Malaysian Journal of Mathematical Sciences 13(2): 251–264 (2019)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Stochastic Mortality Model in a State-Space Framework

Mohd Nor, S. R.[1], Yusof, F. *[1], and Bahar, A.[1,2]

[1] *Universiti Teknologi Malaysia, Department of Mathematical Sciences, Universiti Teknologi Malaysia, Skudai, Johor*
[2] *UTM-Centre for Industrial and Applied Mathematics (UTM-CIAM), Universiti Teknologi Malaysia, Skudai, Johor*

E-mail: fadhilahy@utm.my
* *Corresponding author*

Received: 10 November 2018
Accepted: 27 April 2019

ABSTRACT

The incorporation of the time-varying parameter in the mortality model has become one the main contributions in the actuarial field since it allows for the stochastic nature of the mortality rates. However, it has also become a growing concern among the researchers since the residuals of the proposed model are evaluated independently. In this study, we extended the existing leading independent stochastic mortality model which is the O′Hare mortality model into the state-space representation of the O′Hare mortality model. The parameters of the extended model are estimated using the Expectation-Maximization algorithm of the maximum likelihood estimation method. Using the Malaysian mortality data, we have found that our proposed model significantly improves the accuracy of the in-sample fitting and the seven-year out-sample forecast as compared to the existing model considered.

Keywords: State-Space, Stochastic mortality model, O′Hare model and Forecast

1. Introduction

A model proposed by Lee and Carter (1992) brought a new impetus to the field of stochastic mortality model by developing the standard statistical time series method into a two factor model. Since then, numerous amounts of literature have been proposed on the modelling of mortality risk including, O′Hare and Li (2012), Plat (2009), Cairns et al. (2009), Renshaw and Haberman (2006), and Tuljapurkar (2008). However, a common feature shared among these approaches is that their errors are estimated independently resulting into too erroneously smaller prediction interval. Therefore, to address the aforementioned problem, several authors such as Husin et al. (2015), Fung et al. (2015), Pedroza (2006), and Lazar and Denuit (2009) introduced the reformulation of Lee and Carter (LC) model into a State-Space model with the aim to estimate the parameters simultaneously.

The single factor in the LC model is limited by the fact that it assumes perfect correlation across ages. This limitation is against the characteristics of the mortality rates which are varied across all age groups. For instance, individuals in the young age group tend to have more volatile mortality patterns as compared to the individuals in the old age group. This might be due to the risky behaviours that young people are engaged in such as highway fatalities, smoking, unhealthy food consumption and physical activity. Ignoring these variances for all ages would lead to inaccurate forecast reading. Hence, to overcome this issue, O′Hare and Li (2012) proposed the O′Hare mortality model which incorporates the quadratic age-effect parameter into the modelling procedure. But, similar to LC, the historical and forecasting performances of O′Hare and Li (2012) mortality model are estimated separately.

Therefore, motivated by the concept of State-Space, this study aims to extend the stochastic mortality model proposed by O′Hare and Li (2012) into a unified estimation. The parameters of our proposed model are estimated using maximum likelihood estimation (MLE) via the expectation-maximization (EM) algorithm. Using Malaysia mortality data for male and female, we evaluated and compared our results with the independent mortality model developed by O′Hare and Li (2012). The structure of the paper is organised as follows: Section 2 describes the models′ forecasting in the State-Space structure. Section 3 summarised the fit performances for both of the methods considered. Finally, Section 4 provides a brief conclusion remark.

2. Methodology

2.1 Data Description

In this study, the stochastic mortality models are applied to two sets of population in Malaysia, which are male population and female population. The data was first transformed into a logarithm of mortality rates in order to minimize the existence of any possible high variances occurring among the oldest populations. The data for male and female mortality rates contain 38-time period intervals from 1980 to 2017 with five-year age span, from ages 0 to 80. The mortality data encompasses of three different dimensions which are the year, age, and population dimensions. Hence, there are a total of 38 time periods, 17 age groups and 2 populations that will be accounted into the modelling framework of this study.

