TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

A Cohort Based Mortality Forecast in England and Wales with Application of K-Nearest Neighbor Classification for Causes of Death

Sak Guo Bin & Raja Rajeswari Ponnusamy
School of Mathematics, Actuarial and Quantitative Studies, Asia Pacific University, Malaysia
Email Address: gordonsak@gmail.com, raja.rajeswari@staffemail.apu.edu.my

Article Info
Volume 82
Page Number: 843 - 854
Publication Issue:
January-February 2020

Article History
Article Received: *14 March 2019*
Revised: *27 May 2019*
Accepted: *16 October 2019*
Publication: *04 January 2020*

Abstract
Continuous increase in overall life expectancy has brought forward the importance of mortality forecasting. Estimates in mortality trends have become extremely crucial and informative in various fields, including the planning and funding of health care systems, pension schemes as well as pricing annuity portfolios and reserves. Furthermore, while it is universally agreed that while age and period effects have significant effects onto mortality modelling, cohort effects have presented mixed results, mainly favorable towards the population of England and Wales. Underlying cause-of-death has also become a topic of interest in terms of mortality modelling, as it brings significance towards the forecasted mortality rates, but have yet to be proven or thoroughly understood. Hence, the aim of this study is to compare several popular mortality models using the population of England and Wales, as well as test the significance of cohort effects within the population. This study also intends to classify deaths according to their respective causes using the k-nearest neighbor algorithm, allowing possible assumptions to mortality data without cause-of-death in future studies.

Keywords: *Mortality Forecasting, Lee-Carter, Cohort Effects, K-Nearest Neighbor (KNN) Classification*

1. INTRODUCTION

Development within the medical field in the 21st century has continuously shown noteworthy increase in life expectancy. Together with the increasing awareness of the public regarding the importance of life insurance, the need of mortality forecasting has been brought forward into the limelight. According to a report by Great Britain's Office for National Statistics (2018), the Department of Health and Social Care (DHSC) are key users of the currently available mortality statistics, analyzing trends based on various causes of death between different age groups, especially towards infant mortality be it stillbirths or neonatal deaths. Mortality data on external causes of death are the topic of interest for other public-sector organizations in identifying and making preemptive measures to reduce these deaths. Forecasted mortality rates are then fed into all kinds of statistical models for the interest of various parties, ranging from risk estimation to the calculation of pensions and benefits.

While mortality forecasting has had its fair share of history, the methods then were simplistic and was subjected to a reasonable amount of judgement from the researcher themselves. The introduction of stochastic methods prompted the revolution in how mortality forecasting is done today. It provided a major advantage in producing forecasts in terms of a probability distribution rather than a deterministic point forecast, allowing the

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

development and application of more sophisticated and complex forecasting methods.

In regard to mortality statistics, the World Health Organization (2018) have defined underlying cause-of-death as "the disease or injury which persisted throughout, or a fatal injury initiated by the circumstances of an accident or violence, leading directly towards death". To date, mortality forecasting has generally been done towards overall mortality rates. However, it is also argued that trends in specific causes of death and future likely changes need to be taken into consideration in mortality forecasting to fully understand the projected mortality rates. Some of these trends include medical advancements, changes in lifestyle as well as accessibility to medical and healthcare. Hence, theoretically speaking, mortality rates should be calculated based on specific cause-of-death groupings before combining them in some manner to make forecasts and estimations. While it sounds simple from a theoretical standpoint, it is almost impossible to carry out practically, mainly due to obstacles in understanding specific trends in various causes of death, especially within disease-based models.

2. LITERATURE REVIEW

2.1 Previous Mortality Studies in England and Wales

The Lee-Carter (LC) model has been performed in multiple studies in the past due to its simplicity in calculation and interpretation as well as its sheer popularity, it being one of the staples for mortality forecasting models of today. Many studies regarding mortality forecasting models have also used the LC model as a main model for comparison in terms of forecasting accuracy, as the results produced are consistent with the actual mortality rates. Throughout most of these studies however, there has always been various adjustments and changes done towards some of the variables, thus producing slight inconsistencies between results even within similar periods. While the LC model is able to capture mortality experience in England and Wales fairly decently, its projection of mortality rates has a tendency to be less accurate compared to some of the later models or even variants of the LC model itself(Renshaw & Haberman, 2003; Booth, et al., 2005).

