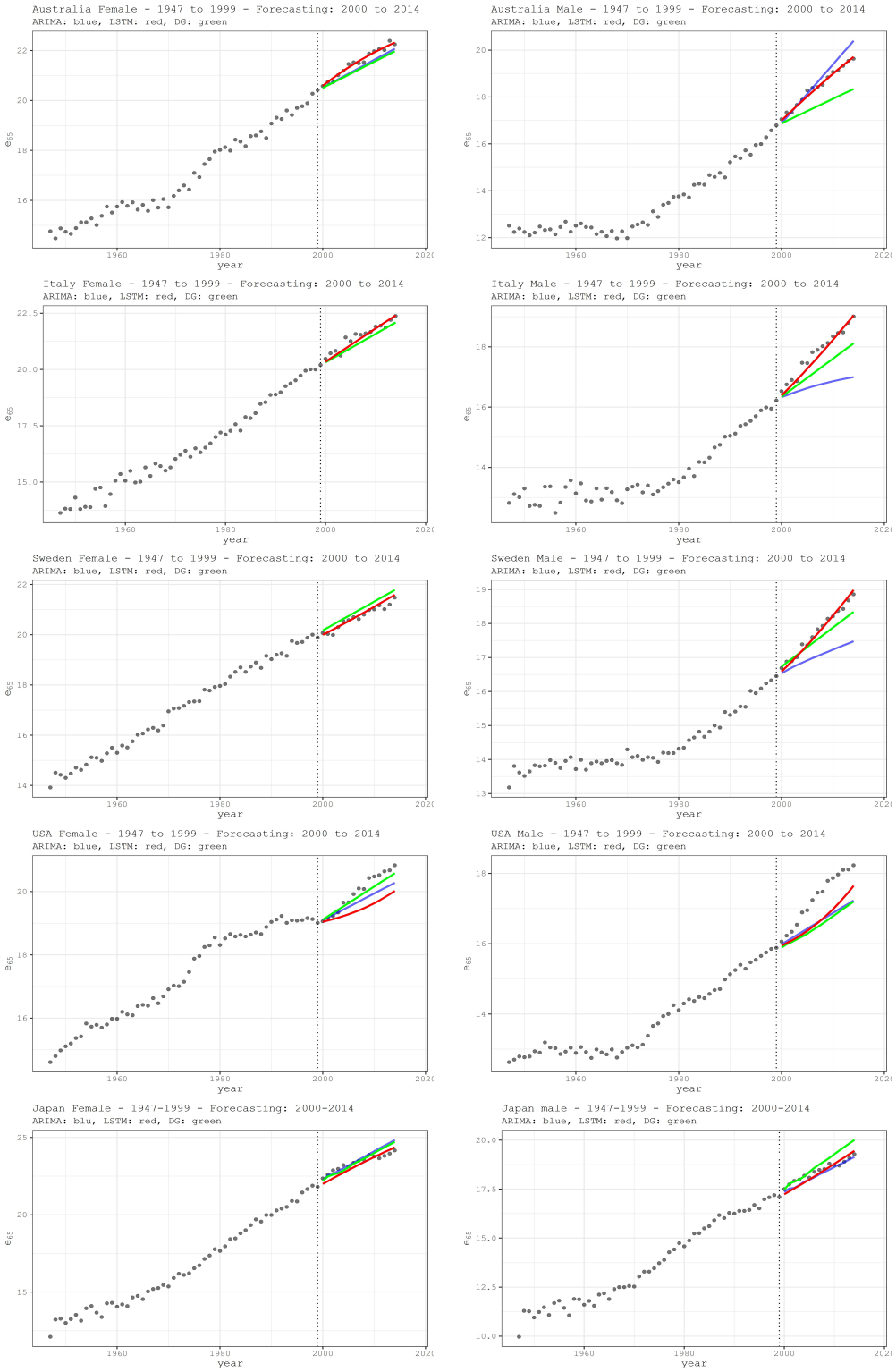


Figure A.2. Historical and forecasted values of $e_{65,t}$ by country and gender (females on the left, males on the right).

