# Backtesting Recurrent Neural Networks with Gated Recurrent Unit: Probing with Chilean Mortality Data

**Jorge M. Bravo[1] Vitor Santos[2]**

[1] niversidade Nova de Lisboa - NOVA IMS & Université Paris-Dauphine PSL & MagIC & CEFAGE-UE, Lisbon, Portugal, ORCID: 0000-0002-7389-5103
jbravo@novaims.unl.pt

[2] Universidade Nova de Lisboa - NOVA IMS) ORCID: 0000-0002-4223-7079
vsantos@novaims.unl.pt

***This is the Author Peer Reviewed version of the following chapter/ conference contribution published by Springer:***

Bravo, J. M., & Santos, V. (2022). Backtesting Recurrent Neural Networks with Gated Recurrent Unit: Probing with Chilean Mortality Data. In M. V. Garcia, F. Fernández-Peña, & C. Gordón-Gallegos (Eds.), Advances and Applications in Computer Science, Electronics, and Industrial Engineering: Proceedings of the Conference on Computer Science, Electronics and Industrial Engineering (CSEI 2021) (pp. 159-174). [9] (Lecture Notes in Networks and Systems; Vol. 433). Springer. https://doi.org/10.1007/978-3-030-97719-1_9

# Backtesting Recurrent Neural Networks with Gated Recurrent Unit: Probing with Chilean Mortality Data

Jorge M. Bravo [1] Vitor Santos [2]

[1] Universidade Nova de Lisboa - NOVA IMS & Université Paris-Dauphine PSL & MagIC & CEFAGE-UE, Lisbon, Portugal, ORCID: 0000-0002-7389-5103
jbravo@novaims.unl.pt

[2] Universidade Nova de Lisboa - NOVA IMS) ORCID: 0000-0002-4223-7079
vsantos@novaims.unl.pt

**Abstract.** Understanding the survival prospects of a given population is essential in multiple research and policy areas, including public and private health care and social care, demographic analysis, pension systems evaluation, the valuation of life insurance and retirement income contracts, and the pricing and risk management of novel longevity-linked capital market instruments. This paper conducts a backtesting analysis to assess the predictive performance of Recurrent Neural Networks (RNN) with Gated Recurrent Unit (GRU) architecture in modelling and multivariate time series forecasting of age-specific mortality rates on Chilean mortality data. We investigate the best specification for one, two, and three hidden layers GRU networks and compare the RNN's forecasting accuracy with that produced by principal component methods, namely a Regularized Singular Value Decomposition (RSVD) model. The empirical results suggest that the forecasting accuracy of RNN models critically depends on hyperparameter calibration and that the two hidden layer RNN-GRU networks outperform the RSVD model. RNNs can generate mortality schedules that are biologically plausible and fit well the mortality schedules across age and time. However, further investigation is necessary to confirm the superiority of deep learning methods in forecasting human survival across different populations and periods.

**Keywords:** Recurrent Neural Networks (RNN); Gated Recurrent Unit (GRU); Mortality modelling and forecasting; Pensions; life insurance; backtesting.

## 1. Introduction

Understanding the dynamics of the survival prospects of a given population is vital in multiple research and policy areas, for instance, in public and private health care planning (e.g., of preventive actions, of long-term care needs, of epidemiological episodes), in demographic analysis (e.g., population projections, ageing assessment), in pension

systems design, reform and solvency analysis, in the pricing and risk management of novel longevity-linked capital market instruments (e.g., q-forwards, longevity bonds, longevity swaps), or the valuation of life insurance and private (individual, occupational) retirement income schemes (Coughlan et al., 2007; Bravo & Silva, 2006; Blake et al., 2019; Bravo, 2016, 2019, 2021a,b; Ashofteh & Bravo, 2020; Simões et al., 2021).

To reduce or eliminate the short- and long-term solvency concerns in retirement income schemes created by continuous life expectancy increases, an upward trend in old-age dependency ratios, and insufficient economic growth, in recent decades most countries have responded with parametric (e.g., increasing the retirement age) or structural pension reforms, including the switch from pay-as-you-go (PAYG) defined benefit (DB) plans towards mandatory fully-funded defined-contribution plans (e.g., the 1981 reform in Chile), the introduction of individual complementary funded accounts (e.g., Romania, Hungary, Poland, China) and the transition from classic DB PAYG plans towards Non-Financial Defined Contribution (NDC) schemes (e.g., Sweden, Italy, Latvia) (OECD, 2019). Another major pension reform trend has been to link earnings-related pension benefits to life expectancy developments. For instance, several countries (e.g., The Netherlands, Slovakia, Denmark, Portugal) automatically indexed their normal and early retirement ages to period life expectancy observed at retirement (Bravo & Herce, 2020; Ayuso et al., 2021a,b). Others have opted to link the first pension benefit to demographic or sustainability factors (e.g., Finland, Portugal) or to transformation (annuity) factors (e.g., Italy, Norway). In France and Italy, the eligibility requirements for a full pension now depend on the number of contribution years linked to longevity trends. Private (and public) retirement income schemes introduced longevity-linked life annuities which differ from the traditional level or inflation-linked annuities in that benefits depend on the dynamics of actual against forecasted survival probabilities (Alho et al., 2013; Bravo & El Mekkaoui, 2018; Bravo, 2021). Retirement income providers are substantially exposed to non-diversifiable longevity (and interest rate) risk.

The increasing use of longevity markers in public policy and private practice led to a growing interest in the development of mortality and longevity forecasting methods. In the actuarial, financial, and demographic literature, the traditional approach to age-specific mortality forecasting is to pursue an empirical identification strategy by which, given some criteria (e.g., BIC information criteria) a unique discrete-time or continuous-time parametric or non-parametric stochastic mortality model is selected from a limited number of methods (see, e.g., Lee & Carter (1992); Dowd et al. (2010); Hyndman et al. (2013); Huang et al. (2009); Zhang et al. (2013); Bravo & Nunes (2021) and references therein). Empirical studies show, however, that there is no single universal mortality forecasting method that performs consistently better across populations. Because of that, and to account for model uncertainty, a recent competing research line recommends the use of model combinations (e.g., Bayesian Model Ensembles) of heterogeneous models (Kontis et al.; 2017; Bravo et al., 2021; Ashofteh & Bravo, 2021; Bravo & Ayuso, 2020, 2021a,b). An emerging modelling approach is to use machine learning and deep learning methods to predict age-specific mortality rates (Deprez et al., 2017; Richman & Wüthrich, 2019; Bravo, 2021c,d; Hong et al., 2021).

