Jurnal
Teknologi

Full Paper

PREDICTION OF MORTALITY RATES USING AUGMENTED DATA

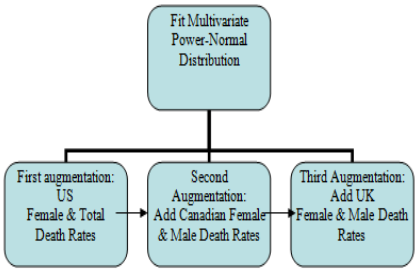
Chon Sern Tan[a*], Ah Hin Pooi[b]

[a]Department of Mathematical and Actuarial Sciences, Universiti Tunku Abdul Rahman, Malaysia
[b]Department of Financial Mathematics and Statistics, Sunway University, Malaysia

Article history
Received
25 October 2015
Received in revised form
14 December 2015
Accepted
9 February 2016

*Corresponding author
tcsern@utar.edu.my

Graphical abstract



Abstract

Prediction of future mortality rate is of significant priority in the insurance industry today as insurers face challenging tasks in providing retirement benefits to a population with increasing life expectancy. A time series model based on multivariate power-normal distribution has been used in the literature on the United States (US) mortality data in the years 1933 to 2000 to predict the future mortality rates in the years 2001 to 2010. To improve the predictive ability, the US mortality data is augmented to include more variables such as death rates by gender and death rates of other countries with similar demographics. Apart from having good ability to cover the observed future mortality rate, the prediction intervals based on the augmented data performed better because they also tend to have shorter interval lengths.

Keywords: Death rates; power-normal distribution; prediction interval; time series model

1.0 INTRODUCTION

Mortality improvement as reflected by lower mortality rates is viewed as a positive change for individuals because they are living healthier and longer. Nevertheless, as lifespan increases, personal savings may end up being insufficient to support their retirement.

From the perspective of insurance providers, longevity improvement presents a challenge for the planning of public retirement systems, life annuities and other longevity-related insurance products of private insurance companies. Prolonged life expectancies lead to the possible risk of underestimating product premiums due to initial assumptions of higher mortality rates, especially in life annuities. Therefore, it has become more crucial for insurers and pension funds to find an appropriate and efficient way to model mortality rates.

In 1992, Lee and Carter pioneered a time series model to make long-run forecasts of age-specific mortality rates [1]. Firstly, logs of central death rates are fitted as a sum of an age-specific constant and the product of a time-varying mortality level index and another age-specific constant. Fitting this model to historical data, singular value decomposition is used to obtain the age-specific constants while the mortality index is modeled as a stochastic time series which in turn is used to forecast future index. Finally, future age-specific central death rates can be forecasted using the forecasted mortality index and age-specific constants.

There have been extensions of the Lee-Carter method that vary according to a number of elements. The more recent extensions are as follows. In [2], the proposed modified Lee-Carter model applied the Lee-Carter model on the difference of log mortality rates, resulting in better performance than the original Lee-Carter model. In this model, a Levy process and the Normal Inverse Gaussian distribution were applied on the mortality index. [3] argued that the Cairns-Blake-Dowd (CBD) model [4] uses the most suitable time-varying model parameters as indexes to indicate longevity risk levels

at different time points. These indexes are jointly modeled with a more general class of multivariate time-series models, instead of a simple random walk that ignore cross-correlations. In [5], the proposed algorithm Multiple Lee-Carter Panel Sieve (MLCPS) combines the Lee-Carter model-based predictions with a bootstrap procedure for dependent data, in order to preserve historical parametric structure and the intra-group error correlation structure. Applying it to estimate the relationship between populations with similar socioeconomic conditions, the empirical results show that it works well in the presence of the dependence structures considered.

In order to reduce problems of model over-parameterization and unjustifiable adding of terms in the model, [6] used a toolkit of functions and expert judgment to design a general procedure to construct mortality models, which can identify sequentially each significant demographic feature in the data and give them a parametric structural form. It produced a relatively parsimonious model with a good fit to the U.K. mortality data. [7] introduced a general framework to model the dynamics of mortality rates of two related populations simultaneously. It prevents forecasts from diverging in the long run by modeling the difference in the stochastic factors between the two populations with a mean-reverting autoregressive process. In [8], improvement on the modeling of the stochastic factors is investigated using a vector error correction model, whose key benefits include eliminating the need to make assumption of which population is dominant.