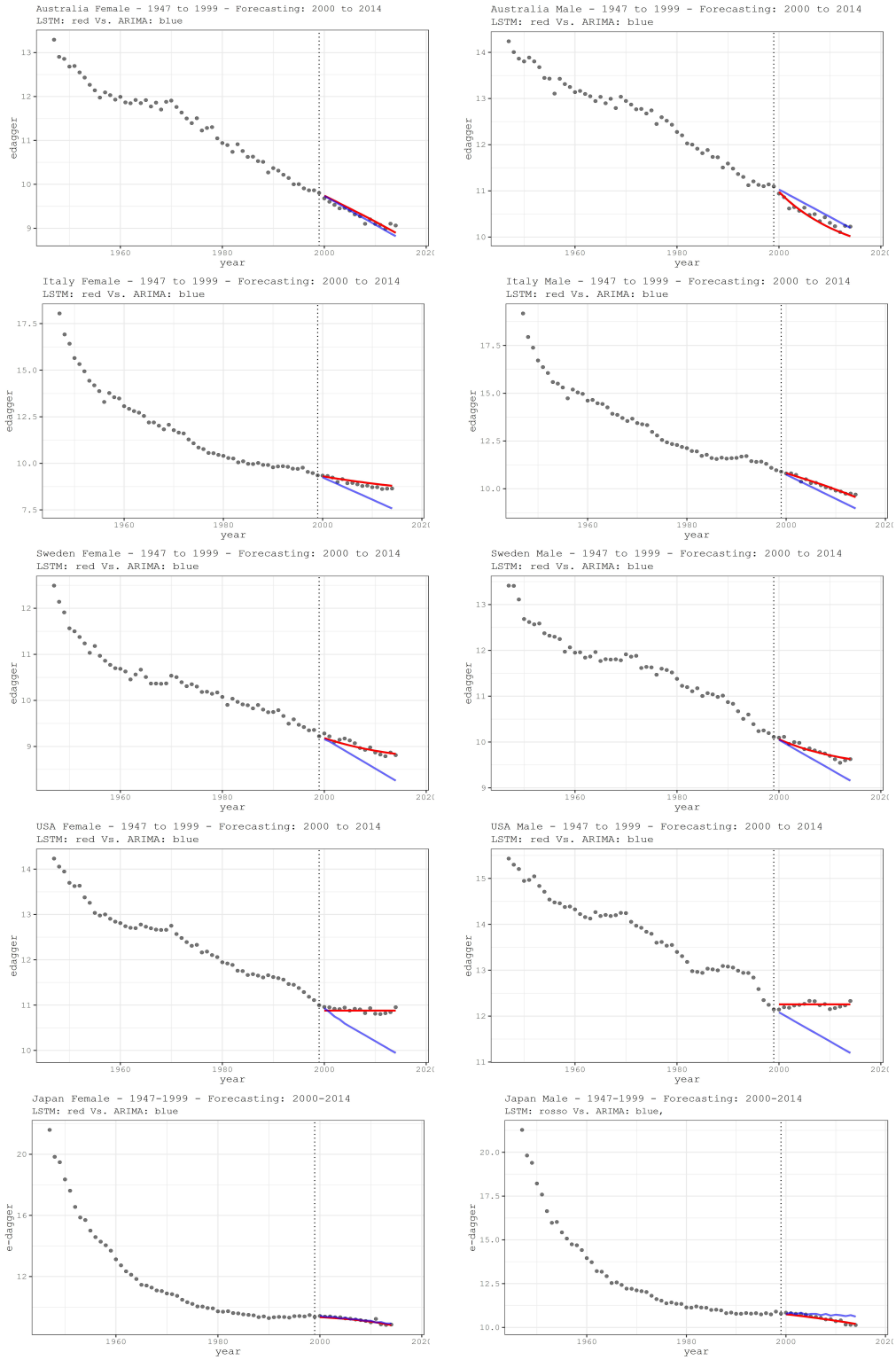


Figure A.3. Historical and forecasted values of $e^{\dagger}_{0,t}$ by country and gender (females on the left, males on the right).