This paper conducts a backtesting analysis to assess the predictive performance of Recurrent Neural Networks (RNN) with Gated Recurrent Unit (GRU) architecture in modelling and multivariate time series forecasting of age-specific mortality rates on Chilean mortality data. RNNs are dynamic neural networks extending Feedforward Neural Network (FNN) to tackle network problems when handling time series. The algorithm incorporates an internal working memory (a loop) to step through learning problems that involve sequential input data. This overcomes the limitations of plain vanilla RNN which tend to exhibit reduced capacity to seize long-term trends dependencies in mortality data, generating poor forecasts. RNNs with GRU architecture is one of the most popular RNN structures that aim to solve the vanishing gradient problem when training networks using back-propagation.

We adopt a fixed horizon backtesting approach considering a common medium-term (10-year) lookforward window to train the networks and to produce forecasts of age-specific mortality rates by sex. It is well-known that neural networks are critically dependent on the choice of hyperparameters. To identify the optimum hyperparameters combination for the RNN-GRU neural network, we carried a preliminary fine-tuning round and empirically investigated the sensitivity of the forecasting results against alternative choices in one, two, and three hidden layers RNN-GRU models (number of hidden neurons, number of epochs, optimizer, batch size). To evaluate the predictive precision of RNN-GRU models, we selected a traditional principal component method, namely the Regularized Singular Value Decomposition (RSVD) model proposed by Huang et al. (2009) and Zhang et al. (2013). The in-sample and out-of-sample forecasting error is measured by the Mean Squared Error (MSE) metric. The study goes deepens the preliminary investigations in Bravo (2021c) by examining the importance of the number of hidden neurons in the overall network performance. The empirical strategy is based on mortality (deaths classified by sex, age, calendar year, and birth cohort) and population (exposure-to-risk) data for Chile from 1992 to the latest available year (2017). The data source is the Human Mortality Database (HMD, 2021). A common difficulty when modelling Latin American countries' longevity is the lack of sufficient past information on mortality trends disaggregated by individual age. Chile is the sole Latin American country available in HMD and this exercise could serve as a point of reference for life table preparation in this region.

The empirical results suggest that the performance of the RNN-GRU network depends on the number of hidden layers and the choice of the hyperparameters. The addition of hidden layers contributes to improving the model performance (minimizes the forecasting error) up to a certain point, after which further addition of layers reduces the model's accuracy because of overfitting. Two and three hidden layer networks outperformed the RSVD model in the validation dataset. The best RNN-GRU networks can produce consistent and biologically plausible mortality schedules across most ages of the human lifespan. We believe further investigation considering alternative RNN networks, different backtesting approaches (e.g., rolling fixed-length horizon backtests, jumping fixed-length horizon backtests), and alternative populations are however needed to confirm or reject the pilot results obtained in this study.

The remainder of the paper is organized as follows. In Section 2, we describe the data and model specification used in this study, namely the Recurrent Neural Networks

with Gated Recurrent Unit structure, the RSVD model, the methods used to compute period life expectancy, and the learning data. Section 3 presents and briefly discusses the empirical results. Section 4 concludes and sets up the key areas for further research.

## 2. Data and model specification

### 2.1. Recurrent Neural Networks with Gated Recurrent Unit architecture

Cho, et al. (2014) introduced Gated Recurrent Unit (GRU) to, like LSTM, solve the short-term memory problem of plain vanilla RNN. GRU is slightly less complex but is approximately as good as an LSTM performance-wise. GRU share many characteristics with the more complex structure of LSTM networks, namely the basic idea of using a gating mechanism to learn from long-term dependencies in the data and to decide which and how much past information on the time series should be forwarded to the output, but there are some important differences (Figure 1). For instance, contrary to LSTM that have three gates (an input gate, a forget gate, and an output gate), a GRU has only two gates (a reset gate and an update gate) and does not have the output gate that determines how much to reveal of a cell. In addition, Second, a GRU does not include an internal memory differing from the exposed hidden state. The input and forget gates are connected by an update gate $z_t$ and the reset gate $r_t$ is applied directly to the previous hidden state.

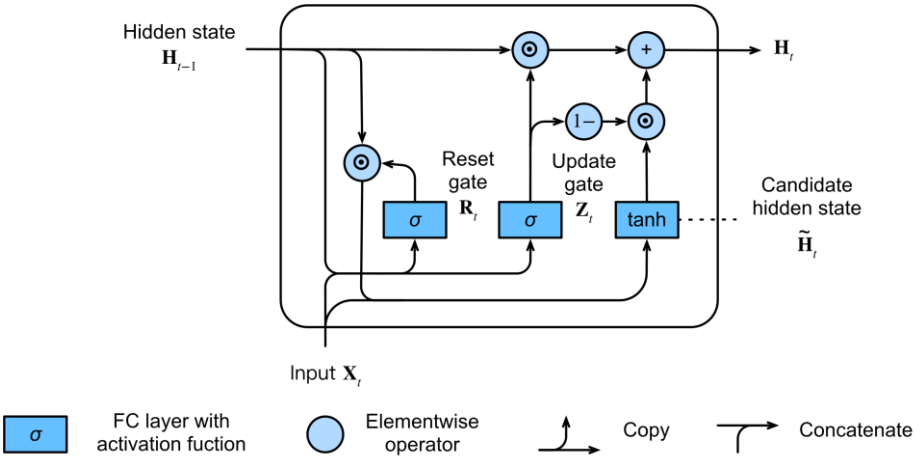


**Fig. 1** – Schematic representation of a Gated Recurrent Unit (GRU) block structure.
Source: Zhang et al. (2021).

Following Richman & Wüthrich (2019), Zhang et al. (2021) and Bravo (2021c), let $(x_1, \ldots, x_T)$ denote a time series of data (age-specific mortality rates) with components $x_t \in \mathbb{R}^{\tau_0}$ observed at times $t = 1, \ldots, T$. Our goal is to use this data as explanatory features to forecast a given output data $y \in \mathcal{Y} \subset \mathbb{R}$. Let $W \in \mathbb{R}^{\tau_0 \times h}$ and $U \in \mathbb{R}^{h \times h}$ denote

