Sains Malaysiana 51(7)(2022): 2237-2247
http://doi.org/10.17576/jsm-2022-5107-24

Extending the GLM Framework of the Lee-Carter Model with Random Forest Recursive Feature Elimination Based Determinants of Mortality

(Perluasan Model Kerangka GLM Lee-Carter dengan Faktor Penyebab Kematian Berdasarkan Eliminasi Ciri Rekursif Hutan Rawak)

NURUL AITYQAH YAACOB[1,2] DHARINI PATHMANATHAN[1]* & IBRAHIM MOHAMED[1]

[1]Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Federal Territory, Malaysia
[2]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Negeri Sembilan, Kampus Kuala Pilah, 72000 Kuala Pilah, Negeri Sembilan Darul Khusus, Malaysia

Received: 23 June 2021/Accepted: 1 January 2022

ABSTRACT

The Lee-Carter (LC) model led to the development of many prominent mortality models. This study aims to modify the generalised linear model (GLM) (Poisson, negative binomial, and binomial) framework of the LC model by incorporating factors that affect mortality into the model. The top three factors which affect the mortality for each of the 14 countries studied were selected using the random forest recursive feature elimination (RF-RFE) method which eliminates the least important factors based on the correlation of the predictors with the log-mortality rate. These selected factors were integrated in the form of additional bilinear variates to the GLM models and compared to their original counterparts. The RF-RFE method is effective in selecting the best determinants of mortality by avoiding multicollinearity among predictor variables. The inclusion of the time-factor modulation based on the factors selected improved the model adequacy significantly. Vast improvement was evident in the Poisson and binomial settings. Furthermore, the modified GLM version fits short-base-period data well. This study shows that the inclusion of exogenous determinants of mortality improves the performance of the model significantly.

Keywords: GLM; Lee-Carter model; mortality; random forest; recursive feature elimination

ABSTRAK

Model Lee-Carter (LC) telah membawa kepada perkembangan banyak model mortaliti yang menyerlah. Kajian ini bertujuan mengubah suai kerangka model linear teritlak (GLM) (Poisson, binomial negatif dan binomial) model LC dengan menggabungkan faktor yang mempengaruhi kematian ke dalam model. Tiga faktor teratas yang mempengaruhi kematian bagi setiap 14 negara yang dikaji, dipilih dengan menggunakan kaedah penghapusan ciri rekursif hutan rawak (RF-RFE) yang berfungsi menyingkirkan faktor yang kurang penting berdasarkan korelasi peramal dengan kadar log kematian. Faktor yang dipilih telah diintegrasikan dalam bentuk bilinear tambahan yang bervariasi dengan model GLM dan kajian perbandingan dengan versi model GLM yang asli telah dijalankan. Kaedah RF-RFE berkesan dalam memilih penentu kematian terbaik dan mengelakkan multikolineariti di antara pemboleh ubah peramal. Modulasi faktor masa yang dimasukkan berdasarkan faktor yang dipilih telah meningkatkan kecukupan model dengan lebih bererti. Peningkatan yang besar dapat dibuktikan pada model Poisson dan binomial. Tambahan pula, versi model GLM yang diubah suai turut sesuai digunakan untuk data jangka masa yang pendek. Kajian ini juga mendedahkan bahawa penggunaan penentu kematian luaran telah meningkatkan prestasi model dengan lebih bererti.

Kata kunci: GLM; hutan rawak; kematian; Model Lee-Carter; penghapusan ciri rekursif

INTRODUCTION

Mortality models are important in describing the demographic structure and health status of a population. The main challenges faced by developed countries are longevity risk and increased correlation between closely related populations such as gender, race, states and countries (Nor et al. 2021). The Lee-Carter (LC) (1992) model which was constructed to fit and predict mortality rates for the United States of America has been a benchmark in modelling and forecasting mortality for many countries since its introduction. This model estimates the period trend using a one-factor stochastic model and explains mortality trends in a stochastic framework by fitting past mortality data and modelling the time trend as a stochastic process (Selecka et al. 2017). Some of the popular extensions of the LC model include the Lee-Miller (2001), Booth-Maindonald-Smith (2002), Renshaw-Haberman (2006), and Hyndman-Ullah (2007) variants. These led to various modifications and improvisations. Hyndman and Shang (2009) improved the functional principal component by Hyndman and Ullah (2007) by using weighted principal component regression which assigns higher weights for more recent data. Hansen (2013) generated mortality table time trajectories and compared the performance of five mortality models, including the LC model and its extensions. The aforementioned models do not have a straightforward setting in parameter estimation and incorporating exogenous determinants into these models which is the aim of this study will be a daunting task. To achieve this aim, the generalised linear model (GLM) framework of the LC model will be of good use.

The application of regression methods on the LC model is not straightforward since the model is not in the regression form (Currie 2013) and not easily identifiable. To achieve this, identifiable constraints must be set. The difficulty in estimating the parameters arises from the bilinear term used in the LC model (see equation 1). Thus, the use of constraints imposed on the individual terms forming the bilinear term are required to overcome the identifiability issue. The weighted least squares and maximum likelihood estimation approaches parameterise the model and overcome the homoscedastic error limitation of the LC model (Wilmoth 1993). This led to the Poisson log-bilinear (Brouhns et al. 2002), binomial (Wang & Lu 2005) and the negative binomial (Delwarde et al. 2007) variants. Cairns et al. (2009) compared eight stochastic models explaining improvements in mortality rates in England and Wales and in the USA. Currie (2016) described these models in the standard model terminology of GLM and Gaussian network model (GNM). Pitt et al. (2018) extended the GLM model by using Lagrange methods and P-splines to improve mortality projection.

