STAT0035: Applying and extending the Lee-Carter model

KQVN9

Last updated: 12/05/2021

Contents

1 Introduction

1.1 Background

The term mortality rate is typically used to refer to the proportion of the number of deaths that have occurred in a population in relation to the size of the population within a given time period. The importance of being able to understand, model and forecast mortality rates within a population has inspired the production of this project. Acknowledging past, present and future changes in mortality plays a vital role in decision making in areas such as public health, the insurance industry and government.

Drawing inference in the analysis of mortality rates can allow for extremely signifcant conclusions to be made. For example: the mortality rate of a population with a specific disease can enable inference to be made on the probability of surviving the disease. The importance in the analysis of mortality is undeniable and, therefore, it is vital that statistical modellling of mortality can be considered reliable and accurate.

An inspiration of the project is the ongoing COVID-19 pandemic. The current pandemic has indicated the important nature of mortality modelling for both business and government applications. A pandemic is often considered an anomaly within mortality modelling. However, a concequence of the pandemic will be the increase in attention surrounding mortality in many business and government sectors.

Stochastic models are a form of mathematical model which can consider various scenarios given an inital condition.[7] Stochastic models are often associated with Monte-Carlo simulation. This is a beneficial feature for a project of this nature because simulation enables the potential to forecast and predict unknown values. This project specifically focuses on the stochastic model known as the Lee-Carter model.

1.2 Project Structure and Aims

Mortality rates can be used to estimate the life expectancy of a population. The Lee-Carter model is a commonly used stochastic model to understand and forecast life expectancy in actuarial applications. The first aim of this project is to gain an understanding of life expectancy in the United Kingdom (UK) by creating and applying the Lee-Carter model.

The UK has an estimated population of approximately 68 million people with the median age in the UK being 40.5 years old.[13] The UK is made up of many local areas (cities, towns, villages, etc) and the age distribution within each of these local areas varies within these communities.[1] The Lee-Carter model has traditionally been used to analyse life expectancy at a national level. Furthermore, the next aim of this project is to adapt the Lee-Carter model to incorporate local area, instead of age, against time. This adaptation of the Lee-Carter model will create an idea of the extent to which different local areas have impacted national mortality.

The final aim of the project is to propose an further adaptation to the Lee-Carter model to consider all variables: age, local area and time. This proposal will be applied to UK mortality data alongside the

original Lee-Carter model. The purpose of this proposal will be to improve the accuracy of the original stochastic model and potentially be a better predictor of mortality.

To summarise the aims of this project:

1) Understanding and applying the Lee-Carter model to UK mortality data.
2) Adapting the Lee-Carter model to incorporate local areas within the UK rather than age groups.
3) Proposing, testing and applying an extension to the Lee-Carter model to consider age, time and local area.

The aims of the project exist to provide direction in terms of the structure of this report. Each section of the report will refer to an aim, with the anticipation of methodically fulfilling each aim throughout the report.

1.3 Literature Review

Each section of the project concentrates on one of the three aims of the project. The first aim looks at applying the Lee-Carter model which was proposed by Lee and Carter in 1992, which is a commonly used stochastic model for mortality modelling. The mathematical foundations of the Lee-Carter model are present in each of the sections of the report. The basic concepts, understanding and assumptions behind the Lee-Carter model are stated in a variety of papers. The 'Understanding the Lee-Carter Mortality Forecasting Method' (Federico Girosi and Gary King; 2007) and 'Actuarial Mathematics for Life Contingent Risks' (David C. M. Dickson, Mary R. Hardy and Howard R. Waters; 2019) both outline similar methodology for the framework and assumptions in the Lee-Carter model. Princeton University's report of Lee-Carter modelling[11] (2017) presents similar methods for estimation (singular value decomposition), modelling and forecasting (Monte-Carlo Markov Chain) to the approaches throughout the project. This article also applies the Lee-Carter model to life table data (US data rather than UK) from the Human Mortality Database.

The second aim involved an adaptation to the Lee-Carter model for United Kingdom mortality that incorporates local areas. The adaptation to the Lee-Carter model was an original approach that built upon the Lee-Carter model and was devised through the assumptions of the Lee-Carter model. The third aim is met though using a classification tree approach to build upon and extend the Lee-Carter model. This was another original approach based on the assumptions and methodology of the Lee-Carter model. 'A random forest algorithm to improve the Lee–Carter mortality forecasting: impact on q-forward' (Susanna Levantesi & Andrea Nigri) looks at applying a random forest algorithm to improve the Lee-Carter model, although this approach involves using decision trees for forecasting whilst the approach in this paper applies classification trees in the modelling stage.

2 The Lee-Carter model applied to United Kingdom life tables

2.1 The Lee-Carter model

The Lee-Carter model is a stochastic longetvity model which was derived through the work of Ronald D. Lee and Lawrence Carter in the modelling and forecasting of US mortality over 30 years ago. The Lee-Carter model is commonly used by actuaries for life expectancy forecasting in insurance/ pension pricing applications.

The Lee-Carter model focuses on the central death rate, denoted m_x, rather than the standard mortality rate calculation (number of deaths divided by population). The central death rate represents the average number of deaths each year at age x last birthday in the relevant time period, divided by the average population at that age over the same period.[8]

An assumption of the Lee-Carter model is that the central death rate varies by both age x and time t. Therefore, we consider the central death rate as a function of both age and time - denoted m_{xt}.

The formula for the Lee-Carter model is as follows:

$$log(m_{xt}) = \alpha_x + \beta_x K_t + \epsilon_{xt}$$

where α_x is the force of mortality at each age group x, β_x represents the change in mortality for each age group x, K_t is the trend in mortality over time t and ϵ_{xt} is the error value.

2.2 Descriptive analysis of data

In order to gain an understanding of the Lee-Carter model we can apply the model to UK life tables. UK life tables are publicly available in the Human Mortality Database (HMD)[5]. The UK life table contain the central death rate m_{xt} for each age from 1922 to 2018 with seperate tables for both males and females.

The life table contain 3 columns of interest: year (t), age (x) and central death rate (m_{xt}). The Lee-Carter model is in terms of $log(m_{xt})$, whilst the UK life table data provides central death rate m_{xt}. Therefore, the logarithm of each m_{xt} value is taken.

A general understanding of the relationships and trends between $log(m_{xt})$, age x and year t can be examined through looking at figures 1 and 2.