the weight matrices for the input and the previous hidden-state result gates, respectively, with $h \in \mathbb{N}$ denoting the number of GRU blocks in a hidden layer. The unit receives as initial information flow the output from the previous GRU unit $h_{t-1} \in \mathbb{R}^h$ and the current input $x_t \in \mathbb{R}^{\tau_0}$. The reset gate controls how much of the previous hidden state we want to remember, capturing short-term dependencies in the time series. The update gate controls how much of the new state resembles the old one, i.e., it captures the long-term dependencies in the data sequences. We use the sigmoid activation function to engineer input values to be in the interval (0, 1). The RNN with GRU architecture can be formally described by the following set of equations:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r), \tag{1}$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z), \tag{2}$$

$$\tilde{h}_t = \phi(W_h x_t + (r_t \circ h_{t-1}) U_h + b_h), \tag{3}$$

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t, \tag{4}$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \in (0,1), \tag{5}$$

$$\phi(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \in (-1,1) \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function deciding how much input data should be used to update the memory of the network, $\phi(\cdot)$ is the hyperbolic tangent function controlling for the importance of the values which are passed, and $b_r$, $b_z$ and $b_h$ are biases. Equation (1) and equation (2) define, respectively, the reset gate (short term memory) and the update gate (long-term memory) mechanisms. Equation (3) describes the dynamics of the candidate hidden state, integrating the reset gate with the regular hidden state updating mechanism. Equation (4) determines the hidden state updating mechanism.

We empirically investigate different choices of the hyperparameters of one, two, and three hidden layers GRU networks (e.g., the number hidden neurons) considering for a fixed 10-year look-forward window, alternative values for the number of epochs, alternative optimizers, and the Mean Squared Error (MSE) as loss function. To calibrate the models, all RNN and principal component approaches are trained on the training set years $\mathcal{D}_1^{train} = \{t \in \mathcal{D} \ , 1992 \leq t \leq 2007\}$. Figure 2 illustrates the decomposition of the training data into the test and validation data as part of the backtesting exercise. The predictive accuracy was assessed on $\mathcal{D}_1^{test} = \{t \in \mathcal{D} \ , 2008 \leq t \leq 2017\}$ using the MSE, computed as

$$MSE_g = \frac{1}{N} \sum_{t=t_{min}}^{t_{max}} \sum_{x=x_{min}}^{x_{max}} \left(\mu_{x,t,g} - \hat{\mu}_{x,t,g}\right)^2, \tag{7}$$

with $N = (x_{max} - x_{min} + 1)(t_{max} - t_{min} + 1)$. Taking the multi-step 10-year forecasts of age-specific mortality rates, we then estimate period life expectancy as follows (Ayuso et al., 2021a):

$$e_{x,g}^p(t) := 0.5 + \sum_{k=1}^{\omega - x} {}_k p_{x,t,g}, \tag{8}$$

where ${}_kp_{x,t,g}$ is the $k$-year survival probability for an individual aged $x$ at time $t$, and $\omega$ is the highest attainable age in the life table, set at age 120 for all years and both sexes.

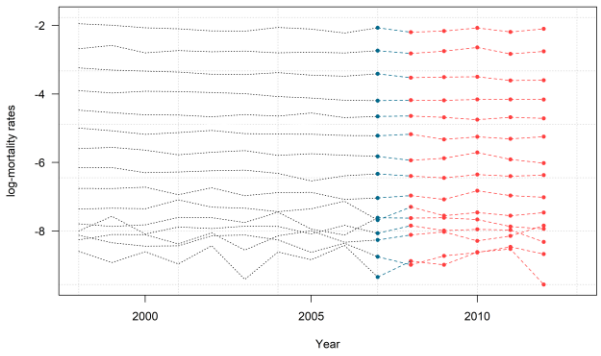


**Fig. 2 –** Backtesting forecasting methods: Training data illustration.

## 2.2. Regularized Singular Value Decomposition model

This paper considers the Regularized Singular Value Decomposition (RSVD) model proposed by Huang et al. (2009) and Zhang et al. (2013) as a benchmark for the RNN-GRU architecture forecasting accuracy. The RSVD model forecasts age-specific mortality rates by extending the one-way functional principal component analysis (PCA) to two-way functional data. This is achieved by introducing regularization of both left and right singular vectors in the SVD of the data matrix. Formally, following Bravo et al. (2021), let $D_{x,t,g}$ be the number of deaths observed at age $x$ during calendar year $t$ from the population (country, sex) $g$ initially ($E^0_{x,t,g}$) or centrally ($E^c_{x,t,g}$) exposed-to-risk. Let $X = \left(m_{x,t}\right)_{n\times p}$ be a data matrix of mortality rates with $n$ ages and $p$ years. The RSVD model assumes that the central mortality rate $m(x,t)$ can be explained in terms of both period $t$ and age $x$ effects as follows:

$$m(x,t) = \sum_{j=1}^{q} \lambda_j U_j(t) V_j(x) + \varepsilon(x,t), \tag{9}$$

where $\lambda_q$ is the singular value, $U_i(\cdot)$ and $V_j(\cdot)$ are smooth functions of period and age, respectively, and $\varepsilon(x,t)$ is a mean zero random noise. The model is fitted iteratively. For instance, the first pair of singular vectors of $X$, $U_1(t)$ and $V_1(x)$, solves the following least-squares problem

$$(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}) = \underset{(u,v)}{\operatorname{argmin}} \|\boldsymbol{X} - \boldsymbol{u}\boldsymbol{v}^T\|_F^2, \tag{10}$$

where $\|\cdot\|_F$ is the Euclidean norm of a matrix. The next pairs are extracted sequentially by removing the effect of the preceding pairs. For two-way functional data, the RSVD of Huang et al. (2009) defines the regularized singular vectors as

$$(\widehat{\boldsymbol{u}}, \widehat{\boldsymbol{v}}) = \underset{(u,v)}{\operatorname{argmin}}\{\|\boldsymbol{X} - \boldsymbol{u}\boldsymbol{v}^T\|_F^2 + \mathcal{P}_\pi(\boldsymbol{u}, \boldsymbol{v})\}, \quad (11)$$