In this study, it is of interest to examine the effects of exogenous determinants in improving the adequacy of mortality models. This is achievable by incorporating exogenous determinants of mortality such as macroeconomic factors, health determinants and so on as predictor variables in the GLM framework. Furthermore, the parameter estimation process for this setting is straightforward as it uses the MLE method. Hanewald et al. (2011) established a link between macroeconomic fluctuations and the mortality index of the LC model to develop a dynamic asset liability model. Hanewald (2011) studied the impact of macroeconomic fluctuations in the LC model. French (2014) suggested that mortality in different populations may be related based on economic literature on technology and knowledge diffusion. Gross domestic product (GDP), health spending as well as lifestyles factors such as alcohol, cigarette, fruit, vegetable, and fat consumption are good factors to be considered in studying mortality rates (French & Ohare 2014). Rasoulinezhad et al. (2020) discovered that mortality could be accounted for by variations in the concentration of carbon dioxide emissions. Poor mental health (Yeh et al. 2019) and diabetes (Chen et al. 2020) also contributed to higher mortality rates. Tulu et al. (2020) used negative binomial regression to predict and compare HIV mortality rates in Thailand. These factors will be considered as exogenous factors to improve the accuracy of the improved model.

The selection of proper exogenous determinants to improve the performance of a mortality model is essential. Random Forest (RF) is a popular ensemble learning approach that has been applied in classification and regression problems (Fawagreh et al. 2014). The RF method identifies strong predictors with the presence of correlation between predictors via the RF-RFE feature by decreasing the estimated importance scores of correlated variables (Darst et al. 2018). The recursive feature elimination algorithm is effective in selecting relevant predictors which affect mortality. It ranks features by models, eliminates the least significant features and keeps the best features which help in prediction.

This study aims to modify and improve the adequacy of the GLM mortality model by incorporating economic and health-related factors which have an influence on mortality and were selected using the RF-RFE approach for each of the 14 countries studied (United States of America (USA), Spain, Japan, Australia, Netherlands, United Kingdom (UK), Sweden, Canada, Belgium, Taiwan, Italy, Chile, South Korea, and Germany). So far, no work from the GLM perspective on mortality modelling has used the RF-RFE approach to select the best exogenous determinants for mortality modelling. This method will be useful in preventing the use of highly correlated variables in the model, and the best predictors (with low correlation) representing exogenous determinants shall be included to further enhance the performance of the model. This paper is organised as follows: Next section discusses the data and methods used, subsequent section reviews the results and the last section concludes the findings.

MATERIALS AND METHODS

THE DATA SET

In this paper, the mortality data for USA, Spain, Japan, Australia, Netherlands, UK, Sweden, Canada, Belgium, Taiwan, Italy, Chile, South Korea, and Germany were obtained from the Human Mortality Database (HMD 2020). The data includes central mortality rates and mid-year populations by individual years up to 110 years of age. In this study, ages above 85 were grouped as 85+ to avoid erratic rates for these ages. Table 1 shows the periods used to study mortality for different countries.

TABLE 1. Total period mortality data for each country

Country	Year
USA	1970-2016
Spain	1971-2016
Japan	1970-2016
Australia	1971-2016
Netherlands	1972-2016
UK	1970-2016
Sweden	1970-2016
Canada	1970-2016
Belgium	1970-2016
Taiwan	1970-2014
Italy	1970-2014
Chile	2000-2016
South Korea	2003-2016
Germany	1992-2016

The period for the mortality data for each country was determined based on the availability of factors contributing to mortality in that country. The factors that affect mortality of each country such as GDP, alcohol and tobacco consumption, health expenditure, fruit and vegetable consumption, fat supply, carbon dioxide emissions, and crude rates (diabetes mellitus, mental and behavioural disorders, and accidents) were obtained from National Account Data (2020), OECD (2020), Our World in Data (2020), and World Bank (2020).

THE LEE-CARTER MODEL

The LC model (1992) is given by:

$$ln(\mu_{x,t}) = a_x + b_x k_t + e_{x,t} \qquad (1)$$

where $\mu_{x,t}$ is the central mortality rate, calculated as the ratio between the number of people aged x who died in year t and the exposure to death for age x in year t. a_x is the average age-specific mortality, and b_x is a deviation in mortality due to changes of k_t. The parameter k_t

represents the index of the level of mortality at time t. $e_{x,t}$ is the residual at age x and time t and $e_{x,t} \sim N(0,\sigma_e^2)$. Lee and Carter (1992) estimated a_x as the average of $ln\ (\mu_{x,t})$ over time, and the b_x and k_t are estimated by singular value decomposition. The constraints

$$\sum_x b_x = 1 \text{ and } \sum_t k_t = 0 \tag{2}$$

were used to obtain a unique solution. k_t is re-estimated so that the observed number of deaths coincide with the estimates. The adjusted k_t is extrapolated by using the ARIMA (0,1,0) model.

$$\hat{k}_t = \hat{k}_{t-1} + \theta + e_t \tag{3}$$

where θ is the drift parameter and e_t are the normally distributed error terms with mean 0 and variance σ_k^2.

