Communications for Statistical Applications and Methods
2021, Vol. 28, No. 5, 521–536

https://doi.org/10.29220/CSAM.2021.28.5.521
Print ISSN 2287-7843 / Online ISSN 2383-4757

6-Parametric factor model with long short-term memory

Janghoon Choi[1,a]

[a]Korea Insurance Research Institute, Korea

Abstract

As life expectancies increase continuously over the world, the accuracy of forecasting mortality is more and more important to maintain social systems in the aging era. Currently, the most popular model used is the Lee-Carter model but various studies have been conducted to improve this model with one of them being 6-parametric factor model (6-PFM) which is introduced in this paper. To this new model, long short-term memory (LSTM) and regularized LSTM are applied in addition to vector autoregression (VAR), which is a traditional time-series method. Forecasting accuracies of several models, including the LC model, 4-PFM, 5-PFM, and 3 6-PFM's, are compared by using the U.S. and Korea life-tables. The results show that 6-PFM forecasts better than the other models (LC model, 4-PFM, and 5-PFM). Among the three 6-PFMs studied, regularized LSTM performs better than the other two methods for most of the tests.

Keywords: Lee Carter model, 6-PFM, accuracy test, vector autoregression, long short-term memory, regularized long short-term memory

1. Introduction

Population aging is a global trend. According to the UN (2019), the life expectancy in the United States at birth is 78.8 years in 2020 and is expected to be 79.8 years in 2030 and 84.4 years in 2060. The life expectancy of Korea is 82.8 years and is expected to be 84.2 years, and 87.9 years for the same periods, respectively.

The prospect of an increase in life expectancy is a desirable phenomenon from a personal perspective, but it also becomes a concern when considering social costs such as pension funds and health insurance that need to be paid to the elderly. In addition, the stability of the social security system may be undermined as economic growth slows due to a shortage of labor and declines in savings, consumption, and investment.

Forecasting mortality is essential in understanding the future aging level and maintaining financial stability of social insurance systems such as the national pension and health insurance systems. Government policies related to such social systems must rely on a forecast to reduce and effectively manage costs incurred by the rapidly increasing elderly population.

The most commonly used model for forecasting mortality is the Lee-Carter model (Lee and Carter, 1992) due to its easy application. However, various studies have been conducted to improve the Lee-Carter model. Some of these studies are Li and Lee (2005), Renshaw and Haberman (2006), Booth and Tickle (2008), Booth *et al.* (2002, 2006), Hyndman and Ullah (2007), Cairns *et al.* (2006), and Wiśniowski *et al.* (2015). More recently, efforts to apply deep learning algorithms have been made

[1] Korea Insurance Research Institute, 38 Gukjeguemyung-ro 6-gil, Youngdeungpo-gu, Seoul, Korea.
E-mail: james021@gmail.com

Published 30 September 2021 / journal homepage: http://csam.or.kr

by Perla *et al.* (2020), Richma and W*ü*thrich (2019), and Nigri *et al.* (2019). In this paper, the 6-parametric factor model (6-PFM) is introduced. 6-PFM is an improved version of 4-parametric factor model (4-PFM) (Haldrup and Rosenskjold, 2019). Also, forecast methods using deep-learning algorithms, which included LSTM and regularized LSTM were applied in addition to the traditional time series method to 6-PFM. The investigation will determine if the newly designed 6-PFM has a better performance than the existing LC model and 4-PFM as well as 5-parametric factor model (5-PFM) which is also newly designed in this paper. The rest of this paper is organized as follows: Section 2 introduces 6-PFM and section 3 reviews the deep-learning algorithms of LSTM and regularized LSTM. Section 4 fits the models with the data from the U.S. and Korea life-tables and performs accuracy tests. After the tests, real forecasts are carried out and the forecast results are shown. Finally, Section 5 presents the conclusion.

2. The 6-parametric factor model

Haldrup and Rosenskjold (2019) designed 4-PFM by applying the dynamic Nelson-Siegel model (Diebold and Li, 2006), which has been very popular for forecasting interest rates in the financial market, to central death rates. 4-PFM can work reliably even when the mortality structure changes. Choi (2021) showed that 4-PFM was more reliable than the LC model when the mortality structure changed due to COVID-19. However, forecast accuracy was not as good as the LC model in normal circumstances. Thus, 5-PFM and 6-PFM were developed to improve this accuracy. 5-PFM is obtained by adding one more factor and 6-PFM is obtained by adding two more factors to 4-PFM. Mathematical expressions of 4-PFM, 5-PFM and 6-PFM are in equation (2.1), equation (2.2), and equation (2.3), respectively.

$$\ln(m_{x,t}) = \beta_{0t} + \beta_{1t}e^{-\lambda_1 x} + \beta_{2t}e^{-\lambda_2(\ln(x)-\ln(k_1))^2} + \beta_{4t}\left(\frac{x}{\omega}\right)^{\lambda_4} + \epsilon_{x,t}, \tag{2.1}$$

$$\ln(m_{x,t}) = \beta_{0t} + \beta_{1t}e^{-\lambda_1 x} + \beta_{2t}e^{-\lambda_2(\ln(x)-\ln(k_1))^2} + \beta_{4t}\left(\frac{x}{\omega}\right)^{\lambda_4} + \beta_{5t}\left(\frac{x}{\omega}\right)^{\lambda_5} + \epsilon_{x,t}, \tag{2.2}$$

$$\ln(m_{x,t}) = \beta_{0t} + \beta_{1t}e^{-\lambda_1 x} + \beta_{2t}e^{-\lambda_2(\ln(x)-\ln(k_1))^2} + \beta_{3t}e^{-\lambda_3(\ln(x)-\ln(k_2))^2} + \beta_{4t}\left(\frac{x}{\omega}\right)^{\lambda_4} + \beta_{5t}\left(\frac{x}{\omega}\right)^{\lambda_5} + \epsilon_{x,t}. \tag{2.3}$$

where x is age, t is year, and ω is limiting age. The log central death rate ($\ln(m_{x,t})$) of 4-PFM is fitted by 4 factors which are β_{0t}, β_{1t}, β_{2t}, β_{4t}, and corresponding factor loadings which are 1, $e^{-\lambda_1 x}$, $e^{-\lambda_2(\ln(x)-\ln(k_1))^2}$, and $(x/\omega)^{\lambda_4}$. The factor loadings are estimated by parameters λ_1, λ_2, λ_4, and k_1. β_{0t} is a factor applying to all ages in common (Siler, 1979) so that the corresponding loading is 1. β_{1t} is a factor showing death rates of infants (Rogers and Little 1994), and its loading function corresponding to β_{1t} has an L-shape, which means that the function decreases very rapidly as age increases from 0 but the function decreases very slowly after some increase away from 0. The bigger the λ_1 is, the more rapidly decreasing the function is. β_{2t} is a factor expressing death rates from the accident humps in young ages (Heligman and Pollard, 1980), and its loading function has a bell shape, which means the function increases toward a peak age k_1 and decreases thereafter. β_{4t} is a factor expressing death rate increase as age increases (Gompertz, 1825). The loading function corresponding to β_{4t} has a linear shape when λ_4 equals 1, is concave when less than 1, and is convex when greater than 1. $\epsilon_{x,t}$ is an error term and is assumed to follow iid N(0,σ^2).