where $\mathcal{P}_\pi(\cdot)$ is a regularization penalty, defined as:

$$\mathcal{P}_\pi(u, v) = \pi_u u^T \Omega_u u \cdot \|v\|^2 + \pi_v v^T \Omega_v v \cdot \|u\|^2 + \pi_u u^T \Omega_u u \cdot \pi_v v^T \Omega_v v, \quad (12)$$

where $\Omega_u$ $(n \times n)$ and $\Omega_v$ $(p \times p)$ are symmetric and nonnegative definite domain-specific penalty matrices. Their objective is to balance the model goodness-of-fit against smoothness in mortality across age and time; $\pi$ is a vector of regularization parameters optimally estimated based on generalized cross-validation criterion. To forecast mortality rates and derive confidence intervals, we use general univariate ARIMA processes to model the time functions $U_i(t)$.

### 2.3. Mortality data

The Chilean mortality data used in this study are publicly available from the HMD. The datasets comprehend the number of recorded deaths together with the corresponding resident population counts (exposure-to-risk), classified by individual age $x$ $(\mathcal{X} = \{x \in \mathbb{N}, 0 \leq x \leq 110+\}$, calendar year $\mathcal{T} = \{t \in \mathbb{N}, 1992 \leq t \leq 2017\}$, year of birth $c = t - x$ and sex. Figure 3 plots the raw log-mortality rates $\widehat{m}_{x,t,g}$ by age in the range 0 to 100 years old and sex (left panel: Male; right panel: Female).

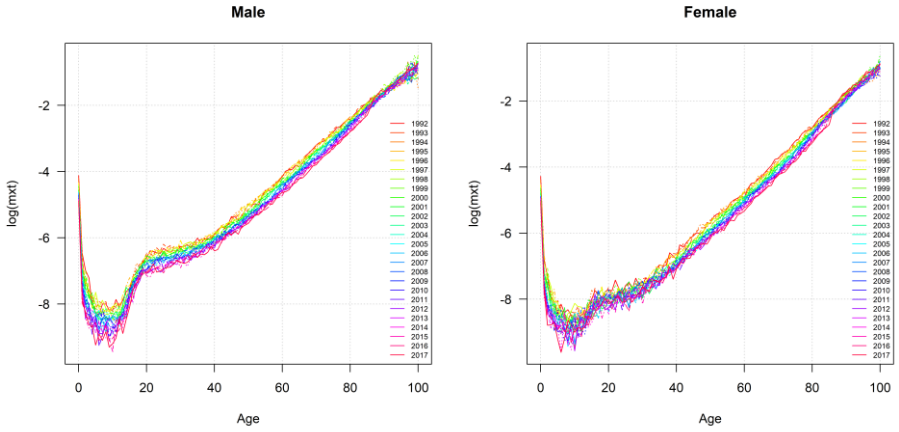


**Fig. 3 –** Crude log-mortality rates by age and sex, Chile, 1992-2017

The longevity trends observed in Chile in the last 25 years are very similar to those observed in developed countries, with a clear downward trend in mortality rates at all ages for both sexes. Chile is one of the countries with the highest life expectancy at all ages on the American continent. The most significant longevity improvements observed in Chile were recorded at younger ages, the exception being the accident hump in the male population in the age range 15-25 years old, and between women. Like in most countries of the world, Chilean women exhibit, on average, higher survival prospects than men of all ages. Increases in life expectancy at adult and old ages are also

important. Figure 4 highlights the dynamics of longevity at all ages through a heatmap and a contour plot of the crude log-mortality rates by age, year, and sex. For both sexes, blue (orange) color represents low (high) mortality. We can observe that longevity gains have been shifting progressively and consistently from early ages to adult and old ages.

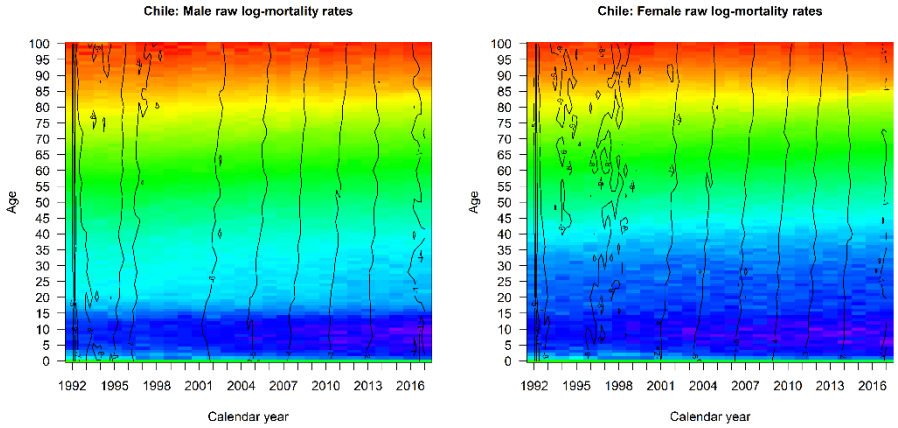


**Fig. 4 –** Heatmap and contour plots of raw log-mortality rates by age and sex, Chile, 1992-2017
Notes: for both sexes, blue (orange) color represents low (high) mortality.

## 3. Results

### 3.1. Hyperparameter calibration

We calibrate models for the male and female populations separately, i.e., we assume as usual that the mortality schedules of both sexes are independent. For very long-term forecasts, this may generate incoherence but given the lookforward window considered in this study, we believe it is a reasonable assumption.[1] For model training and learning, we randomly split the training data into a learning (test) dataset containing 80% of the data and a validation dataset comprehending the residual 20%. To calibrate the Stochastic Gradient Descent algorithm (SGD), we have experimented with different batch sizes and the number of epochs, controlling for the numbers of training samples processed before the model's internal parameters are updated and the number of complete passes through the training dataset. Based on the hyperparameter calibration, we finally opted to run 500 epochs (in batches of 100) of the SGD on the learning dataset. We have experimented with alternative optimizers (Nadam, Adamax, Adam,…) opting finally to use the Adaptive Moment Estimation (Adam) which revealed to be computationally efficient and more reliable, reaching a global minimum when minimising the cost function in training neural nets, and requiring reduced memory given the large

[1] For some examples on the joint modelling of both sexes' mortality schedules see, e.g., Hyndman et al. (2013), Richman & Wüthrich (2019) and Bravo (2021c).

nature of the problem in terms of data and number of parameters. For each model, the optimal calibration is identified by selecting the parameter combination with the lowest MSE loss in the test set.