Figure 1: Change in log(mxt) by age in each year (1920-2019) displayed by decade

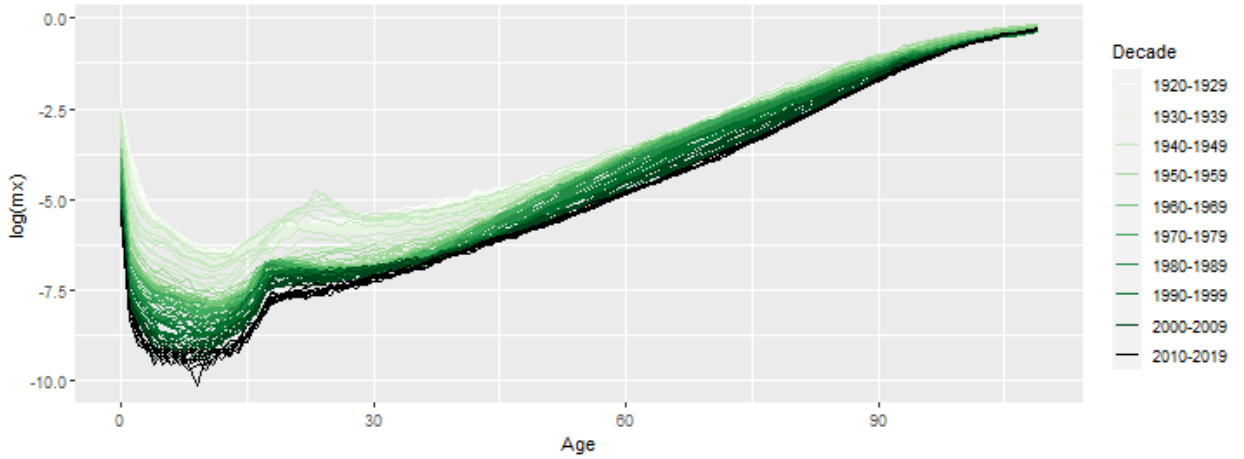
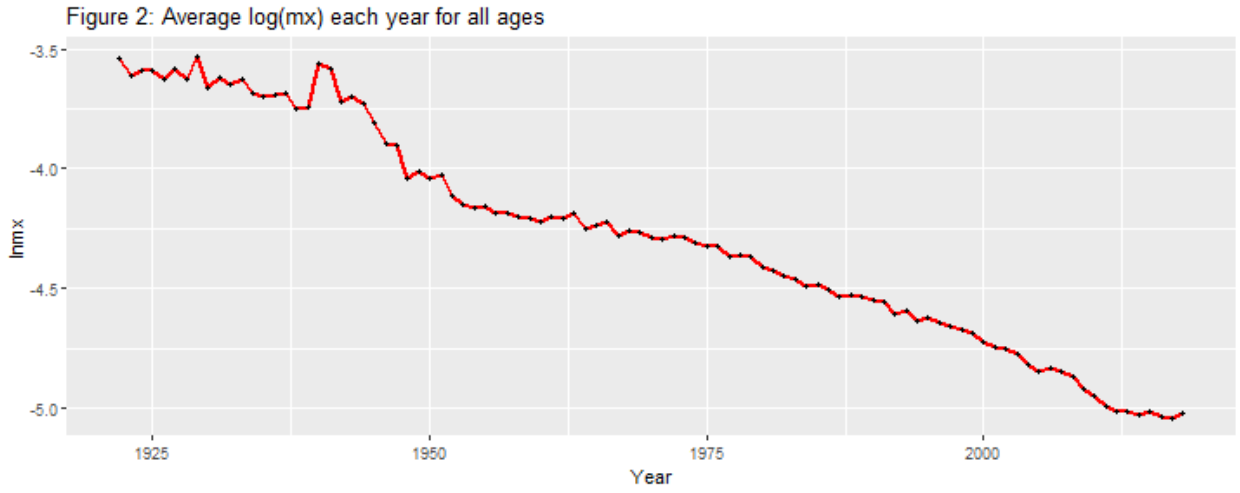


Figure 1 shows a sharp decrease in $log(m_{xt})$ from ages 0 to 9, this decrease is present in all years - this suggests that there is a significantly higher levels of infant mortality in comparison to childhood mortality in the UK. The sharp decrease in $log(m_{xt})$ is followed by a steady increase from ages 10 to 109. Between ages 15 and 30 there seems to be an increase in the rate of growth of $log(m_{xt})$, with a higher rate of growth that ends at an earlier age in the more recent years and a relatively lower increase in the rate of $log(m_{xt})$ growth in the earlier years with the growth period being slightly longer.

Another aspect which can be observed from figure 1 is the general decline in $log(m_{xt})$ over time, this is displayed through the consistent decrease in $log(m_{xt})$ as the decade observed increases. Figure 1 also displays a high spread of $log(m_{xt})$ values within younger age groups, this spread decreases as age increases and there is a relatively small spread of $log(m_{xt})$ within in the older ages. This indicates that $log(m_{xt})$ decreases over time and the rate of decline decreases as age increases.



The indication from figure 1 that $log(m_{xt})$ decreases over time is reinforced in figure 2. Figure 2 displays a clear negative correlation between average $log(m_{xt})$ and year. There seems to be slightly higher variation of $log(m_{xt})$ in the earlier years compared to the later. Figure 2 also provides an idea of what the trend of mortality over time, K_t, will look like in a Lee-Carter model.

2.3 Estimation of variables

In order to fit the Lee-Carter model using the UK life table data, values for α_x, β_x and K_t need to be estimated.

α_x represents the average $log(m_x)$ at each age. This therefore implies that α_x can be estimated by taking the sum of all $log(m_x)$ values at age x and dividing this by total number of years observed (defined n_t):

$$\alpha_x = \frac{\sum_{t=1}^{n_t} log(m_{xt})}{n_t}$$

where α_x is the estimate for α at age x and n_t represents the number of observed years.

Taking the average value of $log(m_{xt})$ for each age in the UK life table, an estimate of α_x can be observed. Figure 3 displays the value of α at each age x:

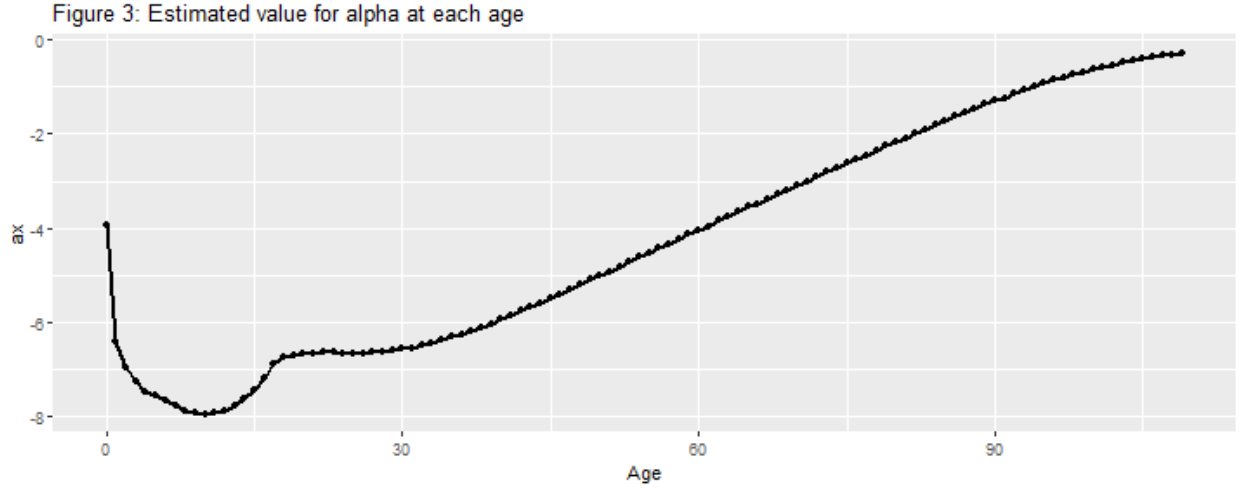


Figure 3 shows that α_x follows a similar trend to $log(m_x)$ in figure 2; there is a relatively large decrease from ages 0-9 followed by a constant increase from age 10 onwards, with a slight increase in rate of growth after age 15 which flattens out around age 20.

Estimations for both β_x and K_t cannot be obtained through the same process as α_x. β_x and K_t appear in the Lee-Carter model as a product. By rearranging the formula for $log(m_{xt})$ in the Lee-Carter model, a formula for $\beta_x K_t$ can be derived:

$$\beta_x K_t = log(m_{xt}) - \alpha_x - \epsilon_{xt}$$

The product of β_x and K_t is approximately equal to $log(m_{xt})$ subtracted by α_x. Furthermore, through subtracting the previously calculated values of $log(m_{xt})$ and α_x, an estimate of the product $\beta_x K_t$ is obtained. The product $\beta_x K_t$ exists in a matrix where the ages, $x = (0, 1, ..., 109)$, are the rows and years, $t = (1920, 1921, ..., 2019)$, are the columns.

This creates an estimate of the product of $B_x K_t$. However, this matrix needs to be split into individual matrices in order to gain individual estimates for each B_x and K_t.

Singular value decomposition (SVD) is a method which allows a matrix A to be factorised into three seperate matrices[12]:

$$A = USV^T$$

where matrices U and V are orthonormal matrices and the matrix S is a diagonal matrix with positive real values[12].

Singular value decomposition (SVD) can be used to obtain the optimal lower-rank value approximation to matrix A. Taking the first column vector of U, the first row vector of V and the first entry in the diagonal matrix S, optimal lower rank values for U, V and S.[11]

Applying SVD to the matrix of $\beta_x K_t$ and taking the optimal lower-rank value approximation, the following form of $\beta_x K_t$ is obtained:

$$\beta_x K_t = S_{(1,1)} u_{(x,1)} v_{(1,t)}$$

where $S_{(1,1)}$ (a constant) is the first value from the diagonal matrix S, u_x is the first column in the matrix u and v_t is the first row of the matrix v.