THE MODIFIED GLM LC MODEL

Before converting (1) to the GLM framework, the LC model is modified with the inclusion of additional bilinear predictor structures representing the exogenous determinants used as follows:'

$$ln\left(\mu_{x,t}\right) = a_x + b_x k_t + c_x^{(i)} d_t^{(i)} + e_{x,t},\ x = 0, \dots, x_m \quad t = 1, \dots, t_n; \tag{4}$$

where $c_x^{(i)} d_t^{(i)}$ allows for a difference between the rate of change in death rates implied by the *ith* factor where i =1, 2, 3. The three most important factors affecting mortality for each country are selected by using the RF-RFE method. A GLM model is written as:

$$g\big(E(Y)\big) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_\kappa x_\kappa, \tag{5}$$

where the random component is Y, the systematic components are $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_\kappa x_\kappa$ and the link function is $g(E(Y))$ (McCullagh & Nelder 1989). Suppose $Y = D_{x,t}$ is the number of deaths, let $\eta = a_x + b_x k_t + c_x^i d_t^i$ be the systematic components and ln (.) as the link function. Equation (4) can be worked on, using the GLM framework (Currie 2016):

$$\eta = g\big(E(Y)\big) = g\left(E\big(D_{x,t}\big)\right) = a_x + b_x k_t + c_x^i d_t^i \tag{6}$$

In equation (6), a_x is a linear covariate while $b_x k_t$ and $c_x^i d_t^i$ form bilinear covariates. Thus, the LC model in the GLM framework is a regression model with the presence of the bilinear terms, $b_x k_t$ and $c_x^i d_t^i$ (Denuit et al. 2019).

Let l(a,b,k,c,d) be the corresponding log-likelihood function of $D_{x,t}$. Then, the parameters of the LC model can be estimated by maximizing $l(a,b,k,c,d)$ with respect to a_x, b_x, k_t, c_x and d_t, specifically by solving the score equation; $\mathrm{U}(\theta_i) = \frac{\partial}{\partial \theta_i} \log \mathrm{l}(\theta_i) = 0$, where i =1,2,3,4,5 and $\theta_1 = a_x$, $\theta_2 = b_x$, $\theta_3 = k_t$, $\theta_4 = c_x$, $\theta_5 = d_t$.

The parameter estimation for the GLM framework in the gnm package (Turner & Firth 2020) in the R software is based on the iteratively reweighted least squares (IRLS) algorithm that uses the Newton scheme. The implementation of this algorithm enables a bilinear model to be linearised at each step of the algorithm as it fixes the current values of the parameters and revises the estimates of other parameters. Hence, if either one of the bilinear terms, b_x or k_t, and c_x or d_t are known, then, the modified LC model is in the GLM framework (Denuit et al. 2019).

The IRLS algorithm gives estimates of parameters with random parameterization due to random initial values assigned if not stated. To get the estimates of b_x and k_t subject to the constraints (2), Currie (2016) provided the following equations:

Let $\hat{b}_{x,R}$ and $\hat{k}_{t,R}$ be the estimates returned by the algorithm and $\hat{b}_x$ and $\hat{k}_t$ be the estimates subject to equation (2). Let $\bar{k}_{t,R} = \sum_t \hat{k}_{t,R} / n_y$ and $\bar{b}_{x,R} = \sum_x \hat{b}_{x,R} / n_a$. Then, and can be re-estimated as follows:

$$\hat{b}_x = \hat{b}_{x,R} / (n_a \bar{b}_{x,R}) \tag{7}$$

$$\hat{k}_t = n_a \bar{b}_{x,R} \left(\hat{k}_{t,R} - \bar{k}_{t,R} 1_{n_y}\right) \tag{8}$$

where n_a is the length of ages, n_y is the length of years and 1_{n_y} is the vector of 1's of the length n_y.

THE BINOMIAL BILINEAR LC MODEL

The binomial bilinear LC model (Wang & Lu 2005) was extended by incorporating factors that potentially affect mortality. Let $r_{x,t}$ be the number of people aged x at the beginning of year t, $q_{x,t}$ be the probability of deaths of these $r_{x,t}$ on the condition that it is a closed group, then the number of deaths $D_{x,t}$ can be assumed to follow the binomial distribution with parameters $n = r_{x,t}$ and $p = q_{x,t}$. $q_{x,t}$ is expressed as follows:

$$q_{x,t} = 1 - exp\left(-exp\left(a_x + b_x k_t + c_x^{(i)} d_t^{(i)}\right)\right) \tag{9}$$

for $x = x_1, \dots, x_m$ and $t = t_1, \dots, t_n$.. Hence, it follows that:

$$D_{x,t} \sim Bin(r_{x,t}, q_{x,t}) \text{ with } q_{x,t} = 1 - exp\left(-exp\left(a_x + b_x k_t + c_x^{(i)} d_t^{(i)}\right)\right), \tag{10}$$

with the log-likelihood function:

$$l_{Bin}\left(a,b,c^{(i)},d^{(i)},k\right)=\sum_{t}\sum_{x}\left(\left(r_{x,t}-D_{x,t}\right)\right. \\ \left. ln\left(1-q_{x,t}\right)+D_{x,t}\,ln\left(q_{x,t}\right)\right)\ +\ constant \tag{11}$$