For the one, two, and three hidden layers RNN-GRU architectures, we calibrated 6, 12, and 24 different networks $\mathcal{M}_j$ , respectively, considering all possible combinations in the array $\mathcal{M}_j = \{\tau_0 = \{1,3\}; \tau_1 = \{5,10,20\}; \tau_2 = \{10,15\}; \tau_3 = \{5,10\}\}$. The RNN-GRU networks with the lowest forecasting error were re-trained on $\mathcal{T}_1^{train}$, from which forecasts of age-specific mortality rates on $\mathcal{T}_1^{test}$ were produced. The model fitting and forecasting procedures have been implemented using a routine running on R software. Table 1 summarizes the average fitting and forecasting loss metrics for all the RNN-GRU hyperparameter combinations tested for the female population of Chile.[2]

**Table 1.** RNN-GRU: fitting/forecasting loss metrics for different hyperparameter combinations

| Parameters | | | | MSE | | CPU | Parameters | | | | MSE | | CPU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau_0$ | $\tau_1$ | $\tau_2$ | $\tau_3$ | Fit | For. | time | $\tau_0$ | $\tau_1$ | $\tau_2$ | $\tau_3$ | Fit | For. | time |
| Panel A: $RNN-GRU_1(\tau_0;\ \tau_1)$ | | | | | | | | | | | | | |
| 1 | 5 | | | 0.54 | 2.74 | 35.5 | 3 | 10 | | | 0.69 | 4.41 | 61.2 |
| 3 | 5 | | | 1.31 | 9.29 | 67.8 | 1 | 20 | | | 0.53 | 3.02 | 39.9 |
| 1 | 10 | | | 1.02 | 5.75 | 46.2 | 3 | 20 | | | 0.46 | 2.84 | 36.7 |
| Panel B: $RNN-GRU_2(\tau_0;\ \tau_1;\ \tau_2)$ | | | | | | | | | | | | | |
| 1 | 5 | 10 | | 0.54 | 2.46 | 71.7 | 1 | 5 | 15 | | 1.50 | 6.42 | 32.2 |
| 3 | 5 | 10 | | 0.93 | 5.47 | 42.1 | 3 | 5 | 15 | | 0.73 | 1.95 | 33.9 |
| 1 | 10 | 10 | | 0.71 | 4.64 | 37.7 | 1 | 10 | 15 | | 0.81 | 1.91 | 30.5 |
| 3 | 10 | 10 | | 0.58 | 2.61 | 31.1 | 3 | 10 | 15 | | 1.41 | 8.38 | 31.8 |
| 1 | 20 | 10 | | 1.54 | 1.82 | 30.5 | 1 | 20 | 15 | | 1.56 | 1.38 | 31.3 |
| 3 | 20 | 10 | | 0.65 | 4.48 | 33.9 | 3 | 20 | 15 | | 0.73 | 4.53 | 33.2 |
| Panel C: $RNN-GRU_3(\tau_0;\ \tau_1;\ \tau_2;\ \tau_3)$ | | | | | | | | | | | | | |
| 1 | 5 | 10 | 5 | 0.55 | 3.31 | 38.3 | 1 | 5 | 10 | 10 | 1.65 | 7.37 | 42.3 |
| 3 | 5 | 10 | 5 | 0.73 | 2.52 | 42.4 | 3 | 5 | 10 | 10 | 0.52 | 3.68 | 42.7 |
| 1 | 10 | 10 | 5 | 0.58 | 3.69 | 41.8 | 1 | 10 | 10 | 10 | 0.69 | 4.47 | 43.7 |
| 3 | 10 | 10 | 5 | 0.57 | 2.63 | 42.7 | 3 | 10 | 10 | 10 | 0.79 | 4.56 | 45.5 |
| 1 | 20 | 10 | 5 | 1.03 | 1.95 | 43.8 | 1 | 20 | 10 | 10 | 0.51 | 3.33 | 44.4 |
| 3 | 20 | 10 | 5 | 0.55 | 2.81 | 46.1 | 3 | 20 | 10 | 10 | 0.51 | 3.79 | 45.7 |
| 1 | 5 | 15 | 5 | 0.99 | 1.74 | 41.9 | 1 | 5 | 15 | 10 | 0.68 | 4.54 | 43.0 |
| 3 | 5 | 15 | 5 | 1.34 | 6.82 | 43.9 | 3 | 5 | 15 | 10 | 1.02 | 6.63 | 43.9 |
| 1 | 10 | 15 | 5 | 2.66 | 9.70 | 42.7 | 1 | 10 | 15 | 10 | 0.54 | 3.69 | 44.4 |
| 3 | 10 | 15 | 5 | 0.54 | 2.94 | 45.5 | 3 | 10 | 15 | 10 | 1.12 | 1.92 | 48.3 |
| 1 | 20 | 15 | 5 | 0.59 | 2.83 | 45.4 | 1 | 20 | 15 | 10 | 0.57 | 3.23 | 50.2 |
| 3 | 20 | 15 | 5 | 1.93 | 2.54 | 47.7 | 3 | 20 | 15 | 10 | 0.55 | 3.15 | 53.8 |

Notes: $\tau_0$. $\tau_1$. $\tau_2$ and $\tau_3$ denote the number of hidden neurons in the hidden GRU layers; Average results for the female population considering for 10-year forecasting horizons. MSE values in $10^{-5}$. CPU time in seconds.

[2] Due to space constraints, the results for the male population are not included in the main manuscript but are available from the authors upon request.