Since mortality data has high random variations across ages and periods, it is necessary to have a lot of observed data fitted to the estimated model in order to obtain accurate forecasts measurements Hyndman and Kostenko (2007). Therefore, the observed Malaysia mortality data are set to a minimum of 31 years to be historically fitted to the mortality models and another 7 years of the observed data are used to be compared with the models′ forecasted values.

2.2 O′Hare Mortality Model

In this section, we adopted the stochastic mortality model proposed by O′Hare and Li (2012) into a general state-space structure as in Holmes et al. (2014). The O′Hare and Li (2012) model is presented as in Equation (1)

$$\ln(m_{x,t}) = \alpha_x + k_t^1 + k_t^2(\bar{x} - x) + k_t^3((\bar{x} - x)^+ + [(\bar{x} - x)^+]^2) + \varepsilon_{x,t} \tag{1}$$

where $k_t^j = k_{t-1}^j + \theta^j + u_t^j$ where $j = 1, 2, 3$ since the model in Equation (1) has only three coefficients of period effects which are k_t^1, k_t^2 and k_t^3.

where $m_(x,t)$ is the observed mortality rates of $x = 1, ..., N$ age-groups in a time period of $t = 1, ..., T$. α_x is defined as the age-effects, whereas k_t^j is defined as the period-effects of the mortality rates. $\bar{x}$ is symbolized as the average measurement of the age-groups whereas $(\bar{x} - x)^+$ takes the same form as $(\bar{x} - x)$ when the value is positive, and else is zero. The non-linear features of

$(\bar{x}-x)^{+}+[(\bar{x}-x)^{+}]^2$ is used to capture for the irregular characteristics happened at the early mortality experiences below 20. $\varepsilon_{x,t}$ and u_t^j is the error terms for the mortality model designed above and are assumed to be independent. O′Hare method is fitted using maximum likelihood estimation (MLE) with the death numbers is assumed to follow Poisson distribution with mean $E_{x,t}m_{x,t}$

$$D_{x,t} = Poisson(E_{x,t}m_{x,t}) \tag{2}$$

where is $D_{x,t}$ the number of deaths and $E_{x,t}$ is the number of exposure. The log-likelihood function of (2) in general is:

$$L(\phi; D, E) = \sum_{x,t} D_{x,t} \ln[E_{x,t}m_{x,t}(\phi)] - E_{x,t}m_{x,t}(\phi) - \ln(D_{x,t}!) \tag{3}$$

where ϕ is referred to as the full set of parameters of the model considered. From the general log-likelihood function in Equation (3), the log-likelihood function for the Lee and Carter method is:

$$L(\alpha, \beta, k; D, E) = \textstyle\sum_{x,t} D_{x,t} \ln[E_{x,t}.\exp(\alpha_x + \beta_x k_t)] - E_{x,t}.\exp(\alpha_x + \beta_x k_t)$$

The third term of Equation (2) is ignored since $\ln(D_{x,t}!)$ do not depend on any parameter values and hence will become zero once the log-likelihood function is derived. The period effects of k_t^j are not directly estimated within the mortality model (1) and are estimated separately in the form of the time series model such as multivariate random walk with drift and ARIMA (p,d,q) models. Hence, to avoid lack of efficiency in the statistical procedures, we recasting the mortality model (1) into a state-space design which is formulated as in (4).

2.3 O′Hare State-Space Mortality Model

O′Hare mortality model in Equation (1) is reformulated into unified estimation of state-space framework as indicated in Equation (4) below:

$$\mathbf{m}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{k}_t + \boldsymbol{\varepsilon}_t \tag{4}$$

where; $\mathbf{m}_t = (\ln(\mathbf{m}_{x_1,t,1}), ..., \ln(\mathbf{m}_{x_N,t,1}), ..., \ln(\mathbf{m}_{x_1,t,M}), ..., ln(\mathbf{m}_{x_N,t,M}))^{'}$, $\mathbf{m}_{x_N,t,M} = (\mathbf{D}_{x_N,t,M}/\mathbf{E}_{x_N,t,M})$ is the mortality rates with $\mathbf{D}_{x_N,t,M}$ is the number of death, $\mathbf{E}_{x_N,t,M}$ is the number of population, $x = x_1, ..., x_N$ is the number of age-groups, $t = 1, ..., T$ is the number of years, $i = 1, .., M$ is the number of population, and a is the parameter intercept.