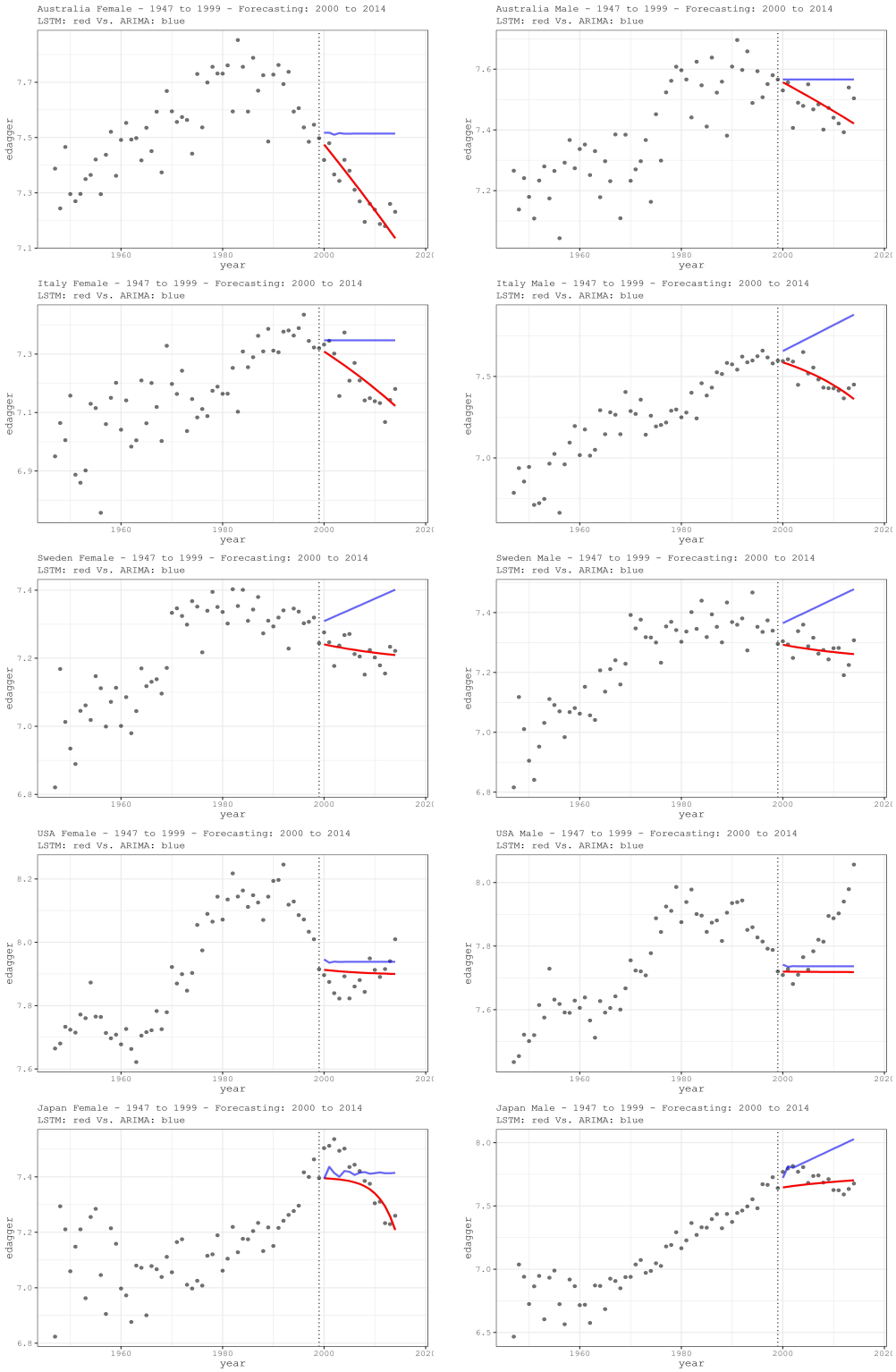


Figure A.4. Historical and forecasted values of $e^{\dagger}_{65,t}$ by country and gender (females on the left, males on the right).