For 5-PFM, $\beta_{5t}(x/\omega)^{\lambda_5}$ is added to 4-PFM, and for 6-PFM, $\beta_{3t}e^{-\lambda_3(\ln(x)-\ln(k_2))^2}$ is added to 5-PFM to increase the forecasting accuracy. Higher factor models such as 7-PFM and 8-PFM could be designed, but there would be too many factors and loadings to fit in those models, so factor models higher than 6-PFM are not considered.

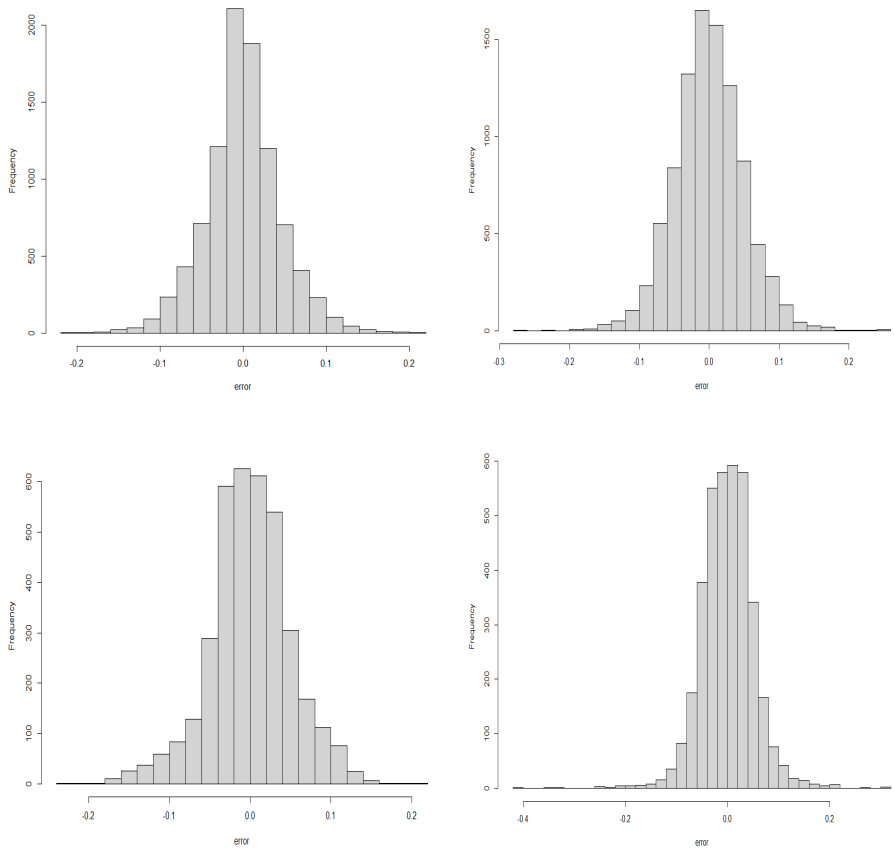

Figure 1: *Histogram of $\epsilon_{x,t}$: above (U.S.) - male(left), female(right) below (Korea) - male(left), female(right).*

In this paper, 6-PFM is selected as a new representative model because 6-PFM performs better than 5-PFM as seen in the accuracy-test results in Section 4.3. Residual analysis also shows that 6-PFM is appropriate. The residual analysis on the error term in equation (2.3), using the 1933~2018 U.S. and 1983~2019 Korea life tables, shows that the errors are normally distributed for 6-PFM as you can see in Figure 1.

6-PFM consists of 6 factors and corresponding 6 factor loadings as in equation (2.3). The factors, which reveal some trends as time passes, are the objects for the forecast. The factor loadings do not have trends as time passes but display patterns according to ages. They are explained as follows:

The two graphs in Figure 2 are the shapes of the loading functions of males and females of the United States fitted by the U.S. life tables from 1933 to 2018. The parameters λ's and k's in the loading functions are estimated by using the 2018 U.S. life-table. The figure above is the shape of a male and the one below is the shape of a female. The common loading functions (black) are the same for both genders because they are fixed at 1. The infant loading functions (red) show infant death rates which have L-shapes as expected. The shapes look similar for both genders but the curve of a

female is a little sharper than the curve of a male because λ_1 for a female is greater than that for a male (2.6 vs. 2.2). This means that the female infant death rate is reduced more sharply than the male infant rate as the age grows near 0. The accident hump loading functions show somewhat different shapes between males and females. Two accident peaks occur at around ages 15(blue) and 20(green) for males and occur at around ages 12(green) and 17(blue) for females. The width of the bell shape for the hump where the accident peak occurs at age 15(blue for male) is quite wide compared with the other 3 curves, which means the accident does not occur narrowly around age 15 compared to the other peak ages at 20, 12, or 17. The aging loading functions (sky-blue and purple) also look a little bit different between the two genders. For males, the sky-blue line is almost linear and the purple line is fairly curved, while the sky-blue line is relatively curved for females. A linear line means that the death rate increases linearly by age, and a curved line means that the death rate increases much faster in later ages. The next two figures in Figure 3 are those of Korean males and females fitted by the Korea life-tables from 1983 to 2019. Like Figure 2, the parameters in the loading functions are estimated by using the 2019 Korea life-table. The figure above is for males and the one below is for females, and the common loading functions (black) are fixed at 1. The shapes of the infant loading functions also look similar for both genders but the curve of females is a little sharper than that of males as in the United States because λ_1 for a female is greater than that for a male(2.6 vs. 1.7). The accident peak occurs at around ages 13(blue) and 18(green) for males and occurs at around ages 11(green) and 25(blue) for females. Compared with those of the United States, the two accident humps occur earlier for the Korean male, and the second accident (blue) occurs later for the Korean female, though the first accident hump occurs at similar ages in both countries. Also, the distance between two accident peak positions for the Korean female is wider than those for the U.S. female. The aging loading functions (sky-blue and purple) also look different between the two genders, which show similar shapes to the U.S. aging loading functions.