Parameter $\boldsymbol{\beta}$ in Equation (4) is equal to the matrix of age-effect parameters of O′Hare model:

$$\boldsymbol{\beta} = \begin{pmatrix} 1 & \bar{x} - x_1 & (\bar{x} - x_1)^+ + [(\bar{x} - x_1)^+]^2 \\ 1 & \bar{x} - x_2 & (\bar{x} - x_2)^+ + [(\bar{x} - x_1)^+]^2 \\ \vdots & \vdots & \vdots \\ 1 & \bar{x} - x_N & (\bar{x} - x_N)^+ + [(\bar{x} - x_1)^+]^2 \end{pmatrix}$$

Parameter $\boldsymbol{k}_t$ is denoted as: $k_t = (k_t^1, k_t^2, k_t^3)'$ where $\boldsymbol{k_t = Dk_{t-1} + u_t}$, with $\boldsymbol{u}_t$ ~MVN(0,$\boldsymbol{Q}$)

$$\text{where } \boldsymbol{D} = \begin{bmatrix} 1 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & & \cdots & 0 & 1 & 0 \\ 0 & & \cdots & 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{\varepsilon}_t = (\varepsilon_{x_1,t}, \varepsilon_{x_2,t}, ..., \varepsilon_{x_N,t})', \text{ where } \boldsymbol{\varepsilon}_t \sim \text{MVN}(0,\boldsymbol{R})$$

$\boldsymbol{\varepsilon}_t$ is assumed to be independent and identically distributed with diagonal and equal variance vector, $\boldsymbol{R}$, whereas $\boldsymbol{u}_t$ is uncorrelated to $\boldsymbol{\varepsilon}_t$, and is also assumed to be independent and identically distributed with a diagonal constant variance with mean zero, $\boldsymbol{Q}$. The estimation procedures has been estimated in a unified way using EM algorithm as in Holmes et al. (2014) steps.

2.4 Mortality Forecasts

There are two mortality models in this study which are O′Hare and O′Hare in a State-Space framework. O′Hare mortality model consists of two stages of modelling procedure. The first stage is to estimate the parameters value in O′Hare model, whereas the second stage is to forecast the estimated parameters obtained from the first stage. Following similar approach of Wan and Bertschi (2015) and Hyndman and Khandakar (2008), the time-varying indexes, k_t^1, k_t^2, and k_t^3 obtained from the first stage of estimation process are forecasted by using appropriate ARIMA (p,d,q) time series models. The forecasted mortality rates obtain for O′Hare mortality model with T as the final year for the observed mortality rates, and h is the number of year ahead is defined as follow:

$$\ln(m_{x,T+h}) = \alpha_x + k^1_{T+h} + k^2_{T+h}(\bar{x}-x) + k^3_{T+h}((\bar{x}-x)^+ + [(\bar{x}-x)^+]^2) + \varepsilon_{x,T+h}$$

In contrast to O′Hare model, O′Hare State-Space has only single stage of modelling procedure. The forecast for mortality model in a state-space representation is obtained by recursive procedure of Kalman filtering combined with EM-algorithm method (Husin et al. (2015)). The mathematical formulation of the mortality forecast for h years is denoted as follows:

$$\boldsymbol{m}_{T+h|T} = \boldsymbol{\alpha} + \boldsymbol{\beta}^h \boldsymbol{k}_{T|T}$$

where $\boldsymbol{k}_{(T|T)}$ is the final state parameter, T is the final time period for the observed mortality rates, and h is the number of year ahead.