THE POISSON BILINEAR LC MODEL

The number of deaths $D_{x,t}$ is assumed to follow the Poisson distribution (Brouhns et al. 2002), with its mean equals to the product of exposure-to-risk, $e_{x,t}$, and the death rate, $\mu_{x,t}$ which is as follows:

$$D_{x,t}\sim Poisson(e_{x,t}\mu_{x,t})\ \text{ where }\ \mu_{x,t}=exp \\ \left(a_x\ +\ b_xk_t\ +c_x^{(i)}d_t^{(i)}\right). \tag{12}$$

The corresponding log-likelihood function is

$$l_{Poisson}\left(a,b,c^{(i)},d^{(i)},k\right) \\ =\sum_{x,t}(D_{x,t}(a_x\ +\ b_xk_t\ +c_x^{(i)}d_t^{(i)})- \\ e_{x,t}exp(a_x\ +\ b_xk_t\ +c_x^{(i)}d_t^{(i)}))+constant. \tag{13}$$

THE NEGATIVE BINOMIAL BILINEAR LC MODEL

In this section, the negative binomial bilinear LC model (Delwarde et al. 2007) was extended to cater macroeconomic variables. The Poisson bilinear model (Brouhns et al. 2002) assumes that the mean and variance are equal where

$$E\left(D_{x,t}\right)=Var\left(D_{x,t}\right)=e_{x,t}exp\left(a_x\ +\ b_xk_t\ +c_x^{(i)}d_t^{(i)}\right)(14)$$

This equidispersion assumption may not be hold in all cases as the number of deaths varies inconsistently by ages and years. For a higher accuracy, a random effect term $\tau_{x,t}$ can be incorporated into the Poisson model (12) when modeling the number of deaths. Thus, a mixed Poisson model will be obtained. Now, $D_{x,t}$ is assumed to follow a Poisson distribution with mean $e_{x,t}$ *exp* ($a_x + b_x k_t + c_x^{(i)}d_t^{(i)} + \tau_{x,t}$) and the variance,

$$Var\left(D_{x,t}\right)=\delta_{x,t}+\nu\delta_{x,t}^2\ \geq E\left(D_{x,t}\right)=\delta_{x,t}, \\ \text{where } \nu=Var\left(exp\left(\tau_{x,t}\right)\right). \tag{15}$$

If *exp* ($\tau_{x,t}$) follows a gamma distribution, $D_{x,t}$ follows a negative binomial distribution with the log-likelihood function:

$$l_{NegBin}\left(a,b,c^{(i)},d^{(i)},k,\nu\right)=\sum_{x,t}\left(\sum_{j=1}^{N}ln\left(\frac{1}{\nu}+D_{x,t}-j\right)\right) \\ -ln\left(D_{x,t}!\right)-\left(D_{x,t}-\frac{1}{\nu}\right)ln\ (1+\nu\delta_{xt}) \\ +D_{x,t}\,ln\ (\nu\delta_{xt}). \tag{16}$$

ν in (15) represents the dispersion parameter. The Poisson distribution is a special case of the negative binomial distribution where $\nu = 0$.

TABLE 2. The link functions and the random components for the variants of the modified GLM LC model

Model	Link function ***g***(***E***(***Y***))	Random component (***Y***)
Binomial	Logit	$\frac{D_{x,t}}{e_{x,t}}$
Poisson	Log	$D_{x,t}$
Negative Binomial	Log	$D_{x,t}$

THE RANDOM FOREST RECURSIVE FEATURE ELIMINATION (RF-RFE) METHOD FOR THE SELECTION OF FACTORS THAT AFFECT MORTALITY

RF is a supervised learning technique that assembles hundreds of decision trees into a single model. It consists of multiple independent decision trees that operate as an ensemble. Bagging, which is an ensemble algorithm, fits multiple models on different subsets of a training dataset. The predictions from all models are then combined. Bagging improves the stability and accuracy of predictions in the RF algorithm by reducing the variances of the decision trees to prevent overfitting in the training dataset (Arif 2020). In RF, each node of a decision tree is considered as a different subset of randomly selected predictors (Darst et al. 2018).

Each tree is built using a different random bootstrap sample which consists of approximately 70% of the total observations and is used as a training set to predict the data in the remaining testing set. The recursive feature elimination (RFE) approach calculates the importance based on the random forest importance criteria after running each subset of features recursively (Kuhn & Johnson 2013). A control object used to specify the details of the feature selection algorithm based on the cross-validation method with *n* resampling iterations was created. The RF-RFE method eliminates predictors in the mortality rate datasets, one at a time, until the desired number (the top three factors are selected for our case) of features is achieved. These features are later incorporated into the GLM mortality models. It works by searching a subset of features from the original training dataset, ranking the variables by their importance scores, omitting the least important ones, and refitting the model until a specific number of features is achieved (Brownlee 2020). The randomForest (Liaw & Wiener 2002), caret (Kuhn 2020), dbplyr (Wickham et al. 2021) and mlbench (Leisch & Dimitriadou 2010) packages in R were used to perform RF-RFE on our data to select the top three factors which affect mortality in the 14 countries studied.