The method based on multivariate power-normal distribution [9] was used to find prediction intervals for future age-specific mortality rates of United States (US). The US mortality data (1933 to 2000) was used to find a multivariate power-normal distribution from which prediction intervals were found for future mortality rates of the years 2001 to 2010. The resulting prediction intervals were found to have good ability of covering the observed future mortality rates.

In this paper, we augment the data used in [9] to include more variables which are female death rates and death rates of Canada and United Kingdom which would have similar demographics to US. Prediction intervals based on the augmented data for US mortality rates are found to have shorter interval lengths while still having good ability of covering the observed future mortality rates.

The paper layout is as follows. In Section 2, we introduce briefly the multivariate time series given in [10] and highlight some results on mortality rate prediction given in [9]. Section 3 gives the results of prediction based on the augmented data. Finally, Section 4 concludes the paper.

2.0 A MULTIVARIATE TIME SERIES MODEL

The multivariate time series model given in [10] makes use of a non-normal distribution called the power-normal distribution given in [11].

The random variable ε is said to have a power-normal distribution with parameters λ^+ and λ^- if

$$\varepsilon = \psi(\lambda^+, \lambda^-, z) = \begin{cases} \dfrac{(z+1)^{\lambda^+} - 1}{\lambda^+} & if\ (z \geq 0, \lambda^+ \neq 0) \\ \log(z+1) & if\ (z \geq 0, \lambda^+ = 0) \\ -\dfrac{\left[(-z+1)^{\lambda^-} - 1\right]}{\lambda^-} & if\ (z < 0, \lambda^- \neq 0) \\ -\log(-z+1) & if\ (z < 0, \lambda^- = 0) \end{cases}$$

where z has the standard normal distribution.

From the univariate power-normal distribution, we may construct a multivariate power-normal distribution for a vector $\boldsymbol{y}$ consisting of *k*-correlated random variables. The vector $\boldsymbol{y}$ is said to have a *k*-dimensional power-normal distribution with parameters $\boldsymbol{\mu}, \boldsymbol{H}, \lambda^+, \lambda^-, \sigma_i, 1 \leq i \leq k$, if $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{H\varepsilon}$, where $\boldsymbol{\mu} = E(\boldsymbol{y})$, $\boldsymbol{H}$ is an orthogonal matrix consisting of the eigenvectors of the variance-covariance matrix of $\boldsymbol{y}$, and $\varepsilon_1, \ldots, \varepsilon_k$ are uncorrelated,

$$\varepsilon_i = \sigma_i \left[\tilde{\varepsilon}_i - E\left(\tilde{\varepsilon}_i\right)\right] / \left\{\mathrm{var}\left(\tilde{\varepsilon}_i\right)\right\}^{1/2},$$

$\sigma_i > 0$ is a constant and ε_i has a power-normal distribution with parameters λ_i^+ and λ_i^-.

From the multivariate power-normal distribution, we may construct a multivariate time series model for a vector $\boldsymbol{x}(t)$ of n_c observations recorded at time *t*. Letting Δt be a small time increment after *t*, an $n_c(l+1)$-dimensional power-normal distribution is found for the vector $\boldsymbol{x}^{(1)} = \left[\boldsymbol{x}(t-(l-1)\Delta t), \ldots, \boldsymbol{x}(t-\Delta t), \boldsymbol{x}(t), \boldsymbol{x}(t+\Delta t)\right]$. An n_c-dimensional conditional distribution of $\boldsymbol{x}(t+\Delta t)$ is next found from the above $n_c(l+1)$-dimensional power-normal distribution. This n_c-dimensional conditional distribution will then specify a lag(*l*-1) multivariate time series model for the vector of n_c time-dependent correlated observations.