2.5 In-Sample and Out-Sample Measurement Errors

In this study, the models are historically fitted to Malaysia mortality rates from the period of 1980 to 2010, then seven years of mortality rates from 2011 to 2017 are forecasts. Five-year age gap for a total of 17 age groups for both male and female mortality rates datasets have been applied to the models.

Following Li et al. (2015), there are three numerical approaches that we have used to examine the respected models′ performances denoted by AE, MAPE and RMSE. The results recorded are in the percentage form. The formulas of the measurement errors are defined as below:

Average Error (AE)= $\frac{1}{NT}\sum_x \sum_t \frac{\widehat{m}_{xt}-m_{xt}}{m_{xt}} \times 100$

Mean Absolute Percentage Error (MAPE)= $\frac{1}{NT}\sum_x \sum_t \frac{|\widehat{m}_{xt}-m_{xt}|}{m_{xt}} \times 100$

Root Mean Squared Error (RMSE)= $\sqrt{\frac{1}{NT}\sum_x \sum_t \left(\frac{\widehat{m}_{xt}-m_{xt}}{m_{xt}}\right)^2} \times 100$

where $\widehat{m}_{xt}$ is the predicted values of the mortality rates and m_{xt} is the mortality rates observed values.

Other than that, the performances of O′Hare mortality model and O′Hare State-Space mortality models are measured by using Akaike Information Criterion (AIC). O′Hare and Li (2012) stated that AIC compares the fit quality of the models with parsimony balances. The formula for AIC is defined as follows:

$$\text{AIC}= L(\phi) - K$$

where $L(\phi)$ is the log-likelihood of the ϕ, and K is the number of estimated parameters.

3. Results and Discussions

This section compares O′Hare and O′Hare State-Space performances based on the in-sample and out-sample measurement errors results. The measurement errors are evaluated in the percentage form and tabulated in Table 1 and Table 2 as below.

Table 1: In-Sample.

Model	O′Hare		O′Hare State-Space	
	Male	Female	Male	Female
AE(%)	-0.0879	-0.0437	-0.0182	-0.0041
MAPE(%)	0.0732	0.0608	0.0637	0.0508
RMSE(%)	0.8179	0.4367	0.4358	0.3186

Table 2: Out-Sample.

Model	O′Hare		O′Hare State-Space	
	Male	Female	Male	Female
AE(%)	0.1506	0.1130	-0.0415	0.0726
MAPE(%)	0.1916	0.1179	0.1073	0.0980
RMSE(%)	0.3713	0.6049	0.1408	0.6373

According to Table 1 and Table 2, our proposed model which is O′Hare State-Space mortality model outperformed the existing O′Hare model for both male and female in terms of in-sample and out-sample estimates since it has the lowest values of AE, MAPE, and RMSE. The potential hidden errors are the errors estimated from the parameters and the regression mortality model. O′Hare State-Space could adequately capture the mortality experiences in male and female as compared to O′Hare since it could capture the potential hidden errors that are being ignored in the separate estimation procedure Pedroza (2006). O′Hare mortality model has more than one estimated errors as compared to O′Hare State-Space model which has only one estimated error. Hence, O′Hare model tend to produce more errors as compared to O′Hare State-Space. As overall, the lowest values of AE, MAPE and RMSE subsequently implies that the proposed models which is O′Hare State-Space is less biased, less magnitude of deviance and less size of deviance, respectively. On the other hand, the quality of fit of the different mortality models is also evaluated by using AIC for Malaysia dataset male and female. The purpose of using AIC as a model comparison technique is because the model accuracy could be analysed conveniently regardless of the number of parameters incorporated in the model. The model with the smallest AIC imply that it has a good fitting performance as compared to others. The approximate AIC values are provided in Table 3 below.

Table 3: AIC values.

Model	O'Hare	O'Hare State-Space
Male	-5430.162	-6480.048
Female	-4336.588	-5150.068

Similar to result in Table 1, Table 3 shows that O′Hare State-Space model is ranked as the best model as compared to O′Hare since it has the smallest AIC values for males and females. Next, the results in Table 2 are then further analysed by illustrating the 7-year forecast of both methods in Figure 1 below.