This form of $B_x K_t$ produced through SVD enables recognition that u_x is an estimate for β_x and the product $S_1 v_t$ is an estimate of K_t. All values of v_t will also be directly proportional to K_t as S_1 is a constant. This implies that also future values for v_t will be directly proportional to K_t, this creates the potential of forecasting of $log(m_{xt})$ as K_t is the only variable in terms of time t in the formula for $log(m_{xt})$.

An assumption behind the estimation of variables in the Lee-Carter model is that α and β are considered as constants over time. Whilst, K is considered as a random variable where the step size is normally distributed with mean μ (drift) and variance σ (volatility).[4] Furthermore, if μ is negative then K tends to decrease meaning that central death rate will follow a general decrease over time. σ represents the the volatility from year to year and will impact the size of the change each year.

another cite for svd: https://www.sciencedirect.com/topics/engineering/singular-value-decomposition

Applying SVD to the estimate for $\beta_x K_t$:

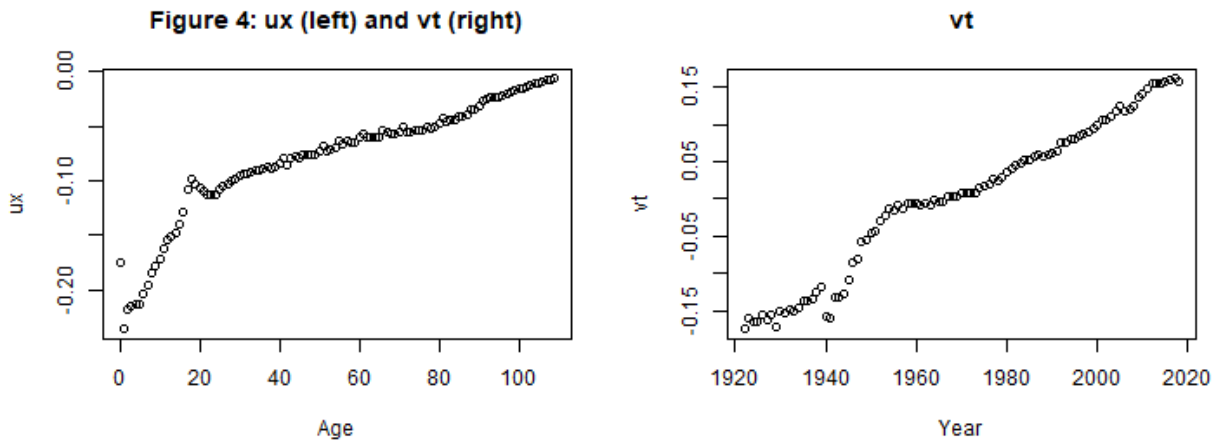


Figure 4 displays both $u(x)$ and $v(t)$. Interestingly, $v(t)$ generally increases as year increases, which is the opposite of what is expected from the previous descriptive analysis of figure 2. Observing figure 4, it can be noticed that $u(x)$ is negative for all x. This will therefore reverse the relationship of $v(t)$ which aligns with the assumption made from looking at figure 2.

Figure 4 highlights a feature of SVD as it approximates $v(t)$ as a normally distributed variable with mean 0, this can be observed through the rough symmetry around 0. This enables the Monte-Carlo simulation of $v(t)$ to be a very simple process.

2.4 Analysis of Lee-Carter estimated values of Central Death Rate

Through substituting estimates for α_x, S_1, $u(x)$ and $v(t)$, an estimate for $log(m_{xt})$ can be formed:

$$log(m_{xt}) = \alpha_x + (S_1 u(x) v(t))$$

These estimates of $log(m_{xt})$ are the fitted values of the Lee-Carter model, and can be compared against the actual value of $log(m_{xt})$ to evaluate the performance of model.

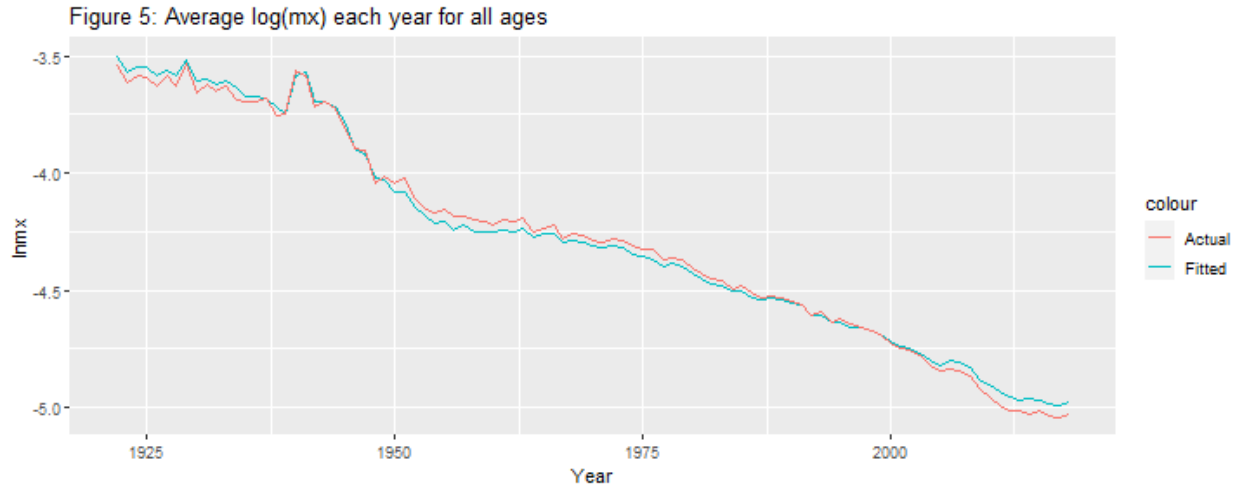
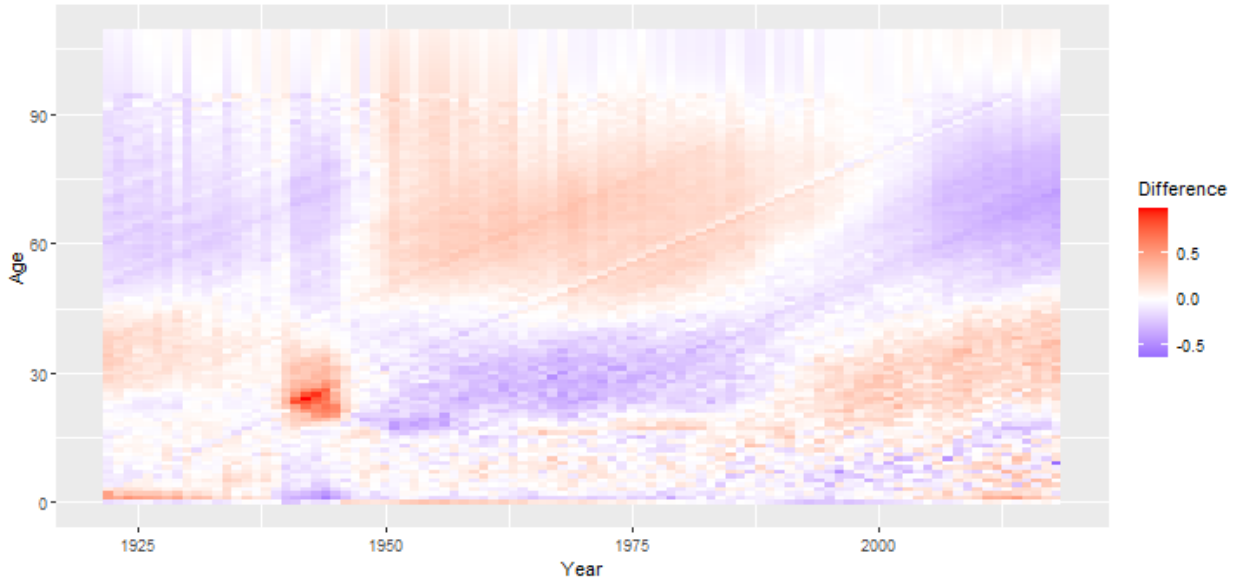


Figure 5 shows the average actual and fitted $log(m_x)$ for all ages in each year. This allows recognition of which years are generally under/ over fitted.