Assuming the multivariate time series is stationary, we may consider that, for $d \geq 2$, the vector $\boldsymbol{x}^{(d)} = \left[\boldsymbol{x}(t+(d-l)\Delta t), \ldots, \boldsymbol{x}(t+(d-1)\Delta t), \boldsymbol{x}(t+d\Delta t)\right]$

has the same distribution as vector $\boldsymbol{x}^{(i)}$. Thus, given the value of $\left[\boldsymbol{x}\left(t+\left(d'-l\right)\Delta t\right),\ldots,\boldsymbol{x}\left(t+\left(d'-2\right)\Delta t\right),\boldsymbol{x}\left(t+\left(d'-1\right)\Delta t\right)\right]$, we may find a conditional distribution for $\boldsymbol{x}\left(t+d'\Delta t\right)$ and later generate a value for $\boldsymbol{x}\left(t+d'\Delta t\right)$ for $d'=2,3,\ldots,d$. In this way, a value of $\boldsymbol{x}\left(t+d\Delta t\right)$ may be generated. The process of generating $\boldsymbol{x}\left(t+d\Delta t\right)$ may be repeated a large number of times. From the generated values of $\boldsymbol{x}\left(t+d\Delta t\right)$, we may find the marginal distribution for the *j*-th component of $\boldsymbol{x}\left(t+d\Delta t\right)$. The prediction intervals with end points given by the $100\left(\alpha/2\right)$ and $100\left(1-\alpha/2\right)$ percentage points of the marginal distribution may be used to predict the value of the *j*-th component of $\boldsymbol{x}\left(t+d\Delta t\right)$.

Table 1 displays nominally 95% prediction intervals obtained in [9], based on constructed lag-0 model for total death rates of age group 60-64. A subset of three age groups (55-59, 60-64 and 65-69) was used. In the table, L_d and U_d are respectively the intervals' lower and upper limits for the mortality rate *d* years after the year 2000.

Table 1 The nominally 95% prediction intervals for future total mortality rate of age group 60-64

d	L_d	U_d	Interval Length
1	0.01120	0.01270	0.00150
2	0.01060	0.01300	0.00240
3	0.01020	0.01290	0.00270
4	0.00990	0.01300	0.00310
5	0.00954	0.01310	0.00356
6	0.00922	0.01310	0.00388
7	0.00894	0.01310	0.00416
8	0.00870	0.01320	0.00450
9	0.00850	0.01310	0.00460
10	0.00821	0.01330	0.00509

3.0 PREDICTION OF US MORTALITY RATES USING AUGMENTED DATA

In this section, the methodology in [9] will be applied to augmented US mortality data. The mortality data of the years 1933 to 2000 will be used as training data (68 years) while the mortality data of the years 2001 to 2010 will be used as testing data (10 years).

The first augmentation is to combine the US training data of total death rates (female and male combined) with US female death rates. For a given year (say year *t*), the age-specific total mortality rates for a subset of N age groups from the 19 age groups 15-19,20-24,…,105-109 are initially used to form a vector $\boldsymbol{m}_t^{(T)}$ of *N* components. The vector $\boldsymbol{m}_t^{(T)}$ is next augmented to the vector $\boldsymbol{m}_t=\left(\boldsymbol{m}_t^{(F)},\boldsymbol{m}_t^{(T)}\right)$ by the inclusion of the US female age-specific mortality rates for the same subset of N age groups.

The second augmentation is to combine the US training data from the first augmentation above with the female death rates and male death rates of Canada. The resulting vector of mortality rates may now be stated as $\boldsymbol{m}_t=\left(\boldsymbol{m}_t^{(F_c)},\boldsymbol{m}_t^{(M_c)},\boldsymbol{m}_t^{(F)},\boldsymbol{m}_t^{(T)}\right)$ where $\boldsymbol{m}_t^{(F_c)}$ and $\boldsymbol{m}_t^{(M_c)}$ are respectively the Canadian female and Canadian male age-specific mortality rates for the same subset of *N* age groups.