Next, let's move onto the factors. The trends of the factors are important because a forecast is conducted based on these trends. While the loadings do not change over time, the factors do. Thus we can forecast the central death rates by analyzing the trends of the factors. The past trends of the 6 factors for the United States and Korea are shown in Figures 4 and Figure 5, respectively. For the U.S., the common factors (black) show some decreasing trends for both male (left) and female (right), but, since they have negative signs the effects of the common factors on the central death rates increase. The infant factors (red) do not show conspicuous trends but steady decreasing trends for both genders. The accident factor 1(green) and the accident factor 2(blue) move in opposite directions for both genders. For males, the accident factor 1 increases, and the accident factor 2 decreases, which means that the accidents of 15-year-old males increase and those of 20-year-old males decrease as time passes. For females, the accident factor 1 decreases, and the accident factor 2 increases, which means that the accidents of 12-year-old females decrease and those of 17-year-old females increase. The aging factor 1's(sky-blue) have the strongest effect on the central death rates among all the factors for both genders because they are at the highest level. However, the aging factor 2's (purple) rarely affect the central death rates for both genders since they are located around 0. Thus, the aging factor2's have the role of fine-tuning the fitting accuracy for the model.

For Korea, the common factors (black) reduce for both males and females as for the U.S. The infant factors rarely change as time goes by, so their effects on the central death rates do not change for both genders. The accident factors do not show conspicuous trends, though the accident factor 2 of females shows a steady increasing trend. Also, the accident factors are located around 0, so the effects of these factors on the central death rates are very small. The aging factor 1 for males (sky-blue) increases towards 2000 and since that year shows some fluctuation; and the aging factor 2 of males

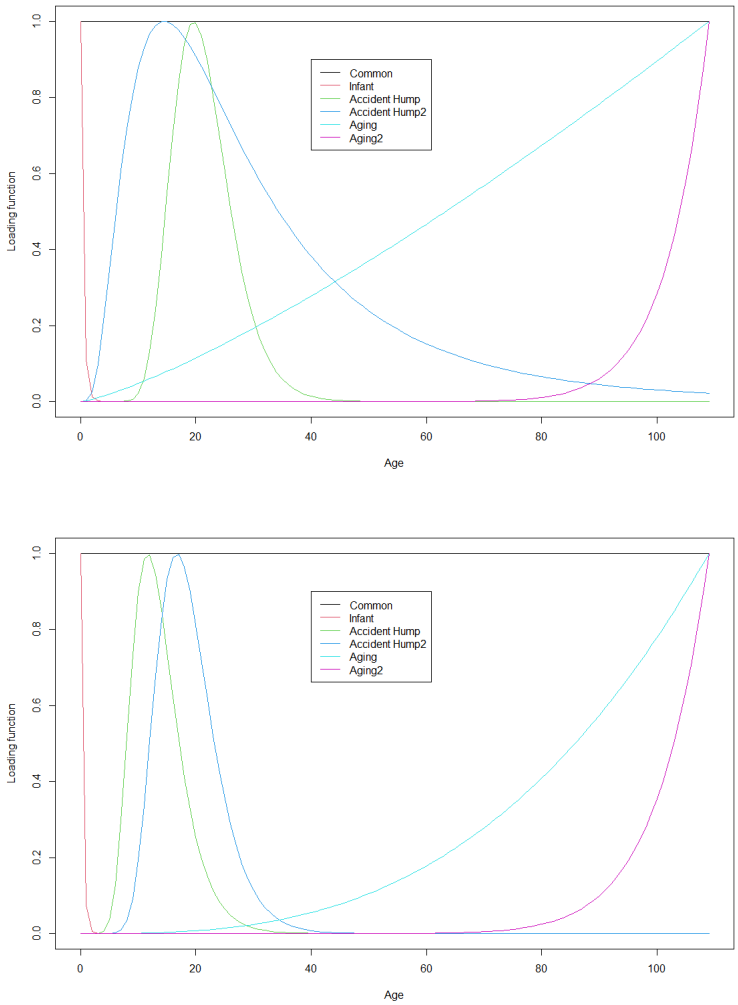

Figure 2: *Shapes of the loading functions in U.S.: above-male, below-female.*

(purple) moves horizontally until around 2000 and shows some increasing trend with fluctuation since the same year. Thus the combined effects of these two factors on the central death rates increase. The aging factor 1 of females (sky-blue) has a strong effect and increases until around 2005, thereafter, it then shows some reduction. The aging factor 2(purple) moves horizontally and moves upwards since around 2005. The aging factors of females are quite differently located compared to those of males. It is not easy to find out why they are so differently located. It might be that the accident occurs more frequently at younger ages (aging factor 1) in Korea.

When we compare the strength of the effects on the central death rates among factors, the infant factor and the aging factors have strong positive effects and the common factors have negative effects for both males and females.

When compared between the two countries, the common factors show negative effects and their