An understanding of the error value of the Lee-Carter model can be gained by comparing the the original values of $log(m_x)$ from the life table with the estimated values from the model. Figure 6 shows a heat map of the difference in the actual and fitted values of $log(m_x)$.

Figure 6: Heat map showing the difference in Actual and Fitted values of log(mx)



The heat map shows which ages and years have been under/ over fitted.

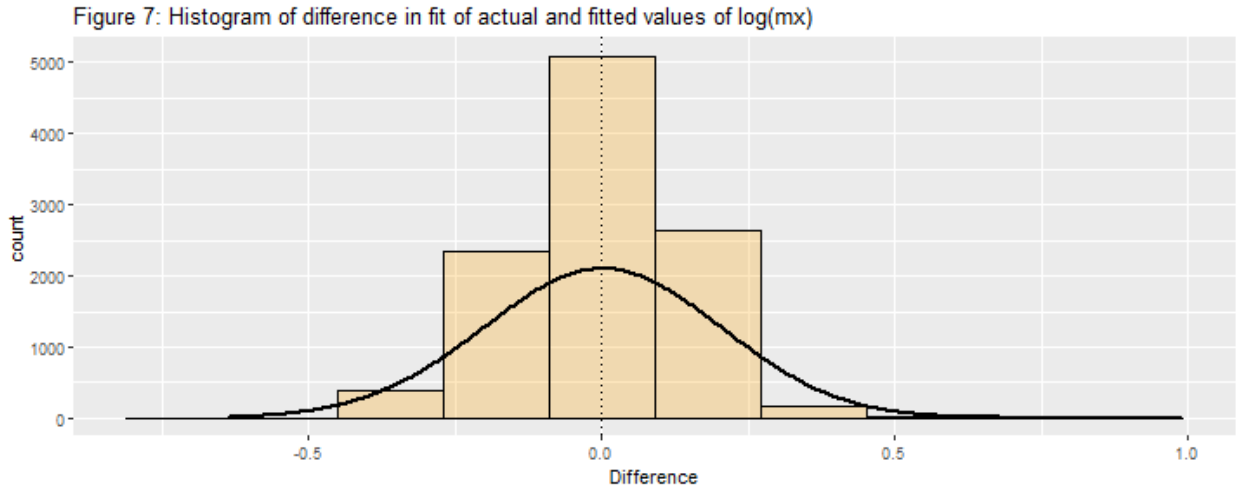
From 1920 until 1938, newborn mortality has been overfitted, there does not seem to be much fitting error for childhood mortality, ages between 28 to 45 have been over fitted older ages and ages older than 45 are generally under fitted.

There is a large cluster during 1939 to 1945 between ages 20 and 35 in which $log(m_x)$ has been underestimated. This is actually expected as this period covers World War 2 which can be considered an anomalous period.

Conversely from the earlier years, from 1945 until 1990 there is generally underfitting in ages 20 to 40 and overfitting in ages over 45. This reverses again from around the year 2000, ages 15-40 are generally overfitted and ages above 45 are underfitted.

The diagonal lines that are present in the chart suggest that $log(m_x)$ is subject to the cohort effect. This is interesting because it implies that there might be a lingering impact on $log(m_x)$ from the year that a person is born. This defies the assumption that the year that a person is born does not have an effect on $log(m_x)$ which is made in the Lee-Carter model.

A sense of the distribution of the error can be gained by plotting a histogram of these differences. This is shown in figure 6.



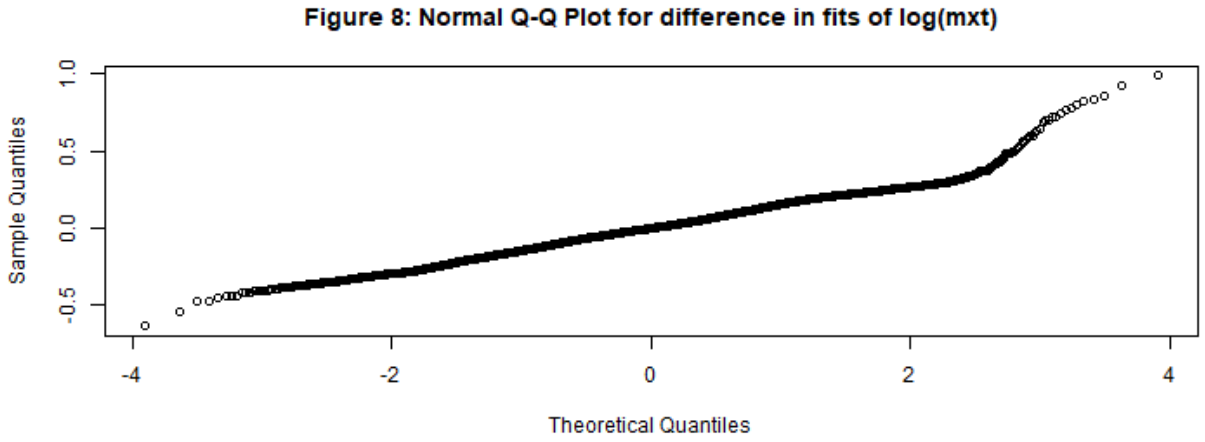


Figure 6 suggests that the error of the Lee-Carter model follows a normal distribution with mean approximately 0. The standard deviation can also be calculated through the formula : $\sigma = \frac{\sqrt{\sum_{i=1}^{N}(x_i - \mu)}}{N}$

The standard deviation of the sample is 0.15 (2 d.p.). Furthermore, we can conclude that the error value is distributed:

$$\epsilon_{xt} \sim N(0, 0.15^2)$$

2.5 Monte-Carlo simulation for v(t)

A Monte-Carlo simulation is a term used to describe repeated random sampling in order to obtain values that are not in the sample. A Markov Chain is a type of Monte-Carlo simulation that simulates a random walk through a probability distribution in order to predict future results.

A Markov Chain Monte-Carlo simulation can be used to forecast future values for $v(t)$. Values for both $u(x)$ and α_x are not dependent on time t, therefore by forecasting future values of $v(t)$ then allows estimation of future values of $\beta_x K_t$. More importantly, estimation of future values of $\beta_x K_t$ and usage of the previously estimated α_x allows estimation of future values of $log(m_{xt})$. This forecasting process is made viable through the assumption in the Lee-Carter model that α and β values can be considered as constants over time (as mentioned previously) meaning that they do not differ for future values of t.

To create a Markvov-Chain Monte-Carlo simulation, a random sample is required. Assigning a new variable D_v: that represents the difference in v at time t and $t-1$ (step size) as the random sample enables simulation of the difference in v_t for each future time t. The following formula can be used to describe D_v:

$$D_v = v(t) - v(t-1)$$

An assumption of the Lee-Carter model is that the differences in K_t (step size) can be considered as a normally distributed random variable (previously mentioned). This implies that D_v is also normally distributed and therefore D_v can be considered:

$$D_v \sim N(E(d_v), \sigma^2_{d_{v(t)}})$$

where $E(d_v)$ represents the expected value/ mean of d_v and is calculated $\frac{\sum_{t=1}^{n_t} d_{v(t)}}{max(t_n)}$, $\sigma^2_{d_{v(t)}}$ is the variance of d_v and can be calculated $\sigma^2 = \frac{\sum_{i=1}^{N}(d(i)-\mu_d)}{N}$

Using this distribution enables a random walk from the final time t which is the year 2019 for the next m years. Setting $m = 20$ produces a random walk for the next 20 years. Reproducing this process 10,000 times creates a sample of 10,000 random walks of $v(t)$ which can be visualised in figure 7:

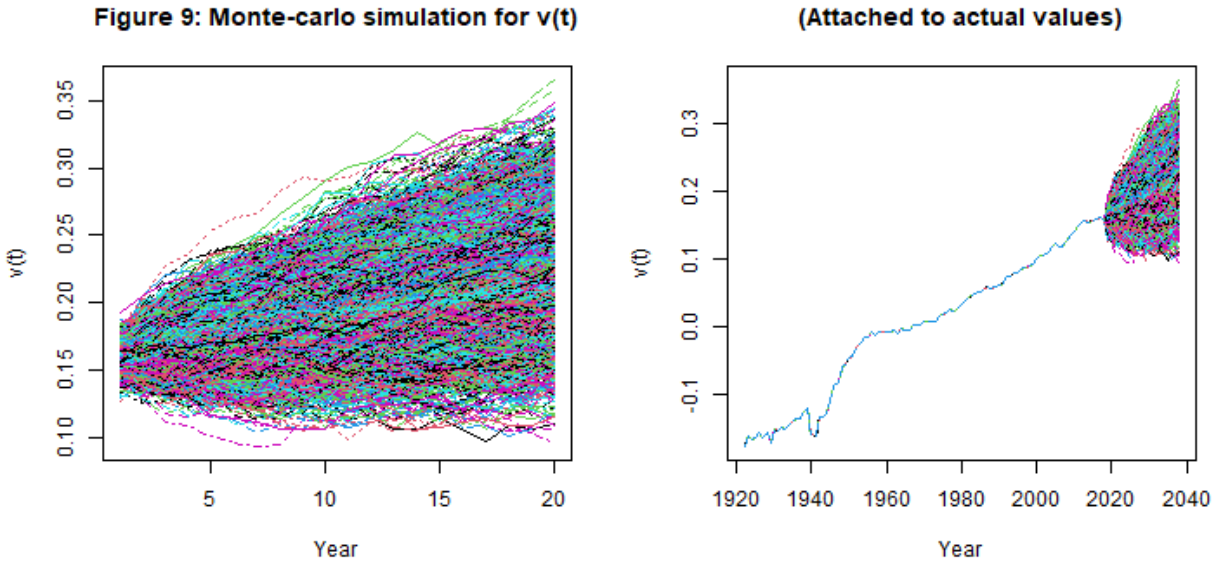


Figure 9 shows 10,000 random walks of $v(t)$ for the next 20 years. Looking at the random walks creates an idea of the maximum and minimum values of $v(t)$ for the next 20 years but does not create an accurate impression of the most likely path of $v(t)$ over the next 20 years. Through taking all values for each year of the random walk and calculating the median and percentiles for each year allows an accurate representation of the path $v(t)$ will take over the next 20 years.

2.6 Analysis of v(t) forecast

The 50th percentile (median) of each forecasted year of $v(t)$ provides an idea of the average path that $v(t)$ can take. Taking both the 5th and 95th perecentiles for each forecasted year of $v(t)$ creates a 90% confidence interval around this average path.

Figure 10: Median of v(t) shown with 5th and 95th percentile ribbon

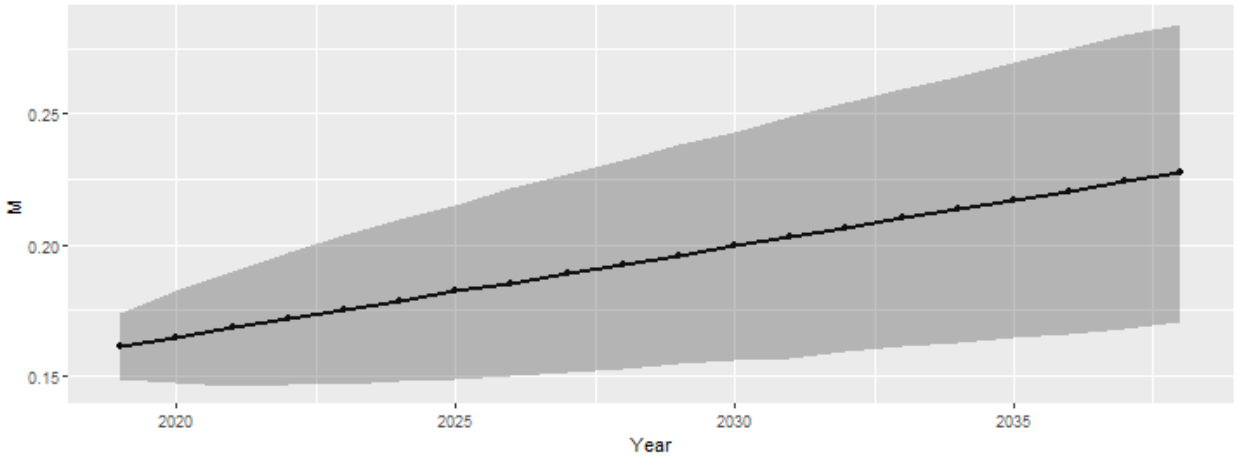


Figure 11: Median of v(t) forecast with 5th and 95th percentile shown with actual data

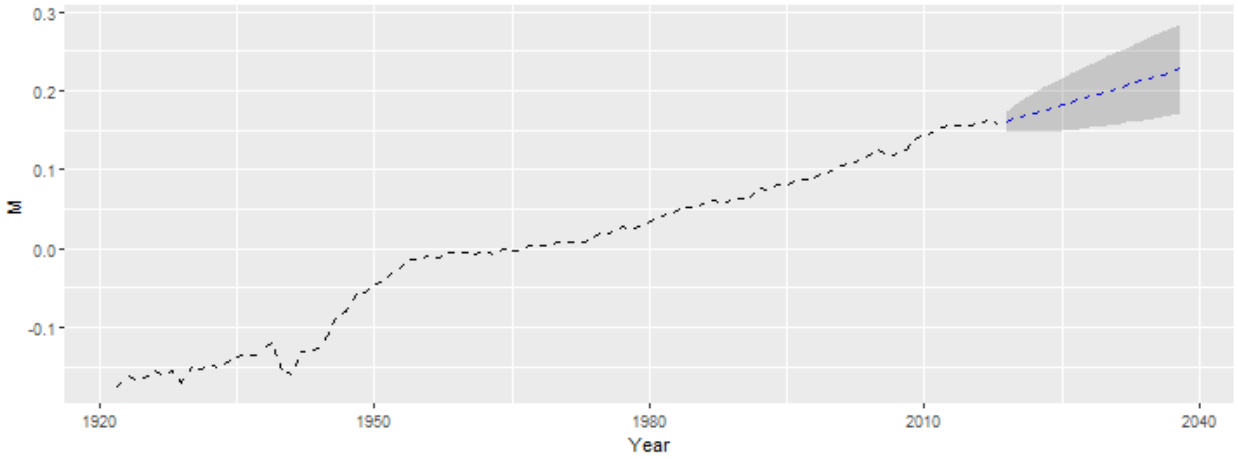


Figure 10 and 11 indicate that $v(t)$ will follow a upwards trend for the next 20 years. The median value represents the average path that $v(t)$ will follow. The shaded region represents the area between the 5th and 95th percentiles which implies that it can be assumed with 90% certainity that the path of $v(t)$ will lie within this region. It can be observed that the interval widens as time t increases, this indicates that there is a larger range of values that $v(t)$ can take with 90% certainty as time t increases. These percentile values computed from the Monte-Carlo simulation can be applied to create a forecast of $log(m_{xt})$.