We also report the run times for each model, measured in seconds on a personal laptop with Intel(R) Core(TM) i7-10510U CPU@2.30GHz with 16GB RAM. For the one hidden layer model (Panel A), the best performing network includes five hidden neurons, i.e., RNN-$GRU_1(\tau_0 = 1;\ \tau_1 = 5)$. For two hidden layer models (Panel B), the highest accuracy is obtained with the specification RNN-$GRU_2(\tau_0 = 1;\ \tau_1 = 20;\ \tau_2 = 15)$. Finally, for three hidden layer models (Panel C), the lowest forecasting error is obtained with the RNN-$GRU_3(\tau_0 = 1;\ \tau_1 = 5;\ \tau_2 = 15;\ \tau_3 = 5)$ model.

Figure 5 illustrates the learning strategy on the best RNN-GRU networks, plotting the early stopping in-sample and the out-of-sample loss on the test dataset. We can observe that the addition of hidden layers helps improve the model performance up to a certain point, after which further addition of layers reduces the model's performance, showing signs of clear overfitting.


**Fig. 5 –** Best RNN-GRU architectures with one, two, and three hidden layers, Women.

Table 2 summarizes the predictive accuracy metrics of the best RNN-GRU models tested in this study and of the RSVD used as the benchmark.

**Table 2.** Forecasting accuracy metrics of the alternative RNN and GAPC models tested

| Model | GRU: #hidden neuron parameters | | | | MSE | | CPU |
|---|---|---|---|---|---|---|---|
| | $\tau_0$ | $\tau_1$ | $\tau_2$ | $\tau_3$ | Fit | Forecast | time |
| RNN-$GRU_1$ | 1 | 5 | | | 0.54 | 2.74 | 35.5 |
| RNN-$GRU_2$ | 1 | 20 | 15 | | 1.56 | 1.38 | 31.3 |
| RNN-$GRU_3$ | 1 | 5 | 15 | 5 | 0.99 | 1.74 | 41.9 |
| RSVD | | | | | 1.19 | 1.85 | 21.2 |

Notes: Results obtained considering 10-year look-forward periods; MSE values in $10^{-5}$ units. CPU time in seconds.

The empirical results suggest that the performance of the RNN-$GRU_2$ network depends on the number of hidden layers and the choice of the hyperparameters, chiefly

the number of hidden neurons per layer. The two and three hidden layer networks outperformed the RSVD model in the validation dataset. The extra computation time required by some RNN networks is not sufficient, at this stage, to discard this model as a valid option for longevity risk modelling.

### 3.2. Forecasts of period life expectancy

Figure 6 illustrates the observed (black dots) and the forecasted (orchid color line) log-mortality rates by individual year generated by the best RNN-GRU for the female population of Chile.

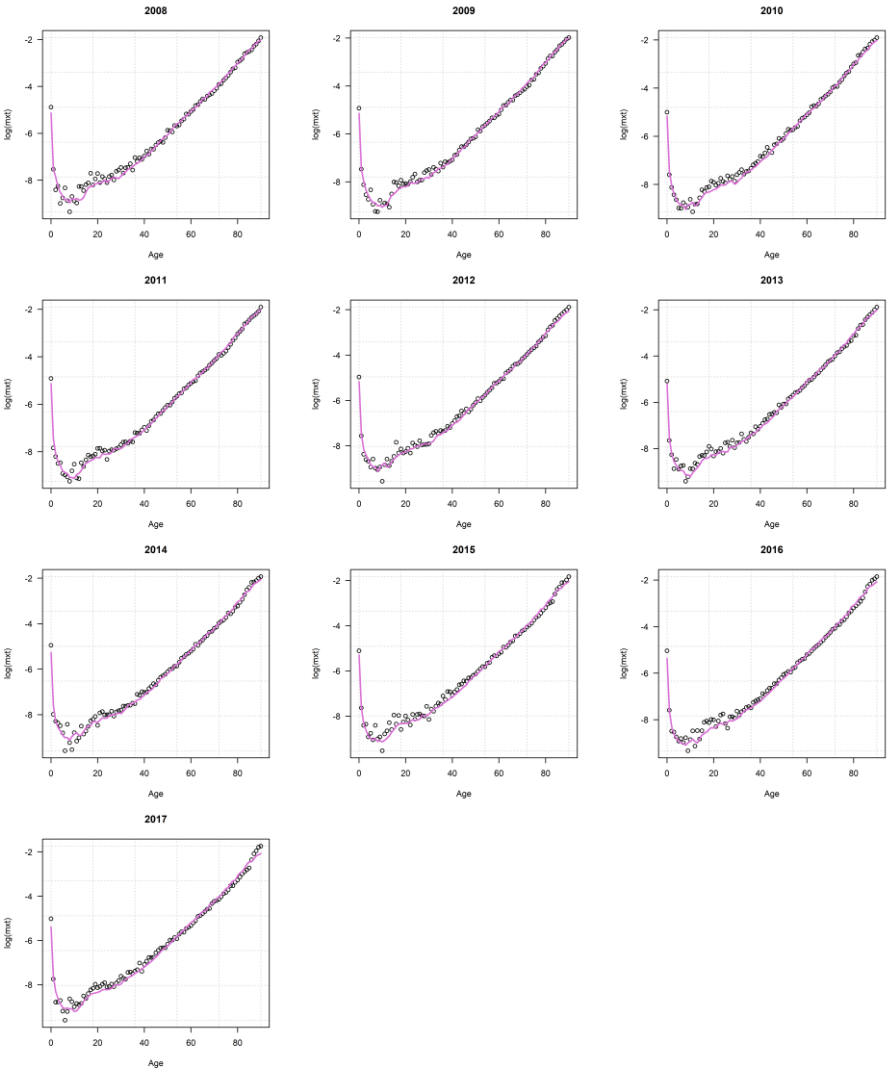


**Fig. 6 –** Best RNN-$GRU_2$ model: Forecasts of log-mortality rates by age and year. Women

Figure 7 represents the corresponding (aggregate) forecasts using the RSVD model. The best RNN-GRU networks were able to produce consistent and biologically plausible mortality schedules across the entire lifespan spectrum, including at younger ages where the volatility is normally higher. Finally, to illustrate the application of forecasts of age-specific mortality in life table computation, we exhibit in Figure 8 the estimates of the period life expectancy computed at birth ($x = 0$) and the benchmark retirement age of 65 ($x = 60$) for Chilean women. The vertical cyan line marks the split between the training and validation datasets. We can observe that the average remaining lifetime in the country has been increasing consistently over the last quarter of a century, from 77.28 (17.57) years at birth (age 65) in 1992 to 81.70 (20.57) years in 2017. Similar results (at a lower level) were obtained for the male population. The increasing life expectancy challenges the solvency of public and private pension schemes and has important implications in individual consumption, saving and, labour market decisions.