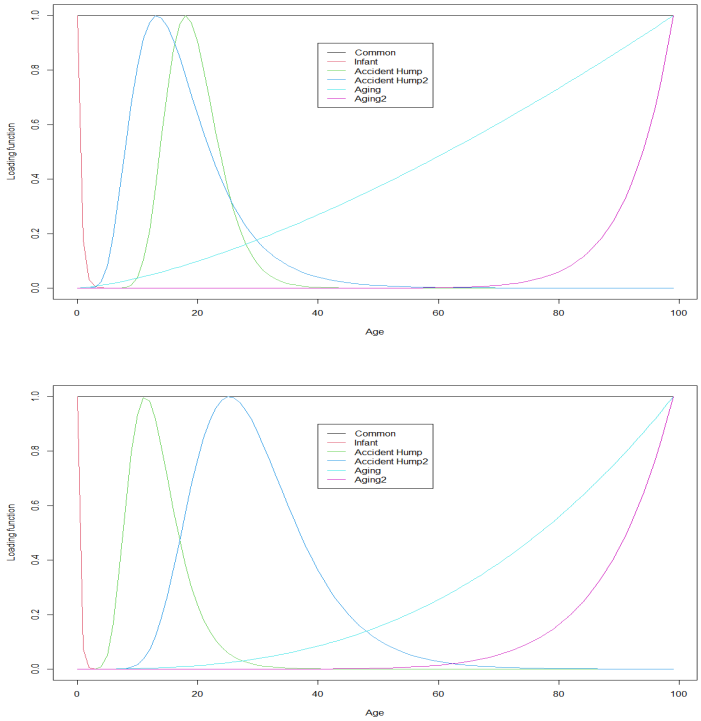

Figure 3: *Shapes of the loading functions in Korea: above-male, below-female.*

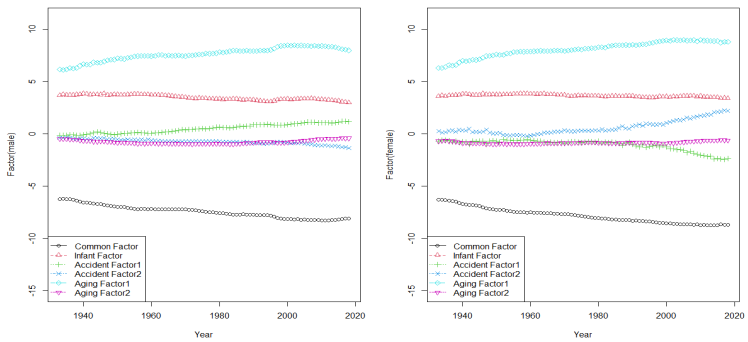

Figure 4: *Trends of 6 factors from yr1933 to yr2018 in U.S.: left-male, right-female.*

strengths increase for both countries. The infant factors also show similar effects for both countries though the trends decrease a little bit for the U.S. The shapes of the accident factor 1 and the accident factor 2 look different between the two countries, but the combined effects are small for both countries since the adding of two factors are near zero. For the aging factors, the linear effect is stronger than the convex effect for both genders in the U.S. since the aging factor 1's for males and females are at higher levels. However, the convex effect is stronger than the linear effect for males in Korea, though the two aging effects for females are similar for both countries.

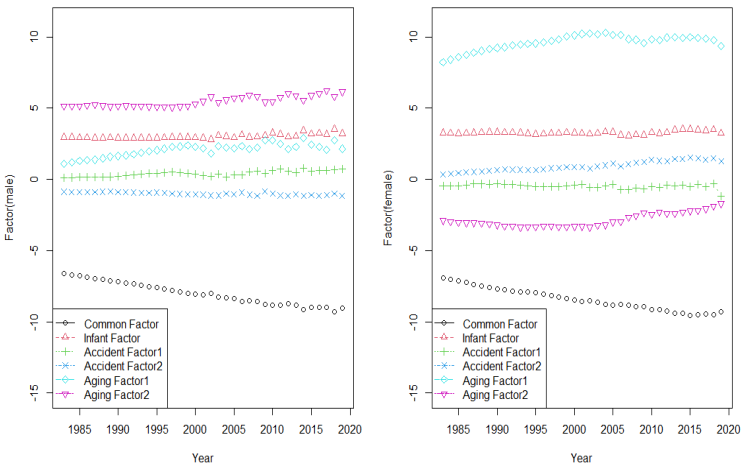

Figure 5: *Trends of 6 factors from yr1983 to yr2019 in Korea: left-male, right-female.*

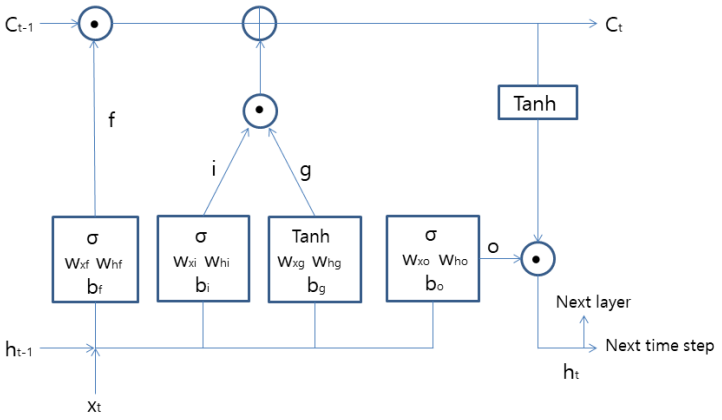

Figure 6: *Structure of LSTM.*

3. Deep-learning method: Long short-term memory and regularized long short-term memory

3.1. Long short-term memory (LSTM)

Long short-term memory (LSTM) is one of the deep learning algorithm methods appropriate for time-series forecasting. LSTM can learn long-term dependencies by learning sequences of observations so that it can forecast a future output based on past features of the data. Raschka and Mirjalili (2019) has shown the structure of LSTM in Figure 6. C_t is a cell state at time t, x_t is input at time t, and h_t is a hidden layer at time t. $\odot$ is element-wise multiplication and $\oplus$ is element-wise addition. An arrow

shows current data streams. 4 boxes contain an activation function (σ) or a hyperbolic tangent (tanh), and weights (e.g. w_{xf}) and bias (e.g. b_f). Each box carries out matrix-vector multiplication with data and linear addition. Gate is a unit computed with an activation function leading to output through $\odot$. LSTM has 3 kinds of gates, forget gate, input gate, and output gate. Forget gate (f) controls how much data to abandon by using the activation function. The activation function has values between 0 and 1 for sigmoid and values ≥ 0 for ReLu. For the sigmoid function, the value near 1 means the gate contains most of the past information, and the value near 0 means it abandons most of them. The ReLu function is similar to the sigmoid but it has a faster learning time. This gate performs element-wise multiplications between the results from the activation function and the past data from a memory cell to determine the level of maintaining past data. Input gate (i) controls the level of maintaining previous output data, and input node (g) updates the cell state with the tanh function to produce new data and perform element-wise multiplications with data from the input gate. The tanh function transfers output results between -1 and 1. Output gate (o) controls the level of reflecting a hidden layer at time t (h_t) by performing element-wise multiplications with data from the memory cell through the tanh function. The mathematical expression of the LSTM is as follows.

$$
\begin{aligned}
f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \\
i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \\
g_t &= \tanh\left(W_{xg}x_t + W_{hg}h_{t-1} + b_g\right) \\
C_t &= (C_{t-1} \odot f_t) \oplus (i_t \odot g_t) \\
o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \\
h_t &= o_t \odot \tanh\left(c_t\right)
\end{aligned}
$$

where f_t, i_t, o_t and C_t are the same size as the hidden layer vector h. For more explanation, refer to Hochreiter and Schmidhuber (1997) and Graves *et al.* (2013). The readers who would like to know more about this subject may also refer to Gers *et al.* (2000), Jozefowicz *et al.* (2015), Azuma (2020), and Chung (2014).