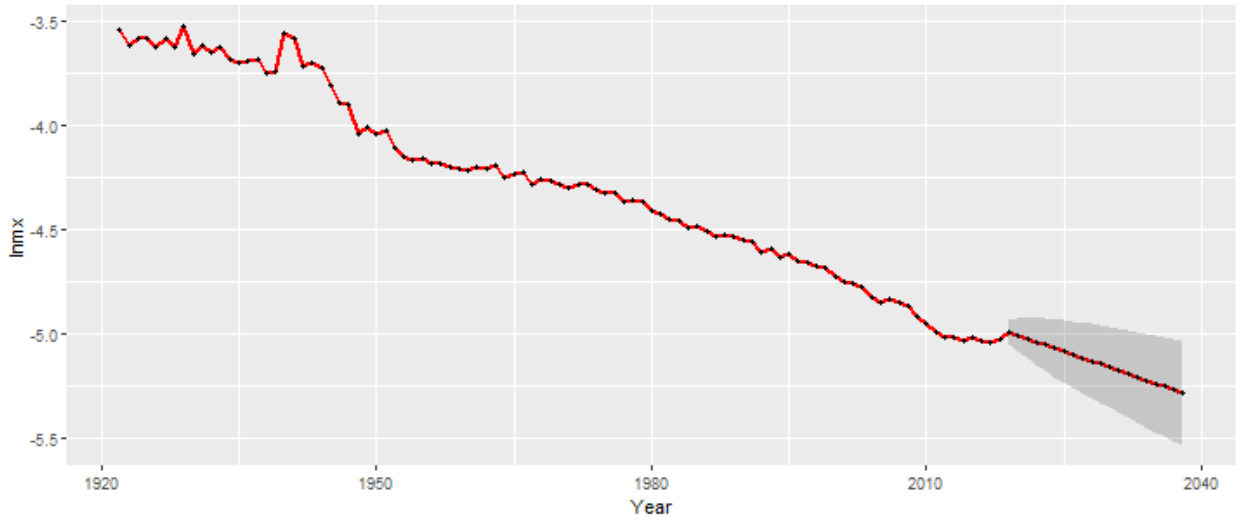
2.7 Forecast of Central Death Rate

Using the prior simulation of $v(t)$ values for the next 20 years, allows $log(m_{xt})$ to be estimated for the next 20 years with the following formula:

$$log(m_{x(t+n)}) = \alpha_x + S_1 u(x) v(t+n)$$

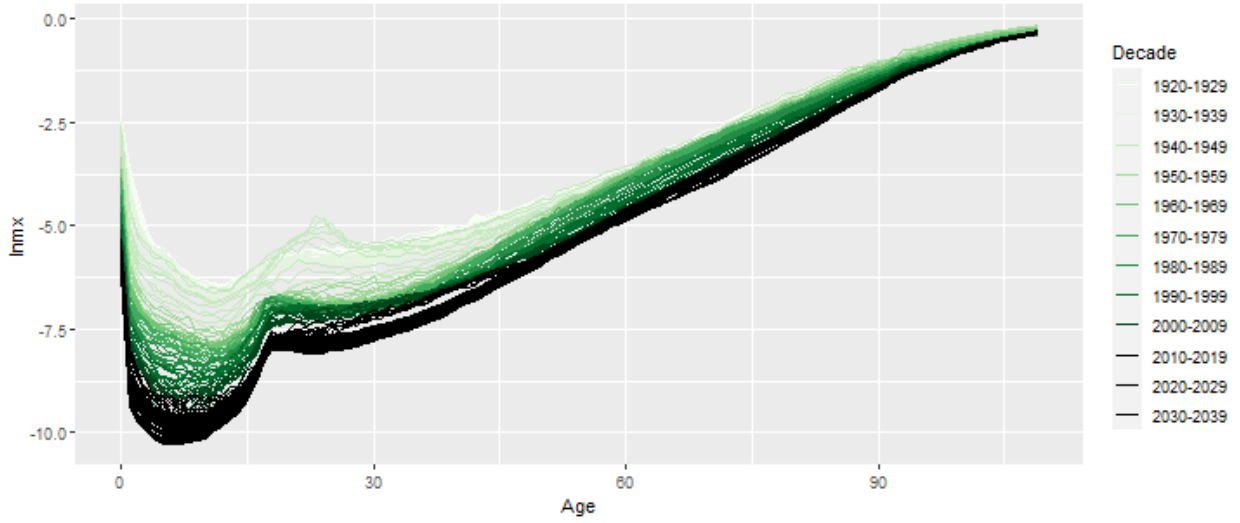
where n represents the forecasted year subtracted by 2018 (max value of t)

Figure 12: Average log(mxt) each year for all ages with forecast



Unsuprisingly, the forecast for $log(m_{xt})$ suggests that it will follow the same downwards trend that it has done since 1920. The shaded area around the forecasted region is the 90% signficance interval which was computed using the 5th and 95th percentiles from the $v(t)$ forecast. This region allows recognition of the range of values which $log(m_{xt})$ is most likely to fall within over the next 20 years. The surface area of the shaded interval increases as time increases (interval widens), implying that there is less certainity in the forecast as time increases.

Figure 13: Change in log(mx) by age in each year with forecast, displayed by decade



$log(m_{xt})$ follows the same trend in the forecasted years than the previous years in terms of its relationship with age. There are high levels of newborn mortality, followed by low levels in childhood mortality which rapidly increases to age 18 which stays relatively constant until age 30 and then $log(m_{xt})$ continues to increase at a steady rate as age increases. As time increases the level of mortality in each age group declines which is present since the 1920s.

The results from both fitting the Lee-Carter model and forecasting allow understanding of how the Lee-Carter model works and the capabilities of the model. It can be recognised from fitting the Lee-Carter model showed that the Lee-Carter model can provide fair estimates for $log(m_{xt})$ and that a general comprehension of future values for $log(m_{xt})$ can be devised through Markov Chain Monte-Carlo simulation.

3 The Lee-Carter Model applied to UK mortality data by location

This section will focus on the second aim of the project: adapting the Lee-Carter model to incorporate local areas within the UK rather than age groups.

Now that an idea of how the Lee-Carter model is constructed and can be applied has been obtained through the previous section, the Lee-Carter model can be adjusted to consider local areas within the United Kingdom instead of age groups. The Office of National Statistics (ONS) publishes UK mortality data each year which provides the number of deaths in each Unitary Authority (UA). ONS Unitary Authorities are groupings of local authorities and are based upon groupings of city councils, borough councils, county councils, or district councils and are a general representation of the local communities in the UK.[9]

The data obtained for this section is a monthly mortality dataset for each Unitary Authority in 2018. The data only covers one year of mortality and therefore we consider time t as intervals of one month. A limitation of the data that was not present in the data used in the previous section is the limited number of time intervals (12 months rather than the 97 years of data) and additionally monthly data is often subject to seasonality. However, the aim of this section is to gain an understanding of how the Lee-Carter model can be adapted to incorporate local areas instead of age. Furthermore, these factors should not be considered an issue but instead acknowledged when analysing the results of the model.

The ONS provides population statistics for each Unitary Authority in the UK with populations estimated from the UK census. The population estimates for 2018 are also used within this section.[10]