The black, red and blue colours are the colours that we used in Figure 1 to represent for the observed mortality rates, O′Hare model and O′Hare State-Space model, respectively. The plots in Figure 1 are consistent with the results that we have obtained in Table 2 since it indicates that the 7-year ahead forecast for method O′Hare State-Space mortality model produce more accurate predictions than the O′Hare mortality model for the given group ages.

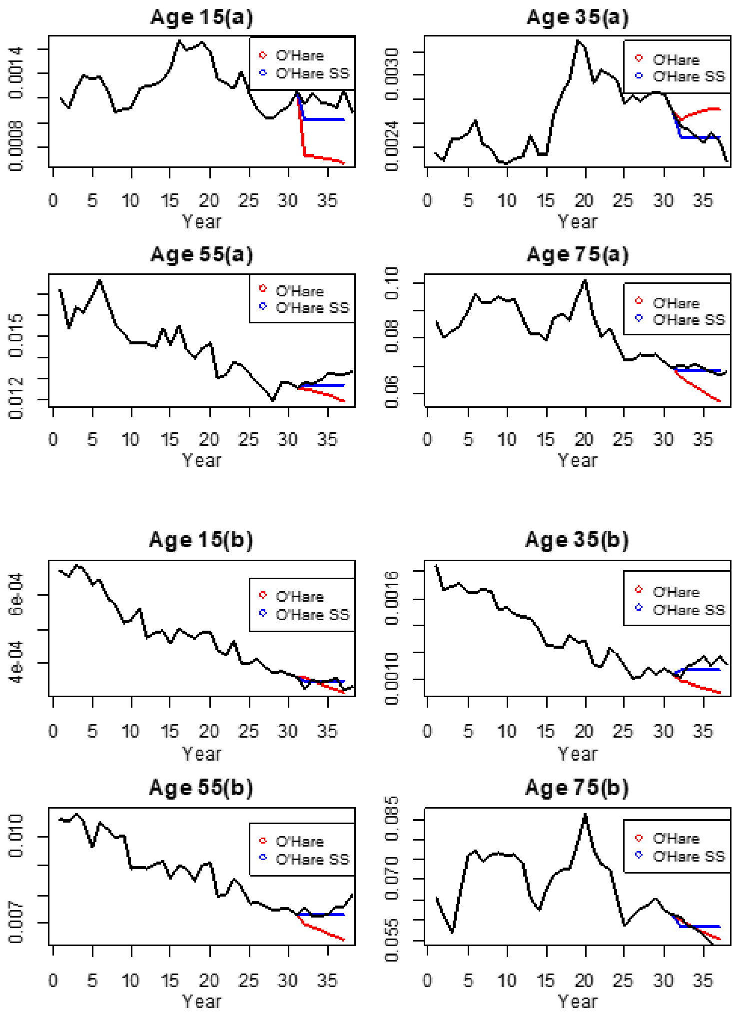

Figure 1: a: 5-years of forecast plots for several selected ages for male; b: 5-years of forecast plots for several selected ages for female.

Figure 2 and Figure 3 illustrates the graphical analysis of standardized residual patterns for each model against the age (first panel) and the fitting period (second panel). The purpose of these plots is to summarize the models′ assumption random errors that need to follow independent and identically distributed normal and to check whether the mortality models could adequately capture the mortality experiences in period effect and age effect.