Booth, Tickle and Smith did a review on the LC model and its two popular variants, the Lee-Miller (LM) variant and the Booth-Maindonald-Smith (BMS) variant, comparing between multiple countries including England and Wales. Both variants differ from the LC method in terms of adjustment in the fitting period, the calculation of the index for overall mortality in each respective year and the jump-off rates. The study concluded that both LM and BMS variants were superior to LC in terms of both forecast accuracy and width of prediction interval(Booth, et al., 2005).

A generalized linear modelling (GLM) regression-based approach was done by Renshaw and Haberman in comparison with the LC method, rooting from retrospective study regarding actuarial mortality reduction factors performed in the United Kingdom. While there was evidence that the GLM approach is more successful at capturing age specific mortality trends, it was significant only towards the males and not the females. The study also showed improvement in overall life expectancy at birth as well as a diminishing gap of life expectancy being forecasted between both genders using the GLM approach(Renshaw & Haberman, 2003).

The LC method was once again used together with its variants, LM and BMS in a study for fourteen developed countries which includes England. Another variant of the LC method was also introduced named the Tuljapurkar-Li-Boe (TLB) method, which restricts the fitting period to 1950 and make no adjustments to the index for overall mortality in respective years. Additional comparisons were also done using a non-parametric method, the Hyndman-Ullah (HU) method including some variants which was proposed based off a functional data analysis technique to model and forecast log mortality rates. Among the forecasting methods performed, the weighted HU method was the most accurate due to the

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

smaller age-specific errors attributed by greater weight to recent data. Other than that, the study also found that the LM method was the least biased, the BMS method performed best for male mortality and the TLB method performed best for female mortality(Shang, et al., 2011).

2.2 Cohort Effects in Mortality Forecasting

During 2006, Renshaw and Haberman then introduced two cohort-based mortality models, one of them being an extension using the LC model, which was applied to mortality data from England and Wales ranging since 1961 to 2003 for each gender. The projection of mortality experience by the suggested LC cohort-extended model (RH model) outperformed the LC model, as it managed to capture characteristic and systematic mortality directly attributable to not only age and period, but also towards cohort effects(Renshaw & Haberman, 2006).Another suggested cohort model (APC model) also performed considerably well compared to the LC model, but trailed behind the RH model. Since then, many mortality models which involved cohort effects have used these cohort models in their research to better understand the significance of cohort effects in the countries of their respective studies, with relative amount of success(Booth & Tickle, 2008).

Several papers were published after that, continuing to compare the RH and APC models with other mortality models using the population of England and Wales as well as other countries. Martina Gustafsson (2011) conducted a study in testing cohort effects on Swedish mortality but also included data from England and Wales as well as Denmark. According to the study, cohort effects in England and Wales were negligible for ages above 60 and remained inconclusive for the rest, which contradicts with another popular study done by Cairns, Blake and Dowd (2006) which developed the CBD model named after themselves. In their study, they found cohort effects in England and Wales of ages above 60 to significant and their model was best fit towards this age group.

2.3 The Relationship between Cause-of-Death and Overall Mortality Rates

The topic of mortality modelling and forecasting have recently shifted from overall mortality towards cause of death. This helps insurance companies and healthcare experts to identify and hopefully, better understand mortality trends in specific causes of death. While the initial appeal of breaking down mortality data into respective cause-of-death can be easily done and understood, the recombination of separate projections by cause-of-death to produce overall mortality rates is extremely difficult. This is because most trends by cause-of-death are not independent and may have cause-and-effect relationships which are easily misunderstood(Richards, 2009).

One of the most recent cause-specific mortality was done in Korea, which consisted up to 12 major causes of death based on past trends from years 1983 to 2012, including ranging from various diseases, accidents to suicides. The estimates of future mortality rates were done using the APC model with slight modifications which has been frequently used mainly in cancer mortality data all over the world. However, the study does have its flaws that there is no mention in any testing of the forecasts' accuracy, mainly focusing on the projected mortality rates along with its interpretations towards the shifting mortality trends on each respective cause-of-death.

The most relevant research done in England and Wales with cause-of-death in mind was done by Cesare and Murphy back in 2009. In their study, they forecasted mortality with various approaches for the cases of lung cancer; influenza, pneumonia and bronchitis (IPB) as well as motor vehicle accidents (MVA). The LC model and BMS variant was engaged in the study, together with the APC model and the Bayesian model. The study made three key conclusions(Cesare & Murphy, 2009):

1. LC based approaches are best selected in situations where drivers of past trends act in a largely linear fashion and copes well with unpredictable changes in trends (MVA) and period effects (IPB).

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

2. APC models produce the best forecast for causes of deaths with clear cohort patterns (lung cancer) but is worse than the LC approaches in terms of forecasting period-driven causes of death.
3. The Bayesian model estimates mortality, but nothing significantly better than the other models considered.