The third augmentation is to combine the training data from the second augmentation with the female death rates and male death rates of the United Kingdom (UK). The resulting vector of mortality rates may now be stated as $\boldsymbol{m}_t=\left(\boldsymbol{m}_t^{(F_u)},\boldsymbol{m}_t^{(M_u)},\boldsymbol{m}_t^{(F_c)},\boldsymbol{m}_t^{(M_c)},\boldsymbol{m}_t^{(F)},\boldsymbol{m}_t^{(T)}\right)$ where $\boldsymbol{m}_t^{(F_u)}$ and $\boldsymbol{m}_t^{(M_u)}$ are respectively the United Kingdom female and United Kingdom male age-specific mortality rates for the same subset of *N* age groups.

For the I_a-th augmentation, we set $n_c=2I_aN,\Delta t=1$ and $\boldsymbol{x}\left(t+k\Delta t\right)=\boldsymbol{m}_{t+k}$ and we use the values of $\boldsymbol{m}_t$ from 1933 to 2000 to estimate the $2I_aN\left(l+1\right)$-dimensional power-normal distribution for $\left(\boldsymbol{m}_{t-(l-1)},\ldots,\boldsymbol{m}_t,\boldsymbol{m}_{t+1}\right)$. From the $2I_aN\left(l+1\right)$-dimensional power-normal distribution, a nominally $100\left(1-\alpha\right)\%$ prediction interval is next obtained for the *j*-th component of $\boldsymbol{m}_{t+d}$ where $1\le j\le N$ and $d\ge1$.

Let $N=3$ and consider the subset formed by the age groups 55-59, 60-64 and 65-69. For the first augmentation with US female death rates, Table 2 displays the nominally 95% prediction intervals based on the constructed lag-0 model for the total death rates of the age group 60-64.

Table 2 The nominally 95% prediction intervals for total death rates of age group 60-64 – first augmentation with US female death rates

d	L_d	U_d	Observed Future Death Rate, O_d	Prediction Interval Length	
				US	Table 1
1	0.01143	0.01287	0.01205	0.00144	0.00150
2	0.01117	0.01263	0.01188	0.00146	0.00240
3	0.01052	0.01273	0.01177	0.00221	0.00270
4	0.01012	0.01262	0.01132	0.00250	0.00310
5	0.00966	0.01257	0.01126	0.00291	0.00356
6	0.00914	0.01264	0.01081	0.00350	0.00388
7	0.00876	0.01253	0.01063	0.00376	0.00416
8	0.00808	0.01268	0.01051	0.00460	0.00450
9	0.00765	0.01261	0.01026	0.00496	0.00460
10	0.00689	0.01270	0.01006	0.00580	0.00509

From Table 2, it can be observed that all the prediction intervals cover the observed future death rates. However, the lengths of prediction intervals are shorter than those in Table 1 for d=1,…,7 only, while the remaining three interval lengths are longer respectively. This indicates that the augmentation with US female death rates may not improve the prediction when *d* is large.

Table 3 displays the corresponding prediction intervals for the second augmentation (Canada-US) while Table 4 shows the corresponding prediction intervals for the third augmentation (UK-Canada-US).

Table 3 The nominally 95% prediction intervals for total death rates of age group 60-64 – second augmentation (Canada-US)

d	L_d	U_d	Observed Future Death Rate, O_d	Prediction Interval Length		
				US	Canada-US	Table 1
1	0.01126	0.01276	0.01205	0.00144	0.00149	0.00150
2	0.01074	0.01218	0.01188	0.00146	0.00145	0.00240
3	0.01012	0.01223	0.01177	0.00221	0.00211	0.00270
4	0.00966	0.01221	0.01132	0.00250	0.00256	0.00310
5	0.00927	0.01216	0.01126	0.00291	0.00289	0.00356
6	0.00889	0.01208	0.01081	0.00350	0.00319	0.00388
7	0.00893	0.01208	0.01063	0.00376	0.00315	0.00416
8	0.00809	0.01183	0.01051	0.00460	0.00375	0.00450
9	0.00801	0.01167	0.01026	0.00496	0.00366	0.00460
10	0.00759	0.01149	0.01006	0.00580	0.00390	0.00509