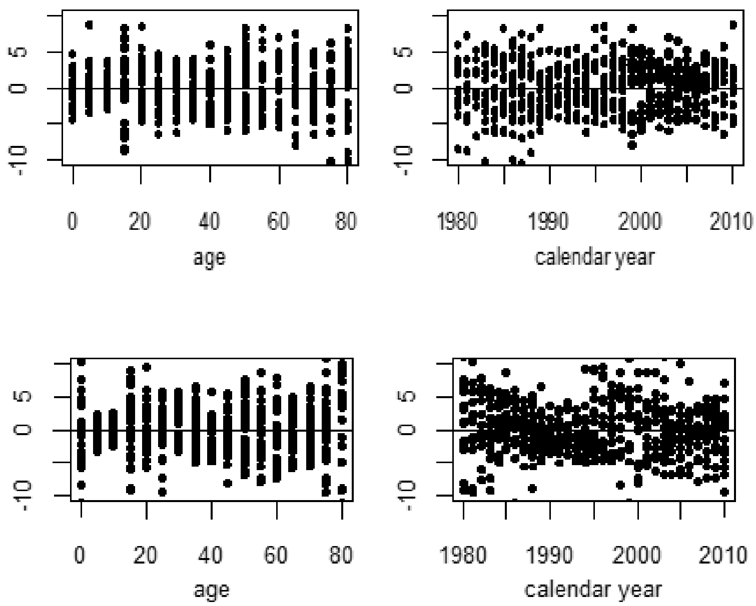

Figure 2: Top Left and Right: Residuals plots for O'Hare male in 2D dimension; Bottom Left and Right: Residuals plots for O'Hare State-Space male in 2D dimension.

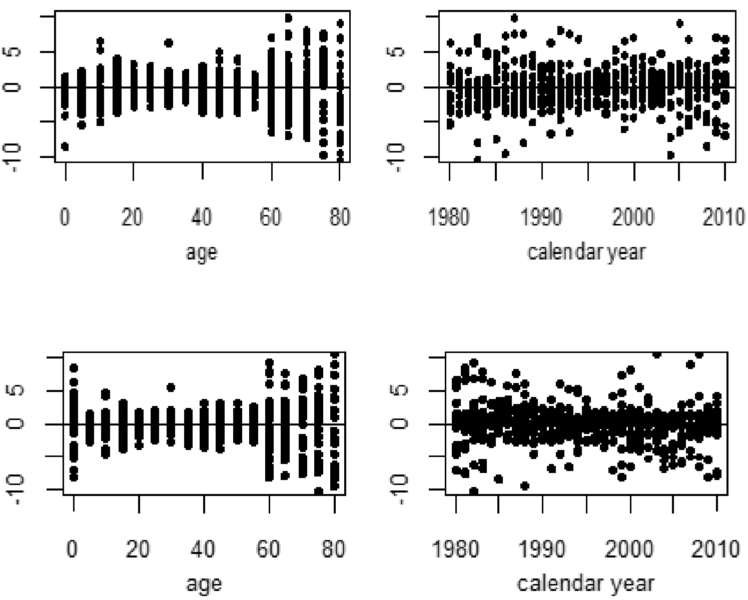

Figure 3: Top Left and Right: Residuals plots for O'Hare female in 2D dimension; Bottom Left and Right: Residuals plots for O'Hare State-Space female in 2D dimension.

As can be seen, although O′Hare State-Space mortality model performed better than O′Hare model in fitting and forecast results, however, the residuals plots of the proposed model for both male and female illustrate the unequal variances in terms of age-effect and calendar-effect dimensions. To validate the graphical assumption, the variance-ratio test and Royston test are applied to O′Hare and O′Hare State-Space model in order to check if the residuals of the respected models follow constant variance and multivariate normal distribution. The results of the tests are tabulated in Table 4 as below:

Table 4: Residual analyses results.

Normality Tests,p-valuess	O'Hare		O'Hare State-Space	
	Male	Female	Male	Female
Variance-ratio	0.064	0.136	0.008	0.381
Royston	0.000	0.000	0.0712	0.0103

Based on Table 4, variance-ratio test p-values indicate that the null hypothesis of constant variance for all stochastic mortality models should not be rejected. However, the results indicate that O′Hare State-Space mortality model for male reject the null hypothesis that the residuals follow equal variance. On the other hand, the p-values of the Royston test suggest that the residuals of O′Hare mortality model do not follow multivariate normal distribution, whereas the residuals of O′Hare State-Space model follow multivariate normal distribution at 1.