Appendix B

Statistical tests to check the noise randomness and normality

Table B.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	0.96352	0.12489***	0.98379	0.82539***
	D'Agostino-Pearson	1.62692	0.44332***	0.85534	0.65203***
	Jarque-Bera	1.55177	0.46030***	0.64381	0.72477***
	ADF	-3.05447	0.15132***	-2.58739	0.34294***
Japan	Shapiro-Wilk	0.96193	0.10710***	0.97511	0.51356***
	D'Agostino-Pearson	8.05556	0.01781*	1.45996	0.48192***
	Jarque-Bera	7.35771	0.02525*	1.20406	0.54770***
	ADF	-3.49574	0.05128**	-2.73088	0.28662***
Spain	Shapiro-Wilk	0.97654	0.41696***	0.95790	0.14191***
	D'Agostino-Pearson	1.83229	0.40006***	2.82652	0.24335***
	Jarque-Bera	1.05350	0.59052***	2.31446	0.31436***
	ADF	-7.55942	0.01000	-4.11879	0.01516*

P-value significance level: $> 0.01^{*}$, $> 0.05^{**}$, $> 0.1^{***}$.

Table B.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	0.21209***	0.96724	0.29319***
	D'Agostino-Pearson	2.52531	0.28290***	0.78319	0.67598***
	Jarque-Bera	1.78204	0.41024***	0.60740	0.73808***
	ADF	-3.07190	0.14432***	-2.50033	0.37711***
Japan	Shapiro-Wilk	0.97452	0.34985***	0.98888	0.95815***
	D'Agostino-Pearson	3.12195	0.20993 ***	0.79814	0.67094***
	Jarque-Bera	2.09605	0.35063***	0.62112	0.73303***
	ADF	-5.14239	0.01000	-3.89596	0.02383*
Spain	Shapiro-Wilk	0.93640	0.02619*	0.97970	0.67844***
	D'Agostino-Pearson	8.69754	0.01292*	1.74855	0.41716***
	Jarque-Bera	7.56206	0.02280*	1.20753	0.54675***
	ADF	-5.80177	0.01000	-3.46488	0.06172***

P-value significance level: $> 0.01^{*}$, $> 0.05^{**}$, $> 0.1^{***}$.

Bibliography

Aburto, J. M., Van Raalte, A. (2018). Lifespan dispersion in times of life expectancy fluctuation: The case of Central and Eastern Europe. *Demography*, 55: 2071-2096.

Aburto, J. M., Villavicencio, F., Basellini, U., Kjærgaard, S., and Vaupel J. W. (2020). Dynamics of life expectancy and life span equality. *PNAS*, 117(10): 5250–5259

Aburto, J. M., Wensink, M., van Raalte, A., Lindahl-Jacobsen, R. (2018). Potential gains in life expectancy by reducing inequality of lifespans in Denmark: an international comparison and cause-of-death analysis. *BMC Public Health*, 18(1): 831.

Aggarwal, C. G. (2018). *Neural Networks and Deep Learning. A Textbook*. Springer Nature 2018. https://doi.org/10.1007/978-3-319-94463-0

Alho, J. M. (1990). Stochastic methods in population forecasting. *International Journal of Forecasting*, 6(4): 521–530.

Bergeron-Boucher, M.-P., Canudas-Romo, V., Oeppen, J., Vaupel, J.W., (2017). Coherent forecasts of mortality with compositional data analysis. *Demographic Research*, 37:527–566.