Table 4 The nominally 95% prediction intervals for total death rates of age group 60-64 – third augmentation (UK-Canada-US)

d	L_d	U_d	Observed Future Death Rate, O_d	Prediction Interval Length		
				Canada-US	UK-Canada-US	Table 1
1	0.01145	0.01282	0.01205	0.00149	0.00137	0.00150
2	0.01106	0.01230	0.01188	0.00145	0.00124	0.00240
3	0.01020	0.01204	0.01177	0.00211	0.00184	0.00270
4	0.00981	0.01211	0.01132	0.00256	0.00230	0.00310
5	0.00938	0.01192	0.01126	0.00289	0.00254	0.00356
6	0.00894	0.01187	0.01081	0.00319	0.00293	0.00388
7	0.00864	0.01170	0.01063	0.00315	0.00305	0.00416
8	0.00850	0.01165	0.01051	0.00375	0.00315	0.00450
9	0.00803	0.01160	0.01026	0.00366	0.00357	0.00460
10	0.00768	0.01149	0.01006	0.00390	0.00380	0.00509

It can be observed that all the prediction intervals in Table 3 and Table 4 cover the observed future death rates while the lengths of prediction intervals are also shorter than those in Table 1 for $d=1,\ldots,10$. This shows that both sets of augmented training data have improved the predictive ability through shorter lengths of prediction intervals. In addition, it can be observed that when the US data is augmented to the Canada-US data and later to the UK-Canada-US data, the interval length tends to decrease with each augmentation when *d* is large.

4.0 CONCLUSION

This paper presents a fairly promising application of a multivariate time series model on the mortality data of the United States. The results indicate improvement effected by augmentation of data on the model's ability of covering the future observed mortality rates with shorter prediction intervals. This is consistent with our intuition that the predictive ability may be improved by enriching the existing data with other *similar* data from the countries with similar demographics. As a further work, the effect of combining more mortality data can be explored by adding death rates of a fourth country (such as Ireland) to the UK-Canada-US combination.

Acknowledgement

We are grateful to UTM for the opportunity to publish in *Jurnal Teknologi*.

References

[1] Lee, R.D., and L.R. Carter. 1992. Modeling and Forecasting US Mortality. *Journal of the American Statistical Association*. 87(419): 659-671.
[2] Brockett, P. L., S.L. Chuang, Y. Deng, and R.D. MacMinn. 2013. Incorporating Longevity Risk and Medical Information into Life Settlement Pricing. *Journal of Risk and Insurance*. 80(3): 799-826.
[3] Chan, W. S., J.S.H. Li, and J. Li. 2014. The CBD Mortality Indexes: Modeling and Applications. *North American Actuarial Journal*. 18(1): 38-58.
[4] Cairns, A. J., Blake, D., and Dowd, K. 2006. A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk and Insurance*. 73(4): 687-718.
[5] D'Amato, V., S. Haberman, G. Piscopo, and M. Russolillo. 2012. Modelling Dependent Data for Longevity Projections. *Insurance: Mathematics and Economics*. 51(3): 694-701.
[6] Hunt, A., and D. Blake. 2014. A General Procedure for Constructing Mortality Models. *North American Actuarial Journal*. 18(1): 116-138.
[7] Cairns, A.J., D. Blake., K. Dowd, G.D. Coughlan, and M. Khalaf-Allah. 2011. Bayesian Stochastic Mortality Modelling for Two Populations. *Astin Bulletin*. 41(01): 29-59.
[8] Zhou, R., Y. Wang, K. Kaufhold, J.S.H. Li, and K.S. Tan. 2014. Modeling Period Effects in Multi-Population Mortality Models: Applications to Solvency II. *North American Actuarial Journal*. 18(1): 150-167.
[9] Pooi, A.H., W.Y. Pan., and Y.C. Wong. 2014. Prediction Intervals for Future Mortality Rates. *Applied Mathematical Sciences*. 8(101): 5039-5051.
[10] Pooi, A.H. 2012. A Model for Time Series Analysis. *Applied Mathematical Sciences*. 6(115): 5735-5748.
[11] Yeo, I.K., and R.A. Johnson. 2000. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*. 87(4): 954-959.