3.2. Regularized long short-term memory

One common problem for fitting models is overfitting. An overfitted model reflects many details and noises of the data, so its forecasting performance is reduced. To improve the model's performance, regularization is adopted to the model (Merity *et al.*, 2017).

There are several regularization methods as L1 and L2, and dropout. L1 and L2 regularizations are the methods to reduce or delete weights, and dropout is a technique that randomly drops units from the neural network (Srivastava *et al.* 2014). These regularizations reduce some units and their connections in LSTM so they prevent overfitting. Dropout might delete the long-term memory connections in LSTM due to its randomness so it may harm forecasting performance if the deleted connections contain necessary information. Both L1 and L2 regularizations can reduce overfitting, but L1 regularization can make the model simpler than L2 regularization because the weights in L1 regularization reduces toward zero's and could be exactly zero's as a regularizing parameter increases, but the weights in L2 regularization could not be exactly zero's (Hastie *et al.* 2009). Since L1 regularization makes the model simpler, L1 regularization is applied to input connections on each LSTM unit by

Table 1: The periods for model input and output for 5 tests

Country	Test	Input periods	Output periods
U.S.	test1	1933~2017	2018
	test2	1933~ 2016	2017~2018
	test3	1933~ 2015	2016~2018
	test4	1933~ 2014	2015~2018
	test5	1933~ 2013	2014~2018
Kor	test1	1983~2018	2019
	test2	1983~ 2017	2018~2019
	test3	1983~ 2016	2017~2019
	test4	1983~ 2015	2016~2019
	test5	1983~ 2014	2015~2019

using Keras in Python. L1 regularization can be expressed as in equation (3.1).

$$(b^*, w^*) = \arg\min_{(b,w)} \sum_j \left(t(x_j) - \left(b + \sum_i w_i h_i \left(x_j \right) \right) \right)^2 + \lambda \left(|b| + \sum_i |w_i| \right) \tag{3.1}$$

where b is a bias, w is a weight, x_j is j^{th} input, h_i is i^{th} hidden layer, $t(x_j)$ is a target function, and λ is a regularization parameter whose value determines the optimization of the results. As λ moves toward zero, the regularization effect is reduced and vice versa.

4. Numerical analysis

4.1. Data and procedure of accuracy test

The U.S. life-tables (human mortality database) from 1933 to 2018 and the Korea life-tables (Statistics of Korea) from 1983 to 2019 are used for the accuracy tests. The Korea life tables before 1983 were not used because the data was not relatively reliable.

The accuracy tests are performed for ages 0 to 109 and 0 to 99 for the U.S. and Korea, respectively. Both mortality rates of the maximum ages of 110(U.S.) and 100(Korea) are fixed at 1, so they are deleted from the life tables. The tests are carried out for 5 different forecasting periods from 1 year to 5 years as in Table 1. Five different forecasting periods are tried for the tests to avoid the coincidence of the test results (i.e. lucky small errors or unlucky large errors). For example, to test for forecasting 5-year log central death rates for the U.S., central death rates from 1933 to 2013 are used to fit the model, and forecasting results are produced for 2014~2018.

After the forecasted log central death rates are produced, the tests are performed by measuring the root mean square error (RMSE) as in equation (4.1).

$$\text{RMSE} = \sqrt{\frac{\sum_{x=0}^{\omega} (Y_x^m - Y_x^r)^2}{\omega + 1}} \tag{4.1}$$

where Y_x^m is a forecasted value of the log central death rate at age x, Y_x^r is a real value of the log central death rate at age x, and ω is the limiting age. Since the future central death rates are not known at some point of the time t_0 in the past, it is assumed to be a current point, and the forecast is carried out based on the assumed current point of the time t_0. After RMSE is obtained by computing the differences between the forecasting results and the real log central death rates for each of the output periods, the average RMSE is finally obtained by averaging the RMSEs for the whole output periods.

4.2. Forecasting process

The process of forecasting 6-PFM is as follows:

- Step 1: fit the model to the latest-year mortality data from the life-tables to determine the parameters as in equation (4.2)

$$(\beta_{0T},\beta_{1T},\beta_{2T},\beta_{3T},\beta_{4T},\beta_{5T},\lambda_1,\lambda_2,\lambda_3,\lambda_4,k_1,k_2) = \arg\min \sum_{x=0}^{\omega} \left(\ln(m_{x,T}) - \ln\left(\hat{m}_{x,T}\right)\right)^2 \tag{4.2}$$

  where $\ln(m_{x,T})$ is a real log central death rate, $\ln(\hat{m}_{x,T})$ is an estimated log central death rate, x is age, ω is the limiting age, and T is the latest year available.

- Step 2: estimate 6 factors year by year using the fixed parameters $(\lambda_1,\lambda_2,\lambda_3,\lambda_4,k_1,k_2)$ estimated in year T in Step 1. It is expressed in equation (4.3).

$$(\beta_{0t},\beta_{1t},\beta_{2t},\beta_{3t},\beta_{4t},\beta_{5t}) = \arg\min \sum_{x=0}^{\omega} \left(\ln(m_{x,t}) - \ln\left(\hat{m}_{x,t}\right)\right)^2 \tag{4.3}$$

  where $t\ (\leq T)$ is year.

- Step 3: forecast 6 factors by applying the vector autoregressive model (VAR) to capture the relationship between multiple quantities as it changes over time, LSTM, or regularized LSTM.

- Step 4: compute the log central death rates by entering the 6 factors forecasted in Step 3 and the parameters estimated in year T in Step 1 into the model.

The accuracies are compared among the 4 models: the LC model, 4-PFM, 5-PFM, and 6-PFM. The 3 forecasting methods of 6-PFM are applied. Thus a total of 6 models are compared: the LC model, 4-PFM with VAR, 5-PFM with VAR, 6-PFM with VAR, 6-PFM with LSTM, and 6-PFM with the regularized LSTM. Forecasting process is excluded and the results are only shown for the LC model, 4-PFM, and 5-PFM. For more details about the LC model and 4-PFM, refer to Lee and Carter (1992) and Haldrup and Rosenskjold (2019).