Bishop, M. C.(1995). *Neural Networks for Pattern Recognition*. Oxford University Press, New York. ISBN:978-0-19-853864-6

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24: 123-140.

Brouhns, N., M. Denuit, J. Vermunt, (2002), A Poisson log-bilinear approach to the construction of projected life tables. *Insurance: Mathematics and Economics*, 31: 373-393.

Bengio, Y., Simard, P., Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2): 157–166.

Bohk-Ewald, C., Ebeling, M. and Rau, R. (2017). Lifespan Disparity as an Additional Indicator for Evaluating Mortality Forecasts. *Demography*, 54: 1559. https://doi.org/10.1007/s13524-017-0584-0

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

Booth, H., R. J. Hyndman, L. Tickle, and P. De Jong. (2006). Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions. *Demographic Research*, 15: 289-310.

Booth, H., Maindonald, J., Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, 56(3):325–336.

Brockwell, P.J., Davis, R.A. (2016). *Introduction to time series and forecasting*. Springer Texts in Statistics. Springer International Publishing, Switzerland. ISBN 978-3-319-29852-8

Brouhns, N., Denuit, M. and Vermunt, J. K. (2002). A Poisson log-bilinear approach to the construction of projected life tables. *Insurance: Mathematics and Economics*, 31: 373-393.

Brouhns, N., Denuit, M., Van Keilegom, I. (2005). Bootstrapping the Poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal*, 3:212-224. DOI: 10.1080/03461230510009754

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

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

Cairns, A.J.G., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Khalaf-Allah, M. (2011). Mortality density forecasts: An analysis of six stochastic mortality models. *Insurance: Mathematics and Economics*, 48(3), 355-367. https://doi.org/10.1016/j.insmatheco.2010.12.005.

Camarda, C. G. (2019). Smooth constrained mortality forecasting. *Demographic Research*, 41: 1091-1130.

Currie I. D., Durban, M. and Eilers, P. H. C. (2004) Smoothing and forecasting mortality rates. *Statistical Modelling*, 4:279-298.

Currie, I. D. (2017). On fitting generalized linear and non-linear models with applications to multidimensional smoothing. *Scandinavian Actuarial Journal*, 4: 356-383.

Colchero, F., Rau, R., Jones, O.R. et al. (2016). The emergence of longevous populations. *Proceedings of the National Academy of Sciences (PNAS)*, 113(48): E7681-E7690.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 303–314.

D'Amato, V., Di Lorenzo, E., Haberman, S., Russolillo, M., 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., Russolillo, M (2012). The stratified sampling bootstrap for measuring the uncertainty in mortality forecasts. *Methodology Computing in Applied Probability*, 14:135–148. https://doi.org/10.1007/s11009-011-9225-z

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

Debonneuil, E., Loisel, S., Planchet, F. (2018). Do actuaries believe in longevity deceleration?. *Insurance: Mathematics and Economics*, 78:325-338. https://doi.org/10.1016/j.insmatheco.2017.09.008.

Deprez, P., Shevchenko, P.V., Wüthrich, M.V (2017). Machine learning techniques for mortality modeling. *European Actuarial Journal*, 7:337–352. https://doi.org/10.1007/s13385-017-0152-4.

Dowd, K., Blake, D., Cairns, A.J.G., Coughlan, G.D., Epstein, D., Khalaf-Allah, M. (2010). Backtesting sthocastic mortality models: an ex-post valuation of multi-period-ahead density forecasts. *North American Actuarial Journal*, 14:281-298.

Efron, B., Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall. ISBN: 0412042312

Edwards, R. D., Tuljapurkar, S. (2005). Inequality in life spans and a new perspective on mortality convergence across industrialized countries. *Population and Development Review*, 31(4): 645-674.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14(2): 179-211.