Results and Discussion

The three GLM frameworks of the LC model as well as their modified versions which incorporated factors that potentially affect mortality were applied to the mortality data of USA, Spain, Japan, Australia, Netherlands, UK, Sweden, Canada, Belgium, Taiwan, Italy, Chile, South Korea, and Germany. The factors affecting mortality for each country are given in Table 3. The top three factors with the highest correlation with mortality for each country were selected via the RF-RFE method and integrated with the three GLM models. This method shortlists the highly influential factors towards mortality even when the predictors are strongly correlated (see Table 4 for the correlation values of the top three factors selected). Multicollinearity was also considered in the selection process. Hence, some highly correlated variables had to be eliminated.

TABLE 3. The top three factors were selected via the RF-RFE for each country (see /)

Factors Study	Factors Selected													
	USA	Spain	Japan	Australia	Netherlands	UK	Sweden	Canada	Belgium	Taiwan	Italy	Chile	South Korea	Germany
Gross Domestic Product (GDP) (per capita)	/	/		/	/			/	/	/				
Alcohol Consumption (Litres per capita)											/	/		/
Tobacco Consumption (grams per capita (15+))				/		/								
Health Expenditure (per capita, current prices)	/	/	/	/	/	/	/	/	/				/	/
Fruit Consumption (kilos per capita per year)						/								
Vegetable Consumption (kilos per capita per year)							/					/		/
Average daily per capita dietary fat supply (grams per capita per day)										/			/	
Diabetes Mellitus (crude rates)												/	/	
Mental and Behavioural Disorders (crude rates)			/		/		/	/	/		/			
Accident (crude rates)	/	/	/								/			
Carbon Dioxide Emissions(metric tons per capita)										/				

TABLE 4. Top three highly correlated factors with log of death rates

Country	Predictor 1	Predictor 2	Predictor 3
USA	GDP (-0.9644)	Healh Expenditure (-0.7951)	Accident (0.9559)
Spain	GDP (-0.8752)	Health Expenditure (-0.3066)	Accident (0.9187)
Japan	Health Expenditure (0.5929)	Mental and Behavioural Disorders (0.7523)	Accident (0.7979)
Australia	GDP (-0.9182)	Tobacco Consumption (0.9649)	Health Expenditure (-0.9678)
Netherlands	GDP (-0.9465)	Health Expenditure (-0.9602)	Mental and Behavioural Disorders (-0.8663)
UK	Tobacco Consumption (0.9572)	Health Expenditure (-0.9945)	Fruit Consumption (-0.8777)
Sweden	Health Expenditure (-0.9699)	Vegetable Consumption (-0.9375)	Mental and Behavioural Disorders (-0.9175)
Canada	GDP (-0.9676)	Health Expenditure (-0.9797)	Health Expenditure (-0.9253)
Belgium	GDP (-0.9513)	Health Expenditure (-0.9643)	Mental and Behavioural Disorders (-0.8134)
Taiwan	GDP (0.9617)	Average daily per capita dietary fat supply (0.8895)	Carbon Dioxide Emissions (0.8476)
Italy	GDP (-0.9533)	Mental and Behavioural Disorders (-0.9548)	Accident (0.9539)
Chile	Alcohol Consumption (0.6519)	Vegetable Consumption (-0.7232)	Diabetes Mellitus (0.4891)
South Korea	Health Expenditure (-0.5479)	Average daily per capita dietary fat supply (-0.5747)	Diabetes Mellitus (0.7618)
Germany	Alcohol Consumption (0.7462)	Health Expenditure (-0.6611)	Vegetable Consumption (-0.8726)

Note: Correlation of the factors with the log of death rates are given in parentheses

The adequacy of these models was assessed via the plots of standardised residuals vs. fitted number of deaths. The residual deviances of the models were examined where the model with the lowest deviance gives the best fit. Results showed that the negative binomial version gives the best fit to the data for cases with and without factors affecting mortality. Table 5 shows that the residual deviance greatly improved for all three variants of the GLM model, with the inclusion of factors that affect mortality, especially the Poisson and binomial variants.

The standardised residual plots and residual deviances show that negative binomial gives the best fit overall for all the 14 countries studied.

The standardised residual plots for the Spanish mortality data are displayed in Figures 1 and 2. The Spanish data is interesting to be analysed due to the presence of extreme values (Azman & Pathmanathan 2020). The impact of incorporating factors that affect mortality shows that the residuals appear to be more random. The dispersion parameter in the negative binomial model gives it an edge especially when the data involves outliers and extreme values. The addition of the time-factor modulation $c_x^i d_t^i$ to the model in (6) improves the adequacy especially in the Poisson and binomial framework. Interestingly, this model also works well for short-base-period data such as Chile (2000-2016), South Korea (2003-2016) and Germany (1992-2016). It is important to identify the factors affecting mortality for each country and this is achieved by using the RF-RFE approach. These factors will represent the time-factor modulation $c_x^i d_t^i$ in (6) and play a key role in improving the accuracy of the model.