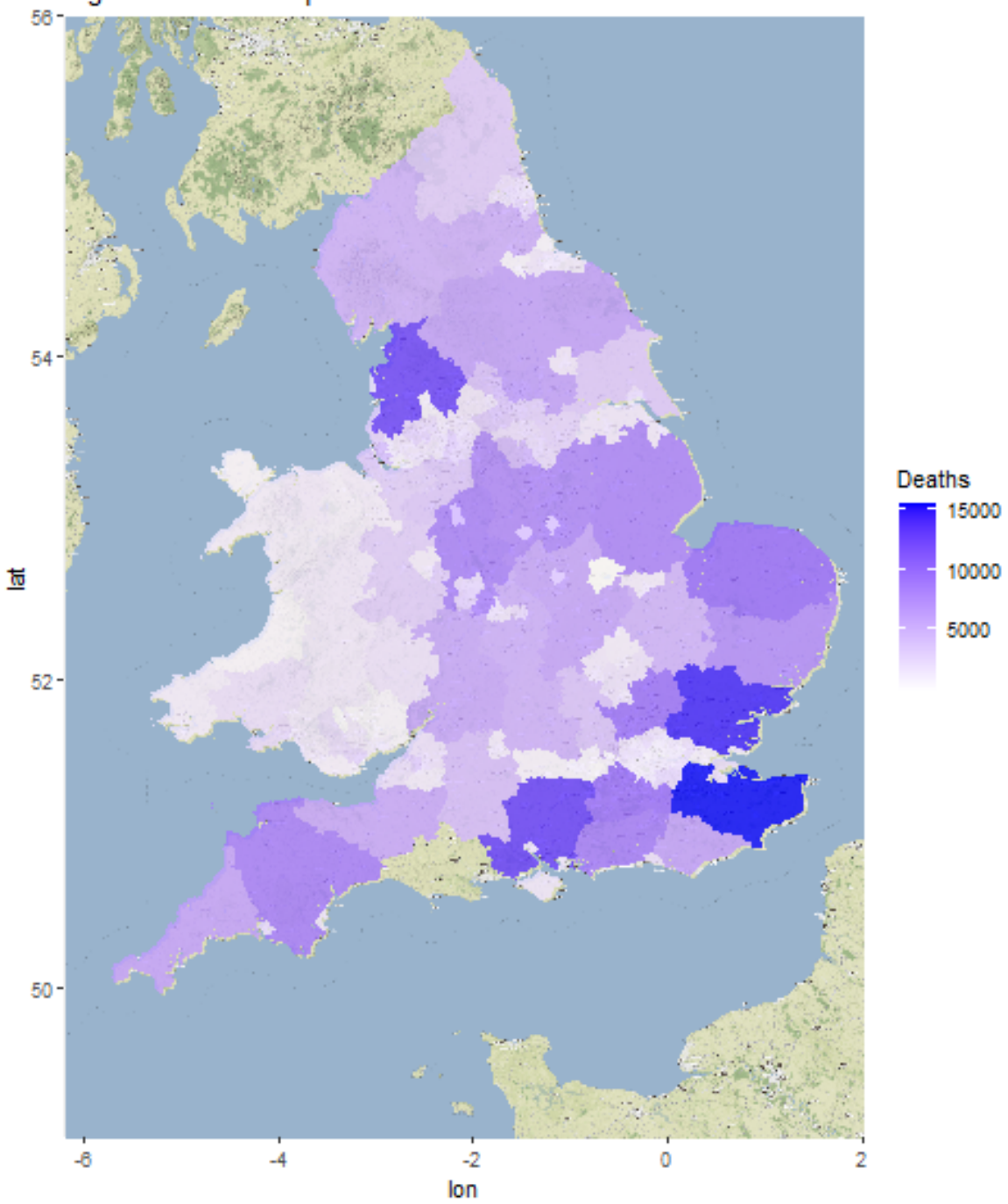
NOTE: ONS datasets did not link for two unitary authorities due to boundary changes and redefinitions of the geographies during the period in which the different datasets have been extracted. Therefore, Hertfordshire has been excluded from map visualisations and Bournemouth has been excluded entirely.

3.1 Descriptive analysis of data

Using ONS data allows geographic visualisation of UAs through spatial plots using ONS published shape files. These shape files can be read by R and adjusted accordingly to produce geo-spatial plots.

Below displays a heat map of the number of deaths within 2018 in each UA:

Figure 14: Heat map for number of deaths in 2018



Looking at the map (figure 14) suggests that there are no signifcant geographic clusters of multiple UAs with high numbers of deaths. There seems to be a high variation in the number of deaths within each UA as indicated by the wide range of total deaths within the year. The high range in the number of deaths in each UA suggests that the population sizes within each UAs vary. Furthermore, deaths should be considered as a rate, as it is in the Lee-Carter model, going forward to avoid population sizes affecting analysis.

3.1.1 Calculation of central death rate

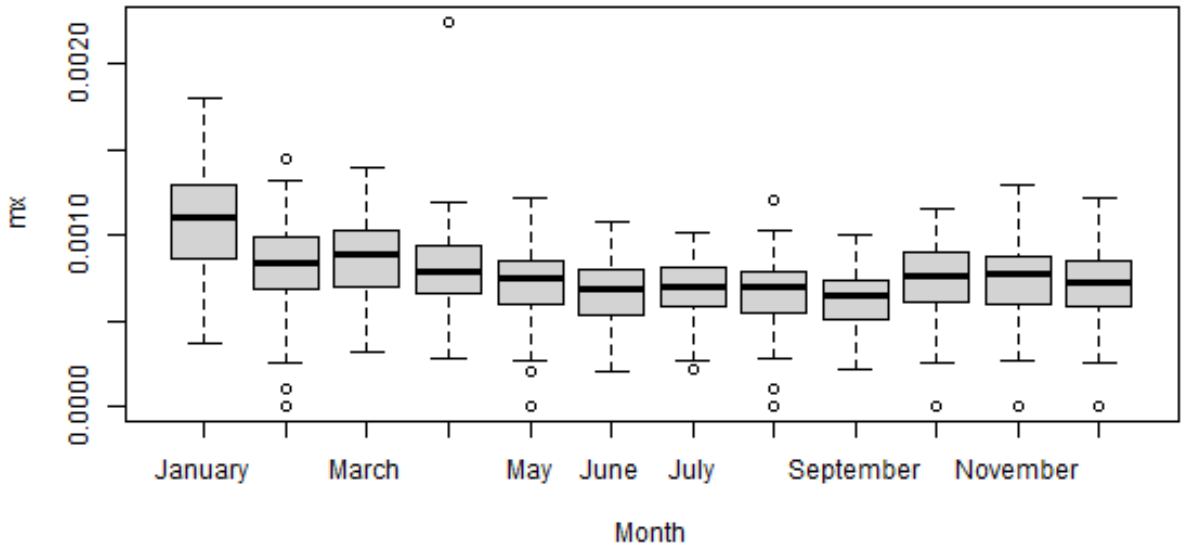
In the previous section the central death rate mxt was provided in the UK life tables. The ONS mortality data for 2018 contains raw counts of mortality, therefore, the central death rate has to be calculated. Due to location being a different variable to age, the process in calculating central death rate in terms of location L (m_{Lt}) can be adjusted to simply be:

$$m_{Lt} = \frac{L_t}{n_L}$$

where L_t is the number of deaths in the population in location l in month t and n is the population size in location l.

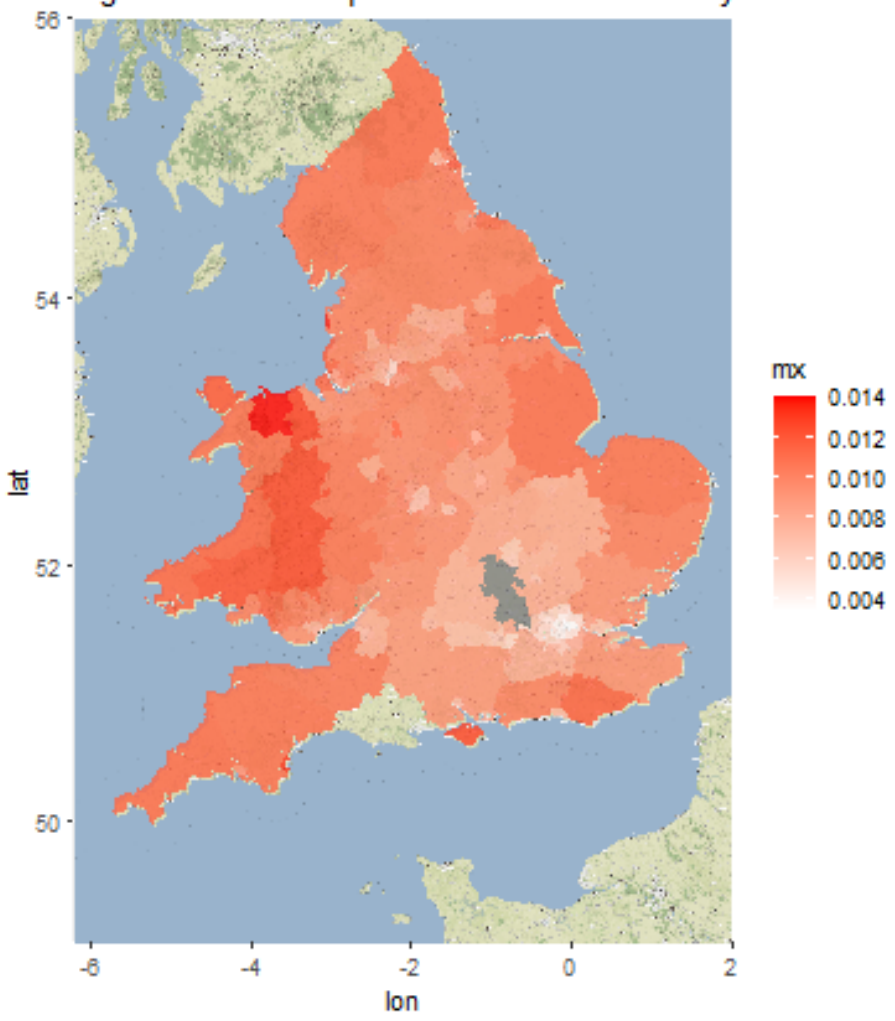
This derivation of central death rate functions due to there being no large fluctuations of deaths and population within a singular location like there is with age groups (as age groups shift as people become older). An assumption made in this process is that the population in each local area remains constant throughout the year; the relative impact of changes in the population size (through people entering and leaving the population) can be considered as negligible in this scenario.