Funahashi, K. I. (1989). On the approximate realization of continuous mappings by neural networks, *Neural Networks*, 2:183–192.

Gers, F. A., Schmidhuber, J., Cummins, F. (1999). Learning to forget: Continual prediction with LSTM. *Artificial Neural Networks*, 470: 850-855.

Gers, F. A., Schmidhuber, J. (2000). Recurrent nets that time and count. *Proceeding of International Joint Conference on Neural Networks*, 24-27: 189-194.

Glorot, X., Bordes, A., Bengio, Y. (2011). Deep sparse rectifier neural networks. In Gordon, G., Dunson, D., and Dudk, M. (eds.), JMLR W&CP: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 15: 315-323.

Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Goldman, N., Lord, G. (1986). A new look at entropy and the life table. *Demography*, 23: 275-282.

Graves, A. et al. (2009). A novel connectionist system for unconstrained handwriting recognition. *Transactions on Pattern Analysis and Machine Intelligence*, 31(5): 855-868.

Haberman, S., Khalaf-Allah, M., Verrall, R. (2010). Entropy, longevity and the cost of annuities. *Insurance: Mathematics and Economics*, 48(2): 197-204.

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. *Advances in Neural Information Processing Systems (MIT Press)*, 9.

Hiam, L., Harrison, D., McKee, M., Dorling, D. (2018). Why is life expectancy in England and Wales 'stalling'?. *Journal of Epidemiology and Community Health*, 72(5): 404-408.

Hyndman, R.J., Ullah, S. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

Ho, J. Y., Hendi, A. S. (2018). Recent trends in life expectancy across high income countries: retrospective observational study. *British Medical Journal*, 362(k2562).

Hochreiter, S., Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9: 1735-1780.

Hornik, K., Stinchcombe, M., White, H. (1989). Multilayer feedforward networks are universal approximators, *Neural Networks*, 2:359–366.

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

Hunt, A., Blake, D. (2014). A general procedure for constructing mortality models. *North American Actuarial Journal*, 18(1):116-138. https://doi.org/10.1080/10920277.2013.852963

Hunt, A., Blake, D. (2015). On the structure and classification of mortality models. *Pension Institute Working Paper*. URL: `http://www.pensions-institute.org/workingpapers/wp1506.pdf`.

Kaakai, S., Hardy, H.L., Arnold, S., El Karoui, N. (2019). How can a cause-of-death reduction be compensated for by the population heterogeneity? A dynamic approach. *Insurance: Mathematics and Economics*, 89: 16-37.

Kannisto V. (2000). Measuring the Compression of Mortality. *Demographic Research*, 3(6).

Kasiviswanathan, K. S., Sudheer, K.P. (2013). Quantification of the predictive uncertainty of artificial neural network based river flow forecast models. *Stochastic Environmental Research and Risk Assessment* 27: 137–146. https://doi.org/10.1007/s00477-012-0600-2

Keskar, N.S., Mudigere, D., Nocedal, J., Smelyanskiy, M.,Tang, P.T.P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Khosravi, A., Nahavandi, S., Srinivasan, D., Khosravi, R. (2015). Constructing optimal prediction intervals by using neural networks and bootstrap method. *IEEE Transactions on Neural Networks and Learning Systems*, 26(8): 1810-1815. doi: 10.1109/TNNLS.2014.2354418.

Khosravi, A., Nahavandi, S., Creighton, D., Atiya, A. F. (2011). Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on Neural Networks*, 22(9): 1341-1356. doi: 10.1109/TNN.2011.2162110.

Koissi, M., Shapiro, A., Hognas, G. (2006), Evaluating and Extending the Lee–Carter Model for Mortality Forecasting Confidence Interval. *Insurance: Mathematics and Economics*.

LeCun, Y., Bottou,L., Orr, G.B., Muller, K.R. (2012). *Efficient backpropagation. in Neural Networks: Tricks of the Trade*. Springer-Verlag, Berlin.