Hence, the graphical and numerical analysis in Figure 2, Figure 3 and Table 4 concludes that O′Hare and O′Hare State-Space mortality models does not follow independent and identically distributed normal residuals. This signifies the failure of the respected model to capture full information of the mortality experiences in the model′s structure. However, the results indicate that residuals of O′Hare State-Space model are converging near to zero for both male and female as compared to O′Hare model, in which it agrees that one of the contribution of the reformulated mortality model in a State-Space framework is that it could capture hidden errors that are failed to capture by the mortality model with a separate type of estimation procedure. The non-normal errors could be improved by extending another advanced stochastic mortality model into a State-Space framework. Such improvements could include reformulation of multi-population mortality model into a State-Space framework. Hence, this study concludes that O′Hare State-Space model outperformed O′Hare mortality model for both in-sample and out-sample estimates.

4. Conclusions

In this study, we proposed an alternative method of the stochastic mortality model by integrating the state and space equations of the existing O′Hare and Li (2012) method. Our approach is evaluated and compared using Malaysia mortality data for both male and female starting from the period of 1980 until 2010. The results of the analysis show that the reformulation of the O′Hare and Li (2012) into a State-Space model produced a better fit in terms of the in-sample and the out-sample forecast as compared to the existing stochastic mortality model of O′Hare and Li (2012) The residuals of the proposed model which is O′Hare State-Space model does not follow independent and identically distributed normal. However, the proposed model still brings a significant contribution to the field of mortality modelling since the residuals are converging near to zero.

Acknowledgment

The authors would like to thank the Department of Statistics Malaysia (DOSM) for providing the data and we would like to acknowledge Universiti Teknologi Malaysia (UTM) and Ministry of Education Malaysia for the financial resources under the Skim Latihan Anak Muda (SLAM) scholarship and research GE STEM grant with vote no. 07397.

References

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., and Balevich, I. (2009). A quantitative comparison of stochastic mortality models using data from england and wales and the united states. *North American Actuarial Journal*, **13**:1–35.

Fung, M. C., Peters, G. W., and Shevchenko, P. V. (2015). A state-space estimation of the lee-carter mortality model and implications for annuity pricing. *Proceeding of 21st International Congress on Modelling and Simulation*, pages 952–958.

Holmes, E. E., Ward, E. J., and Scheuerell, M. D. (2014). Analysis of multivariate time-series using the marss package. *NOAA Fisheries, Northwest Fisheries Science Center*, **2725**:98112.

Husin, W. Z. W., Zainol, M. S., and Ramli, N. M. (2015). Performance of the lee-carter state space model in forecasting mortality. *Proceedings of the World Congress on Engineering*, **1**.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for r. *Journal of Statistical Software*, **27**(3):C3.

Hyndman, R. J. and Kostenko, A. V. (2007). Minimum sample size requirements for seasonal forecasting models. *Foresight*, **Spring**(6):12–15.

Lazar, D. and Denuit, M. M. (2009). A multivariate time series approach to projected life tables. *Applied Stochastic Models in Business and Industry*, **25**(6):806–823.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting u. s. mortality. *Journal of the American Statistical Association*, **87**(419):659–671.

Li, H., O′Hare, C., and X, Z. (2015). A semiparametric panel approach to mortality modeling. *Insurance: Mathematics and Economics*, **61**:264–270.

O′Hare, C. and Li, Y. (2012). Explaining young mortality. *Insurance: Mathematics and Economics*, **50**(1):12–25.

Pedroza, C. (2006). A bayesian forecasting model: Predicting u.s. male mortality. *Biostatistics*, **7**(4):530–550.

Plat, R. (2009). On stochastic mortality modeling. *Insurance: Mathematics and Economics*, **45**(3):393–404.

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

Tuljapurkar, S. (2008). Mortality declines, longevity risk and aging. *Asia-Pacific Journal of Risk and Insurance*, **3**(1):37–51.

Wan, C. and Bertschi, L. (2015). Swiss coherent mortality model as a basis for developing longevity de-risking solutions for swiss pension funds: A practical approach. *Insurance: Mathematics and Economics*, **63**:66–75.