2.4 The K-Nearest Neighbor Algorithm and Classification

The k.-nearest neighbor classification uses the concept that a set of data should be classified based of its nearest neighbor which shares the most similar characteristics with the data being classified. This form of classification is considered as a conventional, non-parametric classifier that when provided an optimal k-value, produces well-performed classifications on the dataset.

The algorithm of the k-nearest neighbor was first proposed by Cover and Hart in 1968 with the calculation of certain measures which include the Euclidean distance, the cityblock (taxicab metric), the cosine distance as well as correlation. Among these measures, the Euclidean is widely used due to its ease of use and interpretation(Aman Kataria, 2013).

The parameter k is another concept whereby the k-value decides the amount of the closest neighbors should be chosen for the algorithm. The k-value impacts the diagnostic performance of the algorithm significantly, as a large k-value may ignore smaller yet important patterns, but greatly reduces the impact of variance caused by random errors and vice versa. Although there have been authors that suggested setting the k-value to equal to the square root of the number of observations in the training dataset for the algorithm, the most optimal k-value can only be obtained through trial and error in most cases(Zhang, 2016).

The k-nearest neighbor classification is one of the most useful forms of classification but does have its shortcomings(Imandoust & Bolandraftar, 2013; Zhang, 2016). The main issue with this form of classification using k-nearest neighbor algorithm is the dependence in choosing an optimal number of neighbors, or the k-values. Different samples may produce varying optimal k-values and may sometimes have multiple k-values of seemingly equal accuracy. Hence, many studies have been performed to solve this problem, with little success(Hassanat, et al., 2014).

Another issue with the k-nearest neighbor classification is the memory requirement and time complexity in running the algorithm(Hassanat, et al., 2014). This is because the algorithm is completely dependent on each example available in the training set. Hence, a large training set may cause the algorithm to take more time to classify any data that is to be validated, which can cause this form of classification to be unfeasible for extremely large datasets to run.

3. METHODOLOGY

3.1 Data Collection

The datasets used in this study are all taken from the official website for the Office for National Statistics of Great Britain, https://www.ons.gov.uk. In terms of overall mortality, the dataset consists of mortality data compiled using the 3-year life tables provided based on ages 0 to 100 taken from years 1980 to 2017. The single-year life tables for years 1980 to 2014 are derived from the already available single-year life tables for years 2015 to 2017. The dataset is then further segregated in terms of central rate of mortality m_x and number of exposures E_x as well as male and female population to fit the *demogdata* function available in R for analysis purposes.

In the case of cause-of-death based mortality, the dataset consists of mortality data compiled based on age groups with intervals of 10 (except age group of 1 to 4) and underlying cause from years 2008 until 2017. Only the data from this 10-year period is publicly available and accessible on the website relative to the data for overall mortality. Mortality data for ages under 1 are ignored as there are extreme values among

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

these data, mainly due to infancy-related diseases. Mortality data for ages 85 and above are also not used as the original data itself does not show a specific limiting age, which may cause inaccuracy in the current study.

3.2 Research Model

3.2.1 Lee-Carter (LC) Model

The LC model is one of the first models that popularized the principal components approach to mortality forecasting, extrapolating age and period, both time-related parameters using time series methods(Booth & Tickle, 2008). Assuming Poisson distribution of deaths similarly to Renshaw & Haberman (2006) in forecasting mortality for the population of England and Wales, the underlying two-factor LC model is given as:

$$ln\, m_{x,t} = \alpha_x + \beta^{(1)}{}_x k_t + \varepsilon_{x,t}$$

$m_{x,t}$ - the central rate of mortality at age x as of year t,

α_x - the age-specific mean value of the log-mortality at age x averaged across the years over which the model is fitted,

$\beta^{(1)}{}_x$- the rate of change in $k_t$based on age x,

k_t- the index of the overall level of mortality as of year t,

$\varepsilon_{i,x}$ - the residuals or error terms for age x as of year t.

The LC model is used with parameter constraints of:

$$\sum_x \beta_x^{(1)} = 1, \qquad \sum_t k_t = 0$$

3.2.2 Renshaw-Haberman's (RH) Lee-Carter Cohort-Extended Model

The RH model with cohort extension that has been selected is a recent addition by Renshaw and Haberman (2006), whereby they extended the LC model to include cohort as a third factor. The number of deaths, $D_{x,t}$ is assumed to follow the Poisson distribution. The model is given as:

$$ln\, m_{x,t} = \alpha_x + \beta^{(1)}{}_x k_t + \beta^{(0)}{}_x \gamma_{t-i} + \varepsilon_{x,t}$$

where similar parameters observed from the LC model have the same meaning, the exceptions being:

$\beta^{(0)}{}_x$- the rate of change in x_{t-x}based on age x,

γ_{t-x}- the overall level of mortality for the cohort born in year t-x.