Forecasting requires time-series analysis. For 6-PFM, 3 different analyses, the VAR, LSTM, and the regularized LSTM are carried out. To forecast from the VAR, the Johansen test, which uses a function ca.jo() in R package is performed for co-integration analysis. For example, the co-integration analysis results for test1 are in Table 2. For the U.S., there are 3 co-integration effects for males and 2 co-integration effects for females because the test value of 45.31, when r (the number of co-integration) ≤ 2 is rejected, but the value of 23.44, when $r \leq 3$ is not rejected at the level of significance of 10% for male, and the value of 68.05, when $r \leq 1$ is rejected but the value of 35.72, when $r \leq 2$ is not rejected at the same level of significance for female. Similarly, we can determine that there are 3 co-integration effects for both males and females in Korea. Refer to Pfaff (2008) for more explanation.

The co-integration results for the 5 accuracy tests are in Table 3. The number in Table 3 is the number of co-integrations. Male in the U.S. and both genders in Korea have 2~3 co-integrations and female in the U.S. has 1~2 co-integrations. Since co-integrations exist for all the tests, the vector error correction model (VECM) should be used and the function vec2var() in R package is used for each test to transform VECM which is the object of the formal class generated by the function ca.jo() into the VAR representation. In Section 4.3 this model is expressed as '6-PFM' instead of '6-PFM

Table 2: Results of Johansen test of 6-PFM for test 1

# co-integration (r)	Test (US)		Test (Kor)		Significance level		
	Male	Female	Male	Female	10%	5%	1%
$r \leq 5$	0.08	0.03	4.95	1.45	6.50	8.18	11.65
$r \leq 4$	7.23	2.31	12.90	6.82	15.66	17.95	23.52
$r \leq 3$	23.44	15.09	22.00	23.45	28.71	31.52	37.22
$r \leq 2$	45.31	35.72	54.11	46.20	45.23	48.28	55.43
$r \leq 1$	89.29	68.05	101.97	79.48	66.49	70.60	78.87
$r = 0$	148.64	103.46	157.62	160.04	85.18	90.39	104.20

Table 3: The co-integration test results of 6-PFM for all 5 accuracy tests

Contry	Gender	test 1	test 2	test 3	test 4	test 5
U.S.	Male	3	2	3	3	3
	Female	2	1	1	1	1
Korea	Male	3	2	2	2	3
	Female	3	3	3	2	2

with VAR'. In the same manner, 4-PFM and 5-PFM, which use VAR, are expressed as '4-PFM' and '5-PFM', respectively.

For LSTM, an algorithm should be set to learn the mapping function from the input (training dataset) to the output (test dataset). For example, a time series data $1, 2, 3, 4, 5, \ldots$, can be transformed into two separate sets X and Y.

$$\begin{aligned} X &: (1,2,3,4), (2,3,4,5), (3,4,5,6), \ldots, \\ Y &: \quad (5,6), \qquad (6,7), \qquad (7,8), \ldots, \end{aligned}$$

where X is a set of input components, and Y is a set of output components. An algorithm learns the output pattern Y from the input pattern X. This is called a sliding window transformation as it is just like sliding a window across the prior observations that are used as inputs to the model in order to forecast the next value in the series. In this case, the window width is 4 time steps.

To forecast the central death rates from LSTM, the time dependence removal of the data is needed by subtracting the previous observations from the current one. Without differentiating , the forecast results may be too sensitively changed to the number of epochs (the number of passes of the entire training dataset the machine learning algorithm completes) or to the values of the regularizing parameter. The differentiating can reduce the instability of the time series data, which leads to stable results. After differentiating, the log central death rates are transformed into a supervised learning format (sliding window transformation) so that the model can learn to fit the target values. The transformed data is inverted into the original scale after the forecast is completed and then it computes RMSE to measure the accuracy. There are 6 factors to be forecasted, and the difference is taken between consecutive observations (lag 1 difference) for each factor. Then, lag 1 differences are transformed into the supervised learning format. We use the window width of 7-time steps in the training data set.

As an example for test 2, the time series of the 6 factors are transformed into two separate sets,

training set (X) and test set (Y) as follows,

$$X: \begin{pmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,7} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,7} \\ d_{3,1} & d_{3,2} & \cdots & d_{3,7} \\ d_{4,1} & d_{4,2} & \cdots & d_{4,7} \\ d_{5,1} & d_{5,2} & \cdots & d_{5,7} \\ d_{6,1} & d_{6,2} & \cdots & d_{6,7} \end{pmatrix}, \begin{pmatrix} d_{1,2} & d_{1,3} & \cdots & d_{1,8} \\ d_{2,2} & d_{2,3} & \cdots & d_{2,8} \\ d_{3,2} & d_{3,3} & \cdots & d_{3,8} \\ d_{4,2} & d_{4,3} & \cdots & d_{4,8} \\ d_{5,2} & d_{5,3} & \cdots & d_{5,8} \\ d_{6,2} & d_{6,3} & \cdots & d_{6,8} \end{pmatrix}, \begin{pmatrix} d_{1,3} & d_{1,4} & \cdots & d_{1,9} \\ d_{2,3} & d_{2,4} & \cdots & d_{2,9} \\ d_{3,3} & d_{3,4} & \cdots & d_{3,9} \\ d_{4,3} & d_{4,4} & \cdots & d_{4,9} \\ d_{5,3} & d_{5,4} & \cdots & d_{5,9} \\ d_{6,3} & d_{6,4} & \cdots & d_{6,9} \end{pmatrix}, \ldots,$$

$$Y: \begin{pmatrix} d_{1,8} & d_{1,9} \\ d_{2,8} & d_{2,9} \\ d_{3,8} & d_{3,9} \\ d_{4,8} & d_{4,9} \\ d_{5,8} & d_{5,9} \\ d_{6,8} & d_{6,9} \end{pmatrix}, \begin{pmatrix} d_{1,9} & d_{1,10} \\ d_{2,9} & d_{2,10} \\ d_{3,9} & d_{3,10} \\ d_{4,9} & d_{4,10} \\ d_{5,9} & d_{5,10} \\ d_{6,9} & d_{6,10} \end{pmatrix}, \begin{pmatrix} d_{1,10} & d_{1,11} \\ d_{2,10} & d_{2,11} \\ d_{3,10} & d_{3,11} \\ d_{4,10} & d_{4,11} \\ d_{5,10} & d_{5,11} \\ d_{6,10} & d_{6,11} \end{pmatrix}, \quad \ldots,$$

where $d_{i,j}$ is a value which subtracts factor $_{i,j-1}$ from factor $_{i,j}$, i is the index of the i^{th} factor, and j is the year.