As shown in Table 3, the RF-RFE approach chose health spending as the most common factor affecting mortality rates in 11 out of 14 countries studied. This finding is consistent with Kim and Lane (2013) who claim that increased government expenditure on medical goods and services is linked to better individual public health outcomes. Kim and Lane (2013) discovered a negative relationship between public health expenditure and the infant mortality rate, as well as a positive relationship between public health expenditure on life expectancy at birth in 17 OECD countries between 1973 and 2000. The remaining factors selected are given in Table 3. In accordance with the present results, previous studies have shown that taking real-world variations in variables such as GDP, health expenditure, and lifestyle (alcohol, tobacco, fruit, vegetable, and fat consumption) into account explains mortality declines and improves mortality rate forecasting (French & O'hare 2014). The inclusion of factors, as verified by Kim and Lane (2013) and French and O'hare (2014), will lead to significant improvement in mortality forecasting. Therefore, all the factors selected are worthy representatives to test in the GLM LC model. Great improvements in the Poisson and binomial framework are seen with the addition of factors which affect mortality (Figures 1, 2 & Table 5). It is important to take note that simulation strategies may not be useful for determining the best model to represent the mortality rates as a whole because factors such as exogenous determinants and the presence of outliers must be considered depending on the country of choice. Undoubtedly, the negative binomial model will supersede its counterparts due to the presence of the dispersion parameter.

TABLE 5. Deviance statistics for the GLM framework of LC models with and without factors (smallest values are bolded)

Country	Models					
	Negative binomial	Negative binomial (Factors)	Poisson	Poisson (Factors)	Binomial	Binomial (Factors)
USA	4049.30	**4009.91**	122275.20	**21574.97**	123881.30	**22160.01**
Spain	4068.70	**3240.03**	40597.71	**5718.36**	41171.08	**5821.79**
Japan	4048.74	**3419.66**	48048.03	**13407.75**	48259.50	**13706.07**
Australia	4333.76	**3134.33**	11227.47	**4576.98**	11270.07	**4661.37**
Netherlands	3257.92	**2571.44**	5849.39	**3394.13**	5940.07	**3424.67**
UK	4050.94	**3266.83**	36232.61	**10551.33**	36630.40	**10896.47**
Sweden	3704.77	**3005.13**	5147.84	**3757.61**	5219.02	**3812.21**
Canada	3802.78	**3151.54**	15572.67	**4640.88**	15706.75	**4710.45**
Belgium	4161.64	**2832.51**	8324.59	**4504.36**	8417.79	**4607.28**
Taiwan	4118.89	**3030.04**	13470.60	**4744.56**	13644.81	**4845.44**
Italy	4038.39	**3455.56**	36207.70	**7512.56**	37224.92	**7673.71**
Chile	1335.88	**1026.86**	2722.21	**1582.69**	2797.90	**1616.05**
South Korea	1267.78	**834.55**	3416.12	**1753.46**	3480.21	**1788.20**
Germany	2127.07	**1446.53**	21656.15	**10209.37**	22437.71	**10678.21**

Parametric Monte-Carlo simulation is not suitable for this study because different choices of constraints are required to fit the LC model result in widely differing confidence and prediction intervals (Haberman & Renshaw 2008). Other simulation strategies to assess these models also have drawbacks. For example, the semiparametric bootstrap approach for the binomial case still requires a two-parameter probability distribution function representing the binomial distribution with dispersion to map the fitted responses onto the simulated responses (Haberman & Renshaw 2008). The fits of the GLM models are expected to differ for every country but the negative binomial model is the best fit for cases with extreme values and outliers due to the presence of the dispersion parameter. The extra parameters of the modified GLM LC model also managed to improve the model fit since they consider the effects of factors on the number of deaths which are not captured by the basic GLM LC model previously.

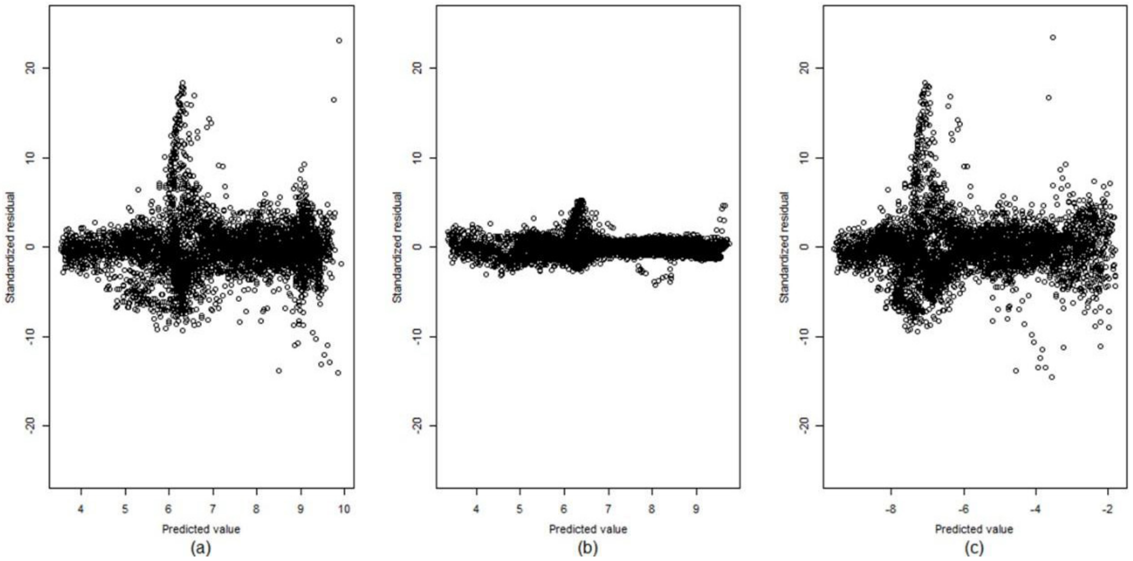

FIGURE 1. Standardised residuals vs. fitted plot for the Spanish population data (without factors): (a) Poisson; (b) Negative binomial; (c) Binomial