Figure 15: Boxplot to show central death rate over time by UA



The boxplot suggests that the highest levels of mortality in each UA is usually seen in January and that $m(x)$ follows a general downwards trend until the mortality rate reaches its lowest level in September before increasing and remaining relatively constant from October through till December. A fair assumption to be made is that mortality is generally higher in winter compared to summer. This chart also creates an idea of what K_t (the trend in mortality over time) will look like in the model.

Figure 16: Heat map for Central Death Rate by ONS Unitary Authority



Observing the heat map for m_{lt} in the UK by Unitary Authority, it is noticable UAs in the North/ midlands of Wales do seem to have higher central death rates and UAs in London and surrounding areas have lower central death rates compared to the rest of England. However, there is no strong evidence of geographical significance between central death rates and UAs.

3.2 The Lee-Carter model with local area instead of age

In the Lee-Carter model, age x is considered as a categorical variable (unordered) rather than a numerical variable as an assumption of the model is that $log(m_{xt})$ can differ for each age x and time t. This suggests that location L, another unordered categorical variable, can be substituted for x.

Replacing the age group x variable with location L in the Lee-Carter results in no changes to the derivation of the logarithm of central death rate:

$$log(m_{Lt}) = \alpha_L + \beta_L K_t + \epsilon_{Lt}$$

where α_L is the force of mortality in each location, β_L is the change in mortality from each location and K_t is the trend in mortality over the year (by month).

3.3 Estimation of variables

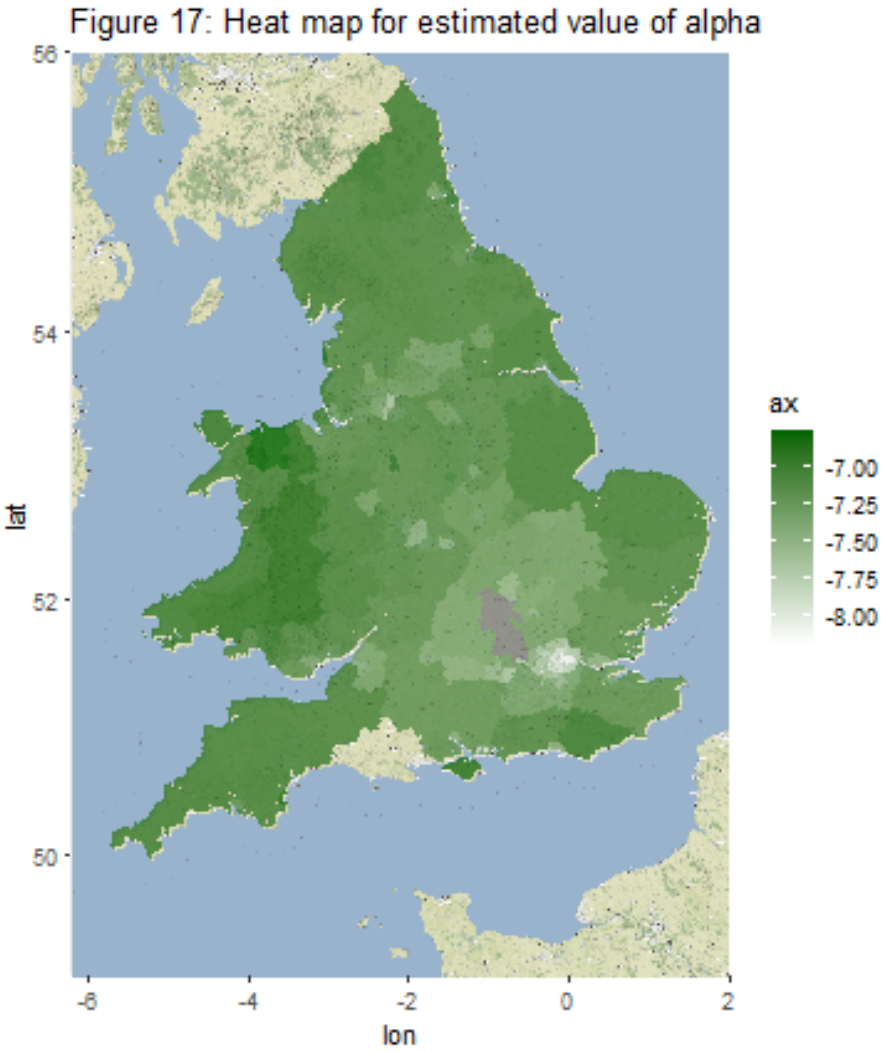
The estimation of α_L is similar to the estimation of α_L in the original Lee-Carter model. α_L is the force of mortality in each UA rather than age group, therefore the formula is adjusted is adjusted to incorporate location instead of age:

$$\alpha_L = \frac{\sum_{t=1}^{n_t} log(m_{Lt})}{n_t}$$

where $log(m_{Lt})$ is the sum of the log of central death rate for each year in each location in month t and n_t is the length of the time period, therefore 12 (months) in this instance.

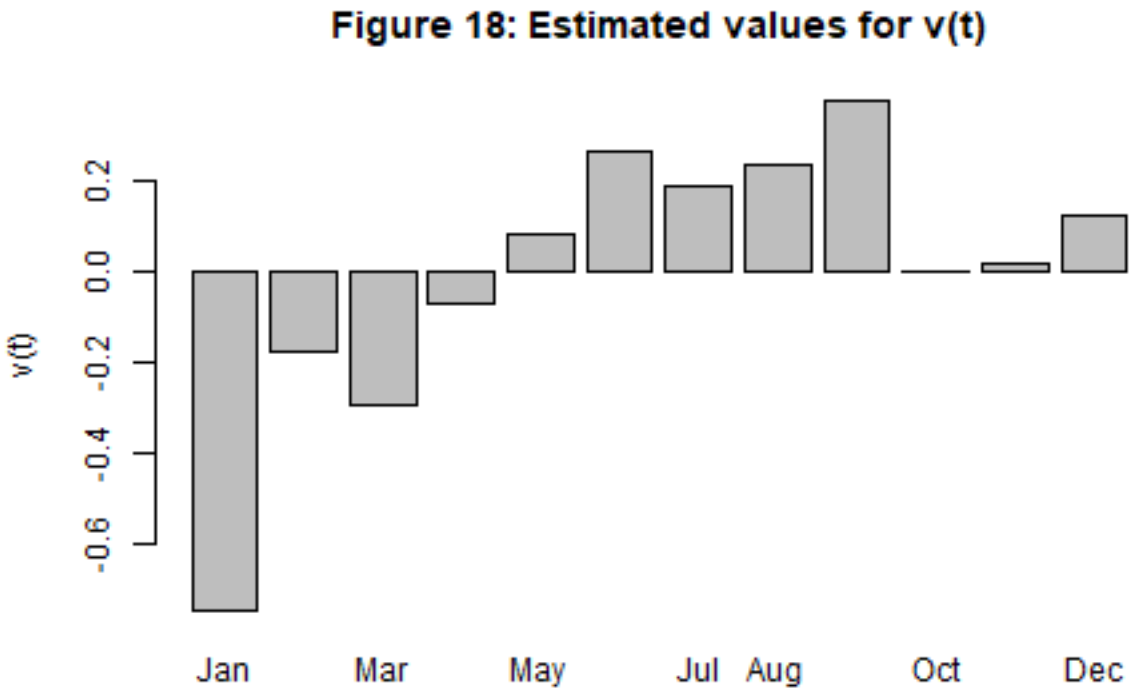
The estimation of β_L and K_t also follow the same process as the estimation of β_x and K_t with location L replacing age x:

$$log(m_{Lt}) - \alpha_L = \beta_L K_t = S_1 u(L) v(t)$$