Lee, R.D. (2006). Mortality Forecasts and linear life expectancy trends. Perspectives on mortality forecasting. *Social Insurance Studies, 3. The Linear Rise in Life Expectancy: History and Prospects*.

Lee, R. D., Carter., L. R. (1992). Modeling and forecasting US mortality. *Journal of the American statistical association*, 87 (419): 659-671.

Lee, R., Miller, T. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography*, 38(4):537–549.

Levantesi, S., Nigri, A. (2020). A random forest algorithm to improve the Lee–Carter mortality forecasting: impact on q-forward. *Soft Computing*, 24:8553–8567. https://doi.org/10.1007/s00500-019-04427-z

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

Li, J., Hardy, M., Tan, K. (2009). Uncertainty in mortality forecasting: An extension to the classical Lee-Carter approach. *ASTIN Bulletin*, 39(1), 137-164. doi:10.2143/AST.39.1.2038060.

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

Li, K., Wang, R., Lei, H., Zhang, T., Liu, Y., Zheng, X. (2018). Interval prediction of solar power using an Improved Bootstrap method. *Solar Energy*, 159:97-112. https://doi.org/10.1016/j.solener.2017.10.051.

MacKay, D. J. C. (1992) A practical bayesain framework for backpropagation networks. *Neural computation*. 4(3): 448-472. `http://www.mitpressjournals.org/doi/10.1162/neco.1992.4.3.448`

Makridakis, S., Spiliotis, E., Assimakopoulos, V. (2020). The M4 competition: Results, findings, conclusions and way forward. *International Journal of Forecasting*, 34(4): 802–808.

Mazloumi, E., Rose,G., Currie, G., Moridpour, S. (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):534-542. https://doi.org/10.1016/j.engappai.2010.11.004.

McCulloch, W., Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5 (4): 115–133. doi:10.1007/BF02478259.

Mitchell, D., Brockett, P., Mendoza-Arriaga, R., Muthuraman, K. (2013). Modeling and forecasting mortality rates, *Insurance: Mathematics and Economics*, 52(2): 275-285. https://doi.org/10.1016/j.insmatheco.2013.01.002.

Nemeth, L. (2017) Life expectancy versus lifespan inequality: a smudge or a clear relationship? PLoS ONE, 12 (9): e0185702.

Nix, D. A., Weigend, A. S. (1994) Estimating the mean and variance of the target probability distribution, *Proceeding of IEEE International Conference on Neural Networks*, 1:55–60.

Oeppen, J. (2008). Coherent forecasting of multiple-decrement life tables: a test using Japanese cause of death data. *Proceeding of Compositional Data Analysis Conference.*

Oeppen, J., Vaupel, J. W., (2002). Broken limits to life expectancy. *Science*, 296(5570): 1029-1031.

Oeppen, J., Vaupel, J. W. (2006). The linear rise in the number of our days. *Social Insurance Studies, 3. The Linear Rise in Life Expectancy: History and Prospects.*

Pascanu, R., Tomas Mikolov, T., Bengio, Y. (2013). On the difficulty of training Recurrent Neural Networks. *arXiv:1211.5063.*

Pascariu, M. D., Canudas-Romo, V., Vaupel, W. J. (2018). The double-gap life expectancy forecasting model. *Insurance: Mathematics and Economics*, 78: 339-350.

Pitacco, E. (2004). Survival models in a dynami context: a survey. *Insurance: Mathematics and Economics*, 35:279-298.

Pitacco, E., Denuit, M., Habermann, S. and Olivieri, A. (2010). *Modelling Longevity Dynamics for Pensions and Annuity Business.* Oxford University Press, ISBN:9780199547272.

Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45:393–404.

Perla, F., Richman, R., Scognamiglio, S., Wüthrich, M. (2021). Time-series forecasting of mortality rates using deep learning. *Scandinavian Actuarial Journal*. DOI: 10.1080/03461238.2020.1867232

Rabbi, A. M. F., Mazzuco, S. (2020). Mortality Forecasting with the Lee-Carter Method: Adjusting for Smoothing and Lifespan Disparity. *European Journal of Population*, 1-24.