The set of parameter constraints for the RH model are also similar to the original LC model with some additions:

$$\sum_x \beta_x^{(1)} = 1, \qquad \sum_t k_t = 0, \qquad \sum_x \beta_x^{(0)} = 1, \qquad \sum_{c=t_1-x_k}^{t_n-x_1} \gamma_c = 0$$

3.2.3 Renshaw-Haberman's Age-Period-Cohort (APC) Model

The APC model was subsequently developed together with the RH model and shows the same model with only changes corresponding to parameters $\beta^{(1)}{}_x$= 1 and $\beta^{(0)}{}_x$= 1, resulting in the following model with similar explanations to the models previously mentioned:

$$ln\, m_{x,t} = \alpha_x + k_t + \gamma_{t-i} + \varepsilon_{x,t}$$

The following parameter constraints are imposed when using the APC model:

$$\sum_t k_t = 0, \qquad \sum_{c=t_1-x_k}^{t_n-x_1} \gamma_c = 0 \qquad \sum_{c=t_1-x_k}^{t_n-x_1} c\gamma_c = 0$$

3.2.4 Euclidean Distance

In terms of the k-nearest neighbor classification for varying causes of deaths for this study, the Euclidian distance has been employed due to its simplicity and ease of understanding which can be expressed through the following equation(Aman Kataria, 2013):

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

$$D(x_i, x_j) = \sqrt{\sum_{k}^{p} (x_{i,k} - x_{j,k})^2}$$

$x_{i,k}$ is an input (training) sample with k characteristics from a total of m training samples,

$x_{j,k}$ is an input (test) sample with k characteristics from a total of n test samples, and

p is the total number of characteristics, which in this study we have 2.

3.3 Measure of Forecasting Errors

Following previous studies of Renshaw and Haberman (2006) on the population of England and Wales, independence between period and cohort effects have been observed and are also assumed in generating estimations of k_t and γ_{t-x} using univariate ARIMA processes for this study. Various ARIMA processes will be carried out for comparison to best generate estimations of the parameters before ultimately projecting mortality rates.

It is generally agreed upon that higher number of parameters in a model leads to a better fit towards data when evaluating goodness-of-fit for different models. However, it is also that the performance of the model is severely affected by over-parametrization over several possible models. To figure out which ARIMA process is the most suitable for the period index and cohort index, the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC).

AIC: $-2logp(L) + 2p$

BIC: $-2logp(L) + plog(n)$

For both criterion, L refers to the likelihood under the fitted model; p is the number of parameters in the model and; n is the sample size of the fitted data(Aikake, 1973; Schwarz, 1978). Lower values of both criterion for the fitted models are preferred, generally presenting more accurate results(Acquah, 2010).

For the k-nearest neighbor classification of cause-of-death, only k-values of 1 to 10 are tested due to constrictions in time. To identify the most suitable k-value for classification between the male and female population respectively, cross tables are created to cross-validate between the predicted classification and the true classification of the test data, whereby this method is suggested in many k-nearest neighbor algorithm related studies(Aman Kataria, 2013; Anava & Levy, 2016). This allows a better understanding as to the effectiveness of the model varying of k-values, as well as the possible reasons of misclassification for certain diseases.

4. DISCUSSION

4.1 Forecasting Mortality for All Causes of Death

4.1.1 Goodness-of-Fit Test on Models

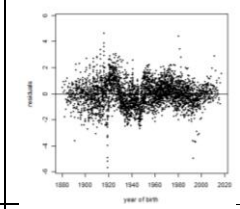
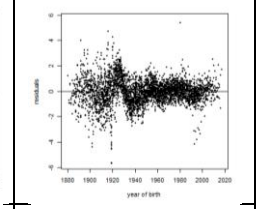
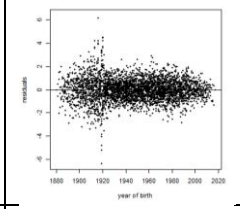
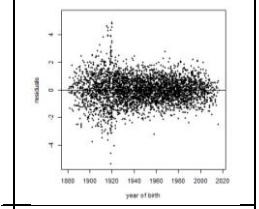
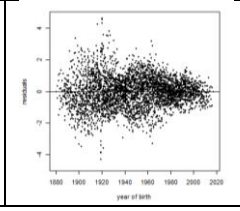
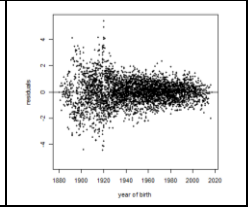
Model	Male	Female
LC		
RH		
APC		