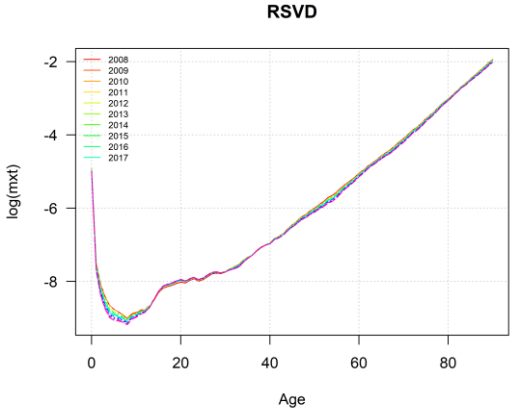


**Fig. 7 –** RSVD model: Forecasts of log-mortality rates by age, Women.


**Figure 8 –** RNN-GRU2: Estimates of the life expectancy at birth and at the age of 65, women.

## 4. Conclusion

Model selection and model ensembles are currently the two main competing approaches when modelling and forecasting age-specific mortality for actuarial, financial, and demographic applications. The pool of individual candidate models includes generalized age-period-cohort stochastic mortality models, principal component methods, and smoothing approaches. More recently, attempts have been made to use machine learning and deep learning methods for multivariate forecasting. This paper follows this latter research trend and conducts a backtesting analysis to assess the predictive performance of RNN with GRU architecture in multivariate time series forecasting of age-specific mortality rates on Chilean mortality data. We compare the RNN performance with that offered by traditional principal component methods (RSVD model), widely used in actuarial and demographic studies. The empirical results obtained on a limited dataset suggest that the forecasting accuracy of RNN-GRU networks outperforms the RSVD model. However, the results also suggest that the RNN-GRU predictive accuracy is critically dependent on hyperparameter calibration and that adding extra hidden layers may lead to model overfitting. This is important since most longevity-linked securities and insurance contracts are typically very long-term contracts with pricing fixed at contract initiation and without the possibility of revision if observed longevity trends deviate from assumed improvements. Conceptual uncertainty (model risk) must be incorporated into pricing, risk management, and inference purposes. One way of doing this is to combine heterogeneous stochastic mortality models using, for instance, a Bayesian model ensemble approach (see, e.g., Bravo et al., 2021).

The mortality schedules produced by RNNs are biologically plausible, which is an important criterion for a good stochastic mortality model, and consistent across all ages of the human lifespan. This is an advantage when compared to some of the classical approaches to mortality forecasting which have proved to perform poorly when applied to both young, adult, and oldest-old age groups.

However, to be able to confirm or reject the claim that RNN models can be added to the toolkit of researchers and professionals working in longevity risk management or public policy analysis, we believe that further investigation is required to investigate extensively the sensitivity of the results to, for instance, hyperparameter choices, the type of network architecture, the lookback and lookforward window, the accuracy metric, or the population characteristics. This is on the agenda for further research.

Extending research to multiple state mortality models accounting for longevity heterogeneity is a priority to tackle actuarial fairness considerations in both social policy (e.g., public pension scheme design) and private insurance contracts (e.g., life insurance). Adopting non-uniform policy approaches considering the ex-ante life expectancy gradient, e.g., implementing differential retirement ages, sustainability factors, or social contribution rates, are some of the possible reform avenues aiming at reducing the redistributive distortions created by longevity heterogeneity (Bravo & Ayuso, 2021b). This can be done by using, e.g., multi-state models, which have been successfully applied to other problems such as long-term care and credit risk modelling (Chamboko & Bravo, 2016, 2020).

Despite the high predictive power of RNN-GRU against RSVD models found in this paper, the unsatisfactory interpretability of neural networks in mortality forecasting is still one of the key obstacles of deep learning techniques in its wide acceptance by the financial industry. For example, the European Union regulations provide customers impacted by tailored pricing algorithms the right to ask and receive an explanation for why a model makes a particular decision under specific circumstances, and the chance to benefit from fair algorithmic competition. Auditors and supervisors need to understand and approve internal risk models. This creates a big challenge when communicating deep learning algorithms and results.

**Acknowledgments**: The authors express their gratitude to the editors and the anonymous referees for his or her careful review and insightful comments, which helped strengthen the quality of the paper. The authors were supported by Portuguese national funds through FCT under the project UIDB/04152/2020 - Centro de Investigação em Gestão de Informação (MagIC).

## References

1. Alho, J., Bravo, J. M., & Palmer, E. (2013). Annuities and life expectancy in NDC. In Holzmann, R., Palmer, E. & Robalino, D. (Eds.) Nonfinancial defined contribution Pension Schemes in a Changing Pension World, Volume 2, Gender, Politics, and Financial Stability, 395-436, World Bank Publications. https://doi.org/10.1596/9780821394786_CH22
2. Ashofteh, A., & Bravo, J. M. (2020). A study on the quality of Novel Coronavirus (Covid-19) official datasets. Statistical Journal of the IAOS, 36 (2), 291–301. https://doi.org/10.3233/SJI-200674
3. Ashofteh, A., Bravo, J. M. (2021). Life Table Forecasting in COVID-19 Times: An Ensemble Learning Approach. CISTI'2021 - 16th Iberian Conference on Information Systems and Technologies, p. 1-6, https://doi.org/10.23919/CISTI52073.2021.9476583.
4. Ayuso, M., Bravo, J. M., Holzmann, R. (2021a). Getting Life Expectancy Estimates Right for Pension Policy: Period versus Cohort Approach. Journal of Pension Economics and Finance, 20(2), 212–231. https://doi.org/10.1017/S1474747220000050
5. Ayuso, M., Bravo, J. M., Holzmann, R., & Palmer, E. (2021b). Automatic indexation of pension age to life expectancy: When policy design matters. Risks, 9(5), 96. https://doi.org/10.3390/risks9050096
6. Blake, D., Cairns, A.J.G., Dowd, K., Kessler, A.R. (2019). Still living with mortality: The longevity risk transfer market after one decade. British Actuarial Journal, 24, 1–80.
7. Bravo, J. M. (2016). Taxation of Pensions in Portugal: A Semi-Dual Income Tax System. CESifo DICE Report - Journal for Institutional Comparisons. 14 (1), 14-23.
8. Bravo, J. M. (2019). Funding for Longer Lives: Retirement Wallet and Risk-Sharing Annuities. Ekonomiaz, 96 (2), 268-291.
9. Bravo, J. M. (2021a). Pricing participating longevity-linked life annuities: A Bayesian Model Ensemble approach. European Actuarial Journal. https://doi.org/10.1007/s13385-021-00279-w
10. Bravo, J. M. (2021b). Pricing Survivor Bonds with Affine-Jump Diffusion Stochastic Mortality Models. In 2021 The 5th International Conference on E-commerce, E-Business and E-