The number of epochs is set to be 10,000 and the input unit is set to be 42 (=6 × 7) for all the tests, and the output unit is set to be 6 (=6 × 1) for test1, 12 (=6 × 2) for test 2, ..., 30 (=6 × 5) for test 5. For the regularized LSTM, the regularizing parameter λ of 0.05 is used. In Section 4.3 these two models are expressed as '6-PFM_LSTM' and '6-PFM_reg-LSTM', respectively. A main part of the python code is in the Appendix.

4.3. Test results

The results of the accuracy tests are in Table 4. The upper two parts are the average RMSEs of the U. S. male and female groups and the lower two parts are the average RMSEs of the Korean male and female groups.

For males in the U.S., 6-PFM performs best with the average RMSE of 0.0847, followed by 6-PFM_reg-LSTM, 6-PFM_LSTM, the LC model, 5-PFM, and 4-PFM . For females, 6-PFM_reg-LSTM shows the best performance, followed by 6-PFM_LSTM, 6-PFM, 5-PFM, 4-PFM, and the LC model, in this order. In a comparison between males and females, performances for males are better for the LC model and 6-PFM, and performances for females are better for 4-PFM, 5-PFM, 6-PFM_LSTM, and 6-PFM_reg-LSTM. Differences of average RMSEs between genders are 0.014, 0.076, 0.068, 0.004, 0.007, and 0.006 for the LC model, 4-PFM, 5-PFM, 6-PFM, 6-PFM_LSTM, and 6-PFM_reg-LSTM, respectively. Thus the 6-PFMs show more stable results than the LC model, 4-PFM, and 5-PFM.

For males in Korea, 6-PFM_reg-LSTM performs best with the average RMSE of 0.1026, followed by 6-PFM_LSTM, the LC model, 6-PFM, 5-PFM, and 4-PFM, in this order. For females, 6-PFM_reg-LSTM is the best model as in males. 6-PFM is the next best, and 4-PFM shows the worst performance. All the models work better for males than for females. Differences in the average RMSEs between genders are 0.061, 0.064, 0.063, 0.040, 0.046, 0.041 for the LC model, 4-PFM, 5-PFM, 6-PFM, 6-PFM_LSTM, and 6-PFM_reg-LSTM, respectively. Thus, 6-PFMs show more stable results than the other models as in the U.S.

For both countries, 6-PFMs performs better than the LC model, 4-PFM, and 5-PFM. Among the three 6-PFMs, 6-PFM_reg-LSTM is the best for both countries except for the U.S. male. When comparing two countries, there is no consistency for the LC model, 4-PFM, and 5-PFM, but there is

Table 4: The results of accuracy tests

Country	Gender	Forecasting period	LC	4-PFM	5-PFM	6-PFM	6-PFM_LSTM	6-PFM_reg-LSTM
U.S.	Male	1 year	0.1487	0.1764	0.1511	0.0703	0.0778	0.0652
		2 years	0.1530	0.1817	0.1554	0.0829	0.0733	0.0707
		3 years	0.1584	0.1921	0.1859	0.0800	0.0854	0.0929
		4 years	0.1646	0.2010	0.1818	0.0857	0.1022	0.1070
		5 years	0.1675	0.2100	0.1942	0.1049	0.1179	0.1122
		Average	0.1585	0.1922	0.1737	0.0847	0.0913	0.0896
	Female	1 year	0.1654	0.1006	0.0899	0.0654	0.0656	0.0612
		2 years	0.1689	0.1069	0.0918	0.0693	0.0630	0.0640
		3 years	0.1710	0.1120	0.0936	0.0747	0.0886	0.0874
		4 years	0.1810	0.1230	0.1223	0.1079	0.1065	0.1058
		5 years	0.1773	0.1367	0.1324	0.1260	0.0956	0.1005
		Average	0.1727	0.1159	0.1060	0.0887	0.0839	0.0838
Rep. of Kor	Male	1 year	0.1070	0.1095	0.1028	0.0966	0.0993	0.0983
		2 years	0.1003	0.1133	0.1028	0.0894	0.0929	0.0918
		3 years	0.1100	0.1173	0.1141	0.1000	0.0997	0.0963
		4 years	0.1143	0.1219	0.1218	0.1218	0.1039	0.1018
		5 years	0.1250	0.1625	0.1727	0.1506	0.1332	0.1248
		Average	0.1113	0.1249	0.1228	0.1116	0.1058	0.1026
	Female	1 year	0.1795	0.1923	0.1858	0.1620	0.1892	0.1593
		2 years	0.1776	0.1822	0.1792	0.1520	0.1674	0.1611
		3 years	0.1671	0.1900	0.1865	0.1546	0.1406	0.1373
		4 years	0.1742	0.1998	0.1960	0.1429	0.1370	0.1365
		5 years	0.1651	0.1775	0.1811	0.1473	0.1262	0.1246
		Average	0.1727	0.1884	0.1857	0.1518	0.1521	0.1438

Table 5: The projected life expectancies

Country	Gender	Year	LC model	6-PFM_reg-LSTM
U.S.	Male	2030	77.35	76.51
		2040	78.61	76.55
	Female	2030	80.78	81.59
		2040	80.88	81.63
Kor	Male	2030	82.61	80.67
		2040	84.95	81.04
	Female	2030	85.73	86.81
		2040	86.28	87.37

a consistency for the three 6-PFMs. The 6-PFMs work better for the U.S. life-tables than the Korea life-tables.

One abnormal phenomena is found in the Korean female group. The average RMSEs in the Korean female group show decreasing trends as the forecasting period increases, which is not normal because longer-term forecasting should be more difficult than short-term forecasting. Maybe coincidence of the test results has occurred, as I mentioned in Section 4.1, in which unlucky large errors have occurred in the short-term forecasting and lucky small errors have occurred in the longer-term forecasting in the Korean female group.

4.4. Mortality forecasts

The accuracy test results show that 6-PFM_reg-LSTM was the most accurate model for both the United States and Korea. Thus, mortality forecasts were performed for 6-PFM_reg-LSTM and the

results were compared with the LC model. The future years to which the models are forecasted are 2030 and 2040. The future years are restricted to 2040 due to the short periods of the past mortality data in Korea. The number of epochs is set to be 10,000 and the regularizing parameter λ is set to be 0.05 as in the accuracy tests. The input unit is set to be 90 (=6 × 15).

The results are shown in terms of life expectancy as in Table 5. The future life expectancies of Korea are 4.16~6.34 years longer than those of the United States. When the two models are compared, 6-PFM_reg-LSTM forecasts shorter life expectancies for males than the LC model but forecasts those of females longer than the LC model for both countries. Specifically, for the United States, the 6-PFM_reg-LSTM shows that the future life expectancies in 2030 would be 76.51 and 81.59 for males and females, respectively, which are shorter life expectancy results than those of the LC model. The model also shows that the life expectancies in 2040 would be 76.55 and 81.63 for males and females, respectively, which are shorter life expectancy results than those of the LC model for males, but longer than the LC model for females.