Table 1: *Residual scatter plots for cohort index of fitted models based on gender*

Residuals of each fitted model is calculated using the *StMoMo* package, which is then plotted into scatter plots by age, period and cohort to better visualize any patterns within these residuals. Based on the cohort patterns shown in Table 1, the cohort residuals of the LC

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

model do not appear to be random especially those after the year 1920, revealing that the fitted LC model fails to capture relevant cohort effects. On the other hand, the cohort residuals for the APC as well as the RH model appears to be more random.

Models	Criterion		
	AIC	*BIC*	*Log-likelihood*
LC	50811	52299	-25167
RH	41905	44250	-20577
APC	49354	51067	-24403

Table 2: *AIC, BIC and log-likelihood values for fitted models of male population*

Models	Criterion		
	AIC	*BIC*	*Log-likelihood*
LC	46909	48397	-23216
RH	40112	42456	-19681
APC	44597	46310	-22024

Table 3: *AIC, BIC and log-likelihood values for fitted models of female population*

By applying AIC, BIC and log-likelihood onto the fitted models, the RH model returns the smallest values of AIC and BIC as well as the largest value for log-likelihood for both male and female population compared to the other two models. Thus, we have mathematically concluded that the RH model is the most suitable and appropriate method among the three in forecasting future England and Wales mortality rates using the currently available mortality dataset with the support of various residual plots.

4.1.2 ARIMA Testing on Period and Cohort Index

With the RH model being selected, several independent univariate ARIMA processes are tested to best forecast the period index, k_t for respective genders. Furthermore, due to cohort effects being computed using the RH model, ARIMA processes are also tested on the cohort index $\gamma_{t-x}$for more accurate results.

ARIMA (p, d, q)	Criterion		
	AIC	*BIC*	*Log-likelihood*
0, 0, 0	370	374	-183
0, 0, 1	326	331	-160
0, 1, 0	127	130	-61
1, 0, 0	201	206	-97
0, 1, 1	129	134	-61
1, 0, 1	194	201	-93
1, 1, 0	129	134	-61
1, 1, 1	130	137	-61

Table 4: *AIC, BIC and log-likelihood values of tested univariate ARIMA processes on male period index*

ARIMA (p, d, q)	Criterion		
	AIC	*BIC*	*Log-likelihood*
0, 0, 0	227	231	-111
0, 0, 1	196	201	-95
0, 1, 0	124	127	-60
1, 0, 0	135	140	-64
0, 1, 1	120	125	-57
1, 0, 1	134	141	-63
1, 1, 0	117	122	-55
1, 1, 1	119	126	-55

Table 5: *AIC, BIC and log-likelihood values of tested univariate ARIMA processes on female period index*

In the case for male period index, the final values show a close tie among ARIMA (0,1,0), ARIMA (0,1,1) and ARIMA(1,1,0). Among the three univariate ARIMA processes, ARIMA (0,1,0) was selected as it showed the lowest AIC and BIC, with the log-likelihood trailing slightly behind the other two. The selection of ARIMA (1,1,0) for the female period index was more direct, as it was ranked first in terms of AIC and BIC values as well as second in terms of log-likelihood.

ARIMA (p, d, q)	Criterion		
	AIC	*BIC*	*Log-likelihood*
0, 0, 0	306	311	-151
0, 0, 1	129	138	-61

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

0, 1, 0	-446	-440	225
1, 0, 0	-433	-424	219
0, 1, 1	-459	-450	232
1, 0, 1	-448	-436	228
1, 1, 0	-472	-463	239
1, 1, 1	-418	-406	213

Table 6: *AIC, BIC and log-likelihood values of tested univariate ARIMA processes on male cohort index*

ARIMA (p, d, q)	**Criterion**		
	AIC	*BIC*	*Log-likelihood*
0, 0, 0	231	237	-113
0, 0, 1	58	66	-26
0, 1, 0	-555	-549	279
0, 1, 1	-558	-549	282
1, 1, 0	-561	-552	283
1, 1, 1	-561	-549	284

Table 7: *AIC, BIC and log-likelihood values of tested univariate ARIMA processes on female cohort index*[1]

Based on Tables 6 and 7,ARIMA (1,1,0) is the most appropriate univariate ARIMA process to be applied in forecasting the cohort index for both genders respectively. This is further supported by Renshaw and Haberman's (2006) study, which also uses ARIMA (1,1,0) for projecting the cohort index of the England and Wales population.