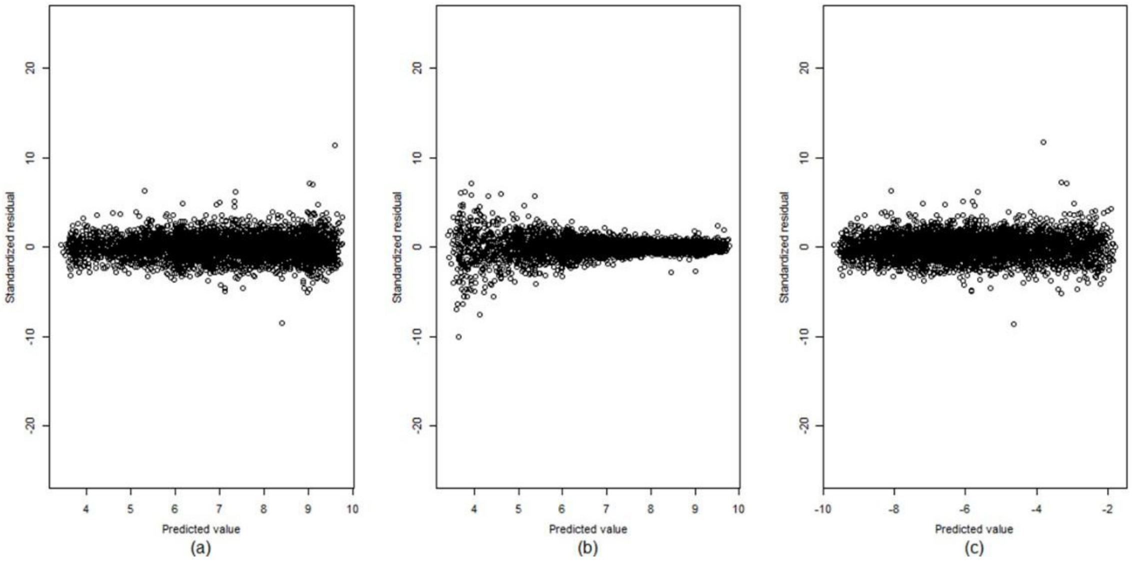

FIGURE 2. Standardised residuals vs. fitted plot for the Spanish population data (with factors): (a) Poisson; (b) Negative binomial; (c) Binomial

Conclusion

The three GLM variants of the LC model were modified by incorporating potential factors that affect mortality based on each selected country. The RF-RFE method played an important role in selecting the factors affecting mortality for each country. The inclusion of factors to the mortality model greatly improved the accuracies especially those of the Poisson and binomial models. Improvement in model accuracy was also seen in the negative binomial extension. However, it did not differ much compared to the original version due to the ability of the negative binomial model to capture overdispersion. It is evident from our study that macroeconomic fluctuations as well as other factors improve the accuracy of mortality models. Integrating exogenous determinants of mortality into the GLM framework of the LC model is feasible due to its straightforwardness in parameter estimation. Furthermore, the proper choice of exogenous determinants using the RF-RFE approach aids in identifying the factors which contribute to the mortality rates of each country. Hyndman and Ullah (2007) used a functional data approach in modelling mortality. This idea sheds light on further investigating mortality models from a generalised functional linear regression model perspective.

Acknowledgements

The authors thank the anonymous referees for their useful suggestions. The authors would also like to convey our heartfelt gratitude to Shafiqah Azman and Aneesha Pillay Balachandran Pillay from Universiti Malaya for sharing relevant work that inspired us to write this paper. This work was supported by University of Malaya, Faculty Research Grant (GPF028B-2018).

References

Arif, R. 2020. *A Simple Introduction to the Random Forest Method.* https://arifromadhan19.medium.com/a-simple-introduction-to-the-random-forest-method-badc8ee6c408

Azman, S. & Pathmanathan, D. 2020. The GLM framework of the Lee-Carter model: A multi-country study. *Journal of Applied Statistics* 49(3): 752-763.

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

Brouhns, N., Denuit, M. & Vermunt, J.K. 2002. A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics* 31(3): 373-393.

Brownlee, J. 2020. *Recursive Feature Elimination (RFE) for Feature Selection in Python. Machine Learning Mastery.* https://machinelearningmastery. com/rfe-feature-selection-in-python/. Accessed on February 15, 2021.

Cairns, A.J., Blake, D., Dowd, K., Coughlan, G.D., Epstein, D., Ong, A. & Balevich, I. 2009. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal* 3(1): 1-35.

Chen, L., Islam, R.M., Wang, J., Hird, T.R., Pavkov, M.E., Gregg, E.W., Salim, A., Tabesh, M., Koye, D.N., Harding, J.L., Sacre, J.W., Barr, E.L.M., Magliano, D.J. & Shaw, J.E. 2020. A systematic review of trends in all-cause mortality among people with diabetes. *Diabetologia* 63(9): 1718-1735.

Currie, I.D. 2016. On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal* 2016(4): 356-383.

Currie, I.D. 2013. Smoothing constrained generalized linear models with an application to the Lee-Carter model. *Statistical Modelling* 13(1): 69-93.