Government (ICEEG 2021). Association for Computing Machinery (ACM), New York, NY, USA, 91–96. https://doi.org/10.1145/3466029.3466037

11. Bravo, J. M. (2021c). Forecasting longevity for financial applications: A first experiment with deep learning methods. The 6th ECML PKDD Workshops, MIDAS 2021 The Sixth Workshop on MIning DAta for financial applicationS. Lecture Notes in Computer Science Series (LNCS), Springer Nature, Switzerland, in press.
12. Bravo, J. M. (2021d). Forecasting mortality rates with Recurrent Neural Networks: A preliminary investigation using Portuguese data. CAPSI 2021 Proceedings (Atas da 21ª Conferência da Associação Portuguesa de Sistemas de Informação 2021), in press.
13. Bravo, J. M., & Ayuso, M. (2020). Mortality and life expectancy forecasts using bayesian model combinations: An application to the portuguese population. RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação, E40, 128–144. https://doi.org/10.17013/risti.40.128–145.
14. Bravo, J. M., & Ayuso, M. (2021a). Forecasting the retirement age: A Bayesian Model Ensemble Approach. Advances in Intelligent Systems and Computing, Volume 1365 AIST, 123 – 135 [2021 World Conference on Information Systems and Technologies, WorldCIST 2021] Springer, Cham. https://doi.org/10.1007/978-3-030-72657-7_12.
15. Bravo, J. M., & Ayuso, M. (2021b). Linking pensions to life expectancy: Tackling conceptual uncertainty in social policy through model averaging. Journal of Social Policy. Submitted.
16. Bravo, J. M., & El Mekkaoui de Freitas, N. (2018). Valuation of longevity-linked life annuities. Insurance: Mathematics and Economics, 78, 212-229.
17. Bravo, J. M., & Herce, J. A. (2020). Career Breaks, Broken Pensions? Long-run Effects of Early and Late-career Unemployment Spells on Pension Entitlements. Journal of Pension Economics and Finance 1–27. https://doi.org/10.1017/S1474747220000189
18. Bravo, J. M., & Nunes, J. P. V. (2021). Pricing Longevity Derivatives via Fourier Transforms. Insurance: Mathematics and Economics, 96, 81-97.
19. Bravo, J. M., & Silva, C. M. (2006). Immunization Using a Stochastic Process Independent Multifactor Model: The Portuguese Experience. Journal of Banking and Finance, 30 (1), 133-156.
20. Bravo, J. M., Ayuso, M., Holzmann, R., & Palmer, E. (2021). Addressing the Life Expectancy Gap in Pension Policy. Insurance: Mathematics and Economics, 99, 200-221. https://doi.org/10.1016/j.insmatheco.2021.03.025.
21. Chamboko, R., & Bravo, J. M. (2016). On the modelling of prognosis from delinquency to normal performance on retail consumer loans. Risk Management, 18(4), 264–287.
22. Chamboko, R., & Bravo, J. M. (2020). A Multi-State Approach to Modelling Intermediate Events and Multiple Mortgage Loan Outcomes. Risks, 8, 64.
23. Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: encoder-decoder approaches. arXiv preprint arXiv:1409.1259.
24. Coughlan, G.D., Epstein, D., Honig, P. (2007). Q-Forwards: Derivatives for Transferring Longevity and Mortality Risks. Working Paper, J. P. Morgan Pension Advisory Group, London.
25. Deprez, P., Shevchenko, P., & Wüthrich, M. (2017). Machine learning techniques for mortality modeling. European Actuarial Journal, 7, 337–352.
26. Dowd, K., Cairns, A., Blake, D., Coughlan, G., Epstein, D., Khalaf-Allah, M. (2010). Backtesting stochastic mortality models. North American Actuarial Journal, 14 (3), 281–298.
27. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735–1780.

28. Hong, W.H., Yap, J.H., Selvachandran, G. et al. (2021). Forecasting mortality rates using hybrid Lee–Carter model, artificial neural network and random forest. Complex & Intelligent Systems, 7, 163–189.
29. Huang, J. Z., Shen, H. & Buja, A. (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. Journal of the American Statistical Association 104 (488): 1609-1620.
30. Human Mortality Database (2021). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).
31. Hyndman, R. J., Booth, H. & Yasmeen, F. (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. Demography 50(1), 261–283.
32. Kontis, V., Bennett, J., Mathers, C., Li, G., Foreman, K., & Ezzati, M. (2017). Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble. Lancet 389 (10076), 1323–1335.
33. Lee, R. D., & Carter, L. (1992). Modeling and forecasting U.S. mortality. Journal of the American Statistical Association, 87, 659–671.
34. Richman, R., & Wüthrich, M. (2019). Lee and Carter go Machine Learning: Recurrent Neural Networks. Available at SSRN: https://ssrn.com/abstract=3441030 (accessed on 10 January 2021).
35. Simões, C., Oliveira, L. & Bravo, J. M. (2021). Immunization Strategies for Funding Multiple Inflation-Linked Retirement Income Benefits. Risks, 9(4): 60; https://doi.org/10.3390/risks9040060
36. Zhang, A., Lipton, Z., Li, M., Smola, A. (2021). Dive into Deep Learning. arXiv:2106.11342.
37. Zhang, L., Shen, H., & Huang, J. Z. (2013). Robust regularized singular value decomposition with application to mortality data. The Annals of Applied Statistics, 7(3): 1540-1561.