4.1.3 Mortality Forecasting and Simulation with the RH Model

The period index, k_t and cohort index, γ_{t-x} are then forecasted 30 years ahead. In terms of the male population, ARIMA (0,1,0) and ARIMA (1,1,0) are applied respectively. ARIMA (1,1,0) is applied to both indexes for the case of the female population.

[1]ARIMA (1,0,0) and ARIMA (1,0,1) are omitted due to error values being return using R.

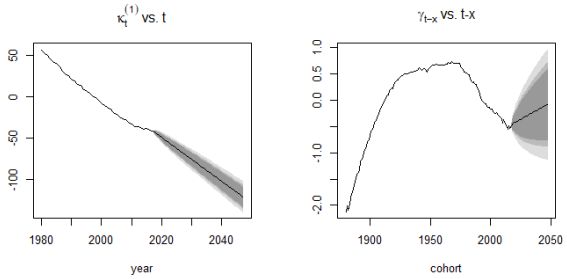


Figure 8:*Forecasted period index, k_t(left) and forecastedcohort index, γ_{t-x}(right) for male population*

At first glance, the forecasted male period index seems to be more consistent across the 90%, 95% and 99% confidence intervals with very little deviance as it continues to slope down. However, the male cohort index seems to spread across a wider range across previously mentioned confidence intervals. Based on the Figure 8, it can also be said that while the forecast favors the cohort index to increase in the next 30 years, it is not impossible for the cohort index to decrease as well.

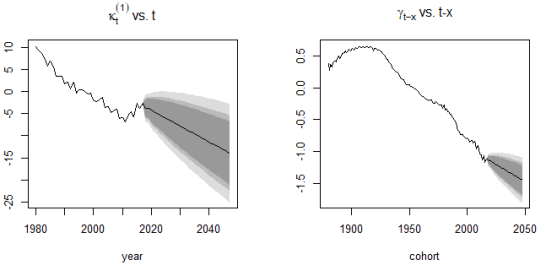


Figure 9:*Forecastedperiod index, k_t(left) and forecasted cohort index,γ_{t-x}(right) for female population*

On the other hand, the forecasted female period index shows a larger deviance across the 90%, 95% and 99% confidence intervals compared to its male counterpart. Although an increase may happen in the few years to come, the overall consensus is that the period index will have decreased 30 years later. The female cohort index shows a much smaller spread compared to its period index as well as its male counterpart. This may be due to the more obvious downward trend displayed in the female cohort index that began in earlier cohorts unlike the male cohort index which have had drastic

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

changes over time. This may also be generalized as an assumption that the female cohort effect is more significant to capture compared to the male population.

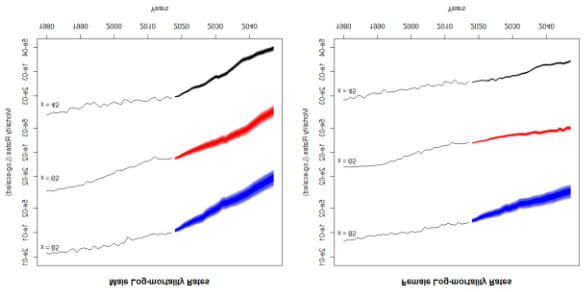


Figure 10: *Forecasted log-mortality rates for ages 45 (black), 65 (red) and 85 (blue) for male population (left) and female population (right)*

Using 1000 simulated trajectories for the period and cohort indexes, mortality rates of all ages for both genders were forecasted at 90%, 95% and 99% confidence intervals. Figure 10 displays the forecasted log-mortality rates of ages 45, 65 and 85 for years 2018 to 2047. It can be observed that as the age of forecast increases, the spread also increases for both genders. This can be interpreted that the actual mortality rate is much more like to differ if we were to use traditional singular-value forecast methods.

When we compare the plots for both genders, it can be said that the forecasted female mortality rates have a much smaller spread compared to the male population for across all ages. This shows that there is less deviation in the forecasted values of female mortality rates compared to male mortality rates, allowing us to better pinpoint the actual mortality rate within the female population in the future. This does not mean the male mortality rates are less accurate. It just simply means that there are more possible values for the male mortality rates to deviate compared to the female mortality rates.