Darst, B.F., Malecki, K.C. & Engelman, C.D. 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genetics* 19(1): 1-6.

Delwarde, A., Denuit, M. & Partrat, C. 2007. Negative binomial version of the Lee-Carter model for mortality forecasting. *Applied Stochastic Models in Business and Industry* 23(5): 385-401.

Denuit, M., Hainaut, D. & Trufin, J. 2019. Some generalized non-linear models (GNMs). In *Effective Statistical Learning Methods for Actuaries I.* Springer, Cham. pp. 363-400.

Fawagreh, K., Gaber, M.M. & Elyan, E. 2014. Random forests: From early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal* 2(1): 602-609.

French, D. 2014. International mortality modelling - An economic perspective. *Economics Letters* 122(2): 182-186.

French, D. & O'Hare, C. 2014. Forecasting death rates using exogenous determinants. *Journal of Forecasting* 33(8): 640-650.

Haberman, S. & Renshaw, A. 2008. On simulation-based approaches to risk measurement in mortality with specific reference to binomial Lee-Carter modelling. In *Society of Actuaries Living to 100 Symposium.*

Hanewald, K. 2011. Explaining mortality dynamics: The role of macroeconomic fluctuations and cause of death trends. *North American Actuarial Journal* 5(2): 290-314.

Hanewald, K., Post, T. & Gründl, H. 2011. Stochastic mortality, macroeconomic risks and life insurer solvency. *The Geneva Papers on Risk and Insurance-Issues and Practice* 36(3): 458-475.

Hansen, H. 2013. The forecasting performance of mortality models. *AStA Advances in Statistical Analysis* 97(1): 11-31.

Human Mortality Database. 2020. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). https://www.mortality.org/.

Hyndman, R.J. & Shang, H.L. 2009. Forecasting functional time series. *Journal of the Korean Statistical Society* 38: 199-211.

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

Kim, T.K. & Lane, S.R. 2013. Government health expenditure and public health outcomes: A comparative study among 17 countries and implications for US health care reform. *American International Journal of Contemporary Research* 3(9): 8-13.

Kuhn, M. 2020. *Caret: Classification and Regression Training. R package version 6.0-86.* https://CRAN.R-project.org/package=caret

Kuhn, M. & Johnson, K. 2013. *Applied Predictive Modeling.* Vol. 26. New York: Springer.

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

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

Leisch, F. & Dimitriadou, E. 2010. Machine learning benchmark problems. *R Package, mlbench*.

Liaw, A. & Wiener, M. 2002. Classification and regression by randomForest. *R news* 2(3): 18-22.

McCullagh, P. & Nelder, J.A. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman and Hall.

National Account Data. 2020. https://unstats.un.org/unsd/snaama/downloads

Nor, S.R.M., Yusof, F. & Norrulashikin, S.M. 2021. Coherent mortality model in a state-space approach. *Sains Malaysiana* 50(4): 1101-1111.

OECD. 2020. *OECD. Stats.* https://stats.oecd.org/

Our World in Data. 2020. https://ourworldindata.org/country/taiwan

Pitt, D., Li, J. & Lim, T.K. 2018. Smoothing Poisson common factor model for projecting mortality jointly for both sexes. *ASTIN Bulletin: The Journal of the IAA* 48(2): 509-541.

Rasoulinezhad, E., Taghizadeh-Hesary, F. & Taghizadeh-Hesary, F. 2020. How is mortality affected by fossil fuel consumption, CO_2 emissions and economic factors in CIS region? *Energies* 13(9): 2255.

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

Seklecka, M., Pantelous, A.A. & O'Hare, C. 2017. Mortality effects of temperature changes in the United Kingdom. *Journal of Forecasting* 36(7): 824-841.

Tulu, H.D., Lim, A., Ma-a-Lee, A., Bundhamcharoen, K. & Makka, N. 2020. Prediction of HIV mortality in Thailand using three data sets from the National AIDS Program Database. *Sains Malaysiana* 49(1): 155-160.

Turner, H. & Firth, D. 2020. *Generalized Nonlinear Models in R: An Overview of the gnm Package*. https://cran.r-project.org/package=gnm

Wang, D. & Lu, P. 2005. Modelling and forecasting mortality distributions in England and Wales using the Lee-Carter model. *Journal of Applied Statistics* 32(9): 873-885.

Wickham, H., Girlich, M. & Ruiz, E. 2021. *dbplyr: A 'dplyr' Back End for Databases. R package version 2.1.0.* https://CRAN.R-project.org/package=dbplyr

Wilmoth, J.R. 1993. *Computational Methods for Fitting and Extrapolating the Lee-Carter Model of Mortality Change.* Technical report, Department of Demography, University of California, Berkeley.

World Bank. World Development Indicators. 2020. https://data.worldbank.org/indicator/EN.ATM.CO2E.GF.KT

Yeh, H.H., Westphal, J., Hu, Y., Peterson, E., Williams, L., Prabhakar, D., Frank, C., Autio, K., Elsiss, F., Simon, G., Beck, A., Lynch, F., Rossom, R., Lu, C., Owen-Smith, A., Waitzfelder, B. & Ahmedani, B. 2019. Diagnosed mental health conditions and risk of suicide mortality. *Psychiatric Services* 70(9): 750-757.

[*]Corresponding author; email: dharini@um.edu.my