Raftery, A. E., Chunn, J.L., Gerland, P., Ševcíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography*, 50(3): 777-801.

Rau, R., Soroko, E., Jasilionis, D., Vaupel, J. W. (2008). Continued reductions in mortality at advanced ages. *Population and Development Review*, 34: 747-768.

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

Richman, R., Wüthrich, M. (2019a). A neural network extension of the Lee-Carter model to multiple populations. *Annals of Actuarial Science*, 1-21. doi:10.1017/S1748499519000071

Richman, R., Wüthrich, M. (2019b). Lee and Carter go Machine Learning: Recurrent Neural Networks. *Available at SSRN: https://ssrn.com/abstract=3441030 or http://dx.doi.org/10.2139/ssrn.3441030.*

Riley J. (2001). *Rising life expectancy: A global history*. Cambridge University Press.

Rojas, R., Feldman, J. (1996). *Neural networks: A systematic introduction*. Springer, Heidelberg.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by backpropagating errors. *Nature*, 323(6088): 533-536.

Siegel, J. W., Xu, J. (2019). On the approximation properties of Neural Networks. *arXiv:1904.02311v1.*

Schäfer, A. M., Zimmermann, H. G. (2007). Recurrent Neural Networks are universal approximators. *International Journal of Neural Systems*, 17(4): 253–263.

Shkolnikov, V.M., Andreev, E.M., Begun, A.Z. (2003). Gini coefficient as a life table function: Computation from discrete data, decomposition of differences and empirical examples. *Demographic Research*, 8(11): 305–358.

Tibshirani, R. (1996). A comparison of some error estimates for neural network models. *Neural Computation*, 8:152-163.

Torri, T. and J. W. Vaupel. (2012). Forecasting life expectancy in an international context. *International Journal of Forecasting*, 28(2): 519–531.

Turing, A. (1948). Intelligent Machinery. *Computers & Thought*, 11-35.

Van Raalte, A.A., Caswell, H. (2013). Perturbation analysis of indices of lifespan variability. *Demography*, 50: 1615–1640.

Van Raalte, A., Sasson, I., Martikainen, P. (2018). The case for monitoring lifespan inequality. *Science*,30 Nov 2018 : 1002-1004.

Vaupel J. W. (1986). How change in age-specific mortality affects life expectancy. *Population Studies*, 40: 147–157.

Vaupel J. W. (1997). The remarkable improvements in survival at older ages. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, 352: 1799–1804.

Vaupel, J. W., Canudas-Romo, V. (2003). Decomposing change in life expectancy: a bouquet of formulas in honor of Nathan Keyfitz's 90th birthday. *Demography*, 40(2): 201-216

Vaupel, J. W., Zhang, Z., van Raalte, A. (2011). Life expectancy and disparity: an international comparison of life table data. *British Medical Journal*, 1(1).

Vapnik, N. V. (1999). An overview of Staistical Learning Theory. *IEEE Transactions on Neural Networks*, 10(5).

Villegas, A. M., Millossovich, P., Kaishev, V. (2015). *Stmomo: An r Package for Stochastic Mortality Modelling*. URL: `https://cran.r-project.org/web/packages/StMoMo/vignettes/StMoMoVignette.pdf`.

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1(4), 339–356.

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

Wilmoth, J.R., Horiuchi, S. (1999). Rectangularization revisited: variability of age at death within human populations. *Demography*, 36(4): 475–495.

Zeiler, M.D. (2012). ADADELTA: An Adaptive Learning Rate Method. `http://arxiv.org/abs/1212.5701`.

Zhou, Z. H., Wu, J., Tang, W. (2002) Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137(1–2): 239-263. https://doi.org/10.1016/S0004-3702(02)00190-X.