4.2 K-Nearest Neighbor Classification on Causes of Death

4.2.1 Fitting and Adjusting the Data

To ensure that the k-nearest neighbor classification can be run effectively, certain limitations were put unto the data to remove possible outliers, whereby this technique is very sensitive towards. The data being used is from years 2008 to 2017, as there is no data publicly published prior to 2008. The diseases being classified are shortlisted to circulatory, digestive,external causes of death, neoplasms, nervous and respiratory as these causes of death have the highest mortality in England and Wales based on their respective causes. Moreover, ages below 1 and above 85 are not considered, the former being heavily affected by infancy and newborn related diseases. The latter is due to no specification on the upper limits of deaths that occur past age 85.

Once all the data have been inputted into the R program, the data is normalized so that the data across all independent variables are consistent with each othersuch that the data will always fall in the range of 0 to 1. The normalized dataset is then segregated into two parts: the training set and validation set. The training set is the dataset that will be used by the program to assist in its initial analytical model building of the k-nearest neighbor classification, whereas the validation set is then inputted to evaluate the performance of the analytical model built by the program. In this case, we initially create partitions of 2/3 and 1/3 for the training and validation sets respectively. To ensure that the partitions remain the same for future use or reference, a constant seed was set beforehand within the R program in creating the required partitions.

4.2.2 Analysis of Data

Using the *knn3Train* function from the *caret* package to perform k-nearest neighbor classification, we first input the training set for the program to create the analytical model in classifying our dataset into the three causes of death previously mentioned. The validation set is also tested with varying k-values to evaluate the performance of the model done using machine learning through trial and error. Evaluation is done via cross tabulation, where we test the classification results of the model against the actual cause-of-death. K-values from

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

the range of 1 to 10 are tested multiple times, as it is possible to have differences in predicted classifications with the same k-value due to possible ties in classification, causing the final classificationto be randomly chosen between the tied classes.

Actual Class	Predicted Class					
	Cir	*Dig*	*Ext*	*Neo*	*Ner*	*Res*
Cir	27	1	0	2	0	0
Dig	0	27	0	0	1	2
Ext	2	1	23	2	0	2
Neo	3	0	2	24	0	1
Ner	2	0	2	1	24	1
Res	2	2	0	0	3	23

Table 11: *Cross-table between estimates and actual results of k-nearest neighbor classification for male population (k = 2)*

It is observed that the k-value of 2 produces the most accurate results for the male population, ranging between 140 and 148 out of 180 (77.78% and 82.22% respectively) possible outcomes correctly predicted. Based on Table 11 which takes one case where the maximum number of correct predictions is achieved for the male population, most deaths caused by circulatory and digestive system-related diseases were correctly predicted at 90% effectiveness, whereas external causes of mortality and respiratory-related diseases had the least predicted correctly at 76.67% effectiveness. Out of the 32 wrong classifications, 9 of them were wrongly predicted (28.13%) as caused by circulatory-related, making it the highest cause-of-death to be wrongly classified as among the male population.

Actual Class	Predicted Class					
	Cir	*Dig*	*Ext*	*Neo*	*Ner*	*Res*
Cir	28	1	0	0	1	0
Dig	3	24	0	0	3	0
Ext	1	0	21	5	1	2
Neo	0	0	3	25	0	2
Ner	0	3	2	2	21	2
Res	2	0	2	0	4	22

Table 12: *Cross-table between estimates and actual results of k-nearest neighbor classification for female population (k = 3)*

The k-value of 3 is more appropriate for the female population with a range of 134 to 141 out of 180 (74.44% and 78.33% respectively) possible outcomes correctly classified. Based on Table 12 which takes one case where the maximum number of correct predictions is achieved for the female population, most deaths caused by circulatory system-related diseases were correctly predicted at 93.33% effectiveness, whereas external causes of mortality and nervous system-related diseases had the least predicted correctly at 70% effectiveness. Out of the 32 wrong classifications, 9 of them were wrongly predicted (28.13%) to be caused by nervous system-related, making it the highest cause-of-death to be wrongly classified as among the female population.

Smaller k-values may have played a part in the inconsistencies of the predicted classifications despite using the same k-value. This is because small k-values are much more likely to cause ties when voting for the classification of the validation data. Although larger k-values may theoretically solve this problem, it has been observed the results of the

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

k-nearest neighbor classification for both the male and female population shows acceptable for k-values of smaller values, but becomes the estimates become increasingly inaccurate as the k-values starting from 5 begin approaching 10, which invalidates this possible solution. On the other hand, correctly classified estimates for the male population seems to be higher compared to the female population. Hence, it can be assumed that mortality data based on cause-of-death in England and Wales for the male population is more consistent compared to the female population for possible machine learning and classification purposes.