It is noted that the life expectancies of the U.S. in 2040 are slightly improved from those in 2030 compared with Korea. For Korea, the 6-PFM_reg-LSTM shows that the life expectancies for males in 2030 would be 80.67, which is a shorter life expectancy result than the results of the LC model, but for females, the life expectancy would be 86.81, which is a longer life expectancy result than the results of the LC model. It also shows that the life expectancy for a male in 2040 would be 81.04, which is a quite shorter life expectancy result than the results of the LC model, but for a female, the life expectancy would be 87.37, which is a longer life expectancy result than the results of the LC model. Therefore, the future life expectancies might be shorter for males but longer for females than we expect since the standard model used for forecasting in both countries is the LC model.

5. Conclusion

6-PFM was developed in an effort to increase the accuracy of the mortality forecast by adding two factors to 4-PFM. With the added two factors, 6-PFM has shown better performance than 4-PFM and 5-PFM as expected. It was also shown that 6-PFM performed better than the LC model in most of the accuracy tests. Among the 3 forecasting methods of 6-PFM, regularized LSTM was shown to have the best performance for both countries except for the U.S. male group. Therefore LSTM is strongly recommended to be adapted when developing forecasting models.

In this paper, the fixed parameter of 0.05 for regularized LSTM was used, but how to determine the appropriate value of the parameter was not developed. The model could perform better if it could be fine-tuned by regularization. Developing fine-tuning methods is proposed for future research.

Appendix: Main part of Python code for LSTM

```
. model = Sequential()

. model.add (LSTM(n_periods_in*n_features, activation='relu', input_shape=
  (n_periods_in, n_features),x activity_regularizer=tf.keras.regularizers.
  L1(l1=\lambda)))

. model.add (RepeatVector(n_periods_out))

. model.add (LSTM(n_periods_out*n_features, activation='relu',
  return_sequences=True))
```

```
. model.add (TimeDistributed(Dense(n_features)))

. model.compile (optimizer='adam', loss='mse')

. hist = model.fit(X, Y, epochs=10,000, verbose=0)

. yhat = model.predict(x_input, verbose=0)
```

where $n\_periods\_in$ is 7, $n\_periods\_out$ is the number of forecasting periods, and $n\_features$ is the number of factors which is 6. The regularizing parameter λ is set to be $l1$=0.0 and $l1$=0.05 for LSTM and reg_LSTM, respectively, inside of the () in $tf.keras.regularizers.L1()$.

References

Azuma Y (2020). Core deep-learning introduction, *SB Creative Corp*.

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

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

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

Cairns A, Blake D, and Dowd K (2006). A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration, *The Journal of Risk and Insurance*, **73**, 687–718.

Choi J (2021). Comparison of accuracy between LC model and 4-parametric factor model when COVID-19 impacts mortality structure, *Communications for Statistical Applications and Methods*, **28**, 233–250.

Chung J, Culcehre C, Cho K, and Bengio Y (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.

Diebold FX and Li C (2006). Forecasting the term structure of government bond yields, *Journal of Econometrics*, **130**, 337–364.

Gers F, Schmidhuber J, and Cummins F (2000). Learning to forget: continual prediction with LSTM, *Neural Computation*, **12**, 2451–2471.

Gompertz B (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies, *Philosophical Transactions of the Royal Society of London*, **115**, 513–583.

Graves A, Mohamed A-r, and Hinton G (2013). Speech recognition with deep recurrent neural networks, *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Haldrup N and Rosenskjold C (2019). *Econometrics*,

Heligman L and Pollard JH (1980). The age pattern of mortality, *Journal of the Institute of Actuaries*, **107**, 49–80.

Hastie T, Tibshirani R, and Friedman J (2009). The elements of statistical learning: data mining, inference, and prediction, *Springer Series in Statistics*.

Hochreiter, Sepp and Schmidhuber, Jürgen (1997), long short-term memory, *Neural Computation*, **9**, 1735–1780.

Human mortality data, Retrieved from Dec 31, 2020, https://www.mortality.org/

Hyndman RJ and Ullah MDS (2007). Robust forecasting of mortality and fertility rates: a functional data approach, *Computational Statistics and Data Analysis*, **51**, 4942–4956.

Jozefowicz R, Zaremba W and Sutskever L (2015). An empirical exploration of recurrent network architectures. In *Proceedings of ICML*, 2342–2350.

Lee RD and Carter LR (1992). Modeling and forecasting U.S. mortality, *Journal of the American Statistical Association*, **87**, 659–671.

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

Merity S, Keskar NS, and Socher R (2017). Regularizing and optimizing LSTM language models.

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

Perla F, Richman R, Scognamiglio S, and Wuthrich M (2021). Time-Series Forecasting of Mortality Rates using Deep Learning, *Scandinavian Actuarial Journal*, **2021**, 572–598.

Pfaff B (2008). VAR, SVAR, and SVEC Models: implementation within R Package vars, *Journal of Statistical Software*, **27**, 1–32.

Raschka S and Mirjalili V (2017). Python machine learning(2nd ed.), *Packt Publishing*.

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

Richman R and W*ü*thrich MV (2019). Lee and Carter go machine learning: recurrent neural networks.

Rogers A, and Little JS (1994). Parameterizing age patterns of demographic rates with the multiexponential model schedule, *Mathematical Population Studies*, **4**, 175–195.

Siler W (1979). A competing-risk model for animal mortality, *Ecology*, **60**, 750–757.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, and Salakhutdinov R (2014). Dropout: a simple way to prevent neural networks from overfitting, *Journal of Machine Learning Rresearch*, **15**, 1929–1958.

Statistics Korea. Life-table, Retrieved from Dec 31, 2020, https://kosis.kr/statisticsList/statisticsList Index.do? vwcd=MT_ZTITLE&menuId=M_01_01#content-group

United Nations (2019), World Population Prospects 2019, Retrieved from Apr 01, 2021, http:// population.un.org/wpp

Wiśniowski A, Smith PWF, Bijak J, Raymer J, and Forster JJ (2015), Bayesian population forecasting: extending the Lee Carter method, *Demography*, **52**, 1035–1059.

Received April 15, 2021; Revised June 28, 2021; Accepted August 27, 2021