5. CONCLUSION

When it comes to forecasting overallmortality in England and Wales, it is safe to assume that the RH model is still one of the best mortality models to project future mortality rates, be it the male or female population. The study has also proved that cohort effects are a significant factor in mortality forecasting for England and Wales besides the two common factors of age and period, further supported that the APC model which considers cohort effects still produced better forecasts than the LC model despite being worse that the RH model according to the information criterion. This encourages the use of mortality models that account for cohort effects to be used in comparison with other mortality models for future mortality studies, as some countries may also have cohort effects that when considered, may improve the mortality literature for said country or population.

In terms of classifying certain deaths to their respective cause-of-death using the k-nearest neighbor algorithm, the study has shown that by plugging an appropriate k-value, an effectiveness of 70% to 80% can be achieved. The success of this form of classifying cause-of-death is considerably subjective.Hence, classification of varying causes of death in England and Wales using the k-nearest neighbor algorithm remains inconclusive due to other considerations that may improve said effectiveness tohigher percentages which are more satisfactory.

REFERENCES

1. Acquah, H. d.-G., 2010. Comparison of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in Selection of an Asymetric Price Relationship. Journal of Development and Agricultural Economics, 2(1), pp. 1-6.
2. Aikake, H., 1973. Information Theory and an Extension of the Maximum Likelihood Principle. 2nd International Symposium on Information Theory, Volume 73, pp. 1033-1055.
3. Aman Kataria, M. D. S., 2013. A Review of Data Classification Using K-Nearest Neighbour Algorithm. International Journal of Emerging Technology and Advanced Engineering, 3(6), pp. 354-360.
4. Anava, O. & Levy, K. Y., 2016. k-Nearest Nighbors: From Global to Local. Barcelona, s.n.
5. Booth, H. & Tickle, L., 2008. Mortality Modelling and Forecasting: A Review of Methods. Annals of Actuarial Science, 3(1-2), pp. 3-43.
6. Booth, H., Ticle, L. & Smith, L., 2005. Evalutaion of the Variants of the Lee-Carter Method of Forecasting Mortality: A Multi-Country Comparison. Special Issue, New Zealand Population Review, 31(1), pp. 13-34.
7. Cesare, M. D. & Murphy, M., 2009. Forecasting Mortality, Different Approaches for Different Cause of Deaths? The Cases of Lung Cancer; Influenza, Pneumonia, and Bronchitis; and Motor Vehicle Accidents. British Actuarial Journal, 15(S1), pp. 185-211.
8. Hassanat, A. B., Abbadi, M. A. & Altarawneh, G. A., 2014. Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. International Journal of Computer Science and Information Security, 12(8), pp. 33-39.
9. Imandoust, S. B. & Bolandraftar, M., 2013. Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background. International

January - February 2020
ISSN: 0193 - 4120 Page No. 843 - 854

Journal of Engineering Research and Applications, 3(5), pp. 605-610.

10. Office for National Statistics of Great Britain, 2018. Mortality Statistics in England and Wales Quality and Methodology Information Report, Newport: Office for National Statistics of Great Britain.
11. Renshaw, A. E. & Haberman, S., 2003. Lee-Carter Mortality Forecasting with Age-Specific Enhancement. Insurance: Mathematics and Economics, Volume 33, pp. 255-272.
12. Renshaw, A. E. & Haberman, S., 2006. A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors. Insurance Mathematics and Economics, Volume 38, pp. 556-570.
13. Richards, S., 2009. Selected Issues in Modelling Mortality by Cause and in Small Populations. British Actuarial Journal, 15(Supplement), pp. 267-283.
14. Schwarz, G., 1978. Estimating the Dimension of a Model. The Annals of Statistics, 6(2), pp. 461-464.
15. Shang, H. L., Booth, H. & Hyndman, R. J., 2011. Point and Interval Forecasts of Mortality Rates and Life Expectancy: A Comparison of Ten Principal Component Methods. Demographic Research, 25(5), pp. 173-214.
16. Willets, R. C., 2004. The Cohort Effect: Insights and Explanations. British Actuarial Journal, Volume 10, pp. 833-877.
17. World Health Organization, 2018. WHO - Mortality. [Online] Available at: http://www.who.int/topics/mortality/en/[Accessed 25 August 2018].
18. Zhang, Z., 2016. Introduction to Machine Learning: K-Nearest Neighbors. Annals of Translational Medicine, 4(11), pp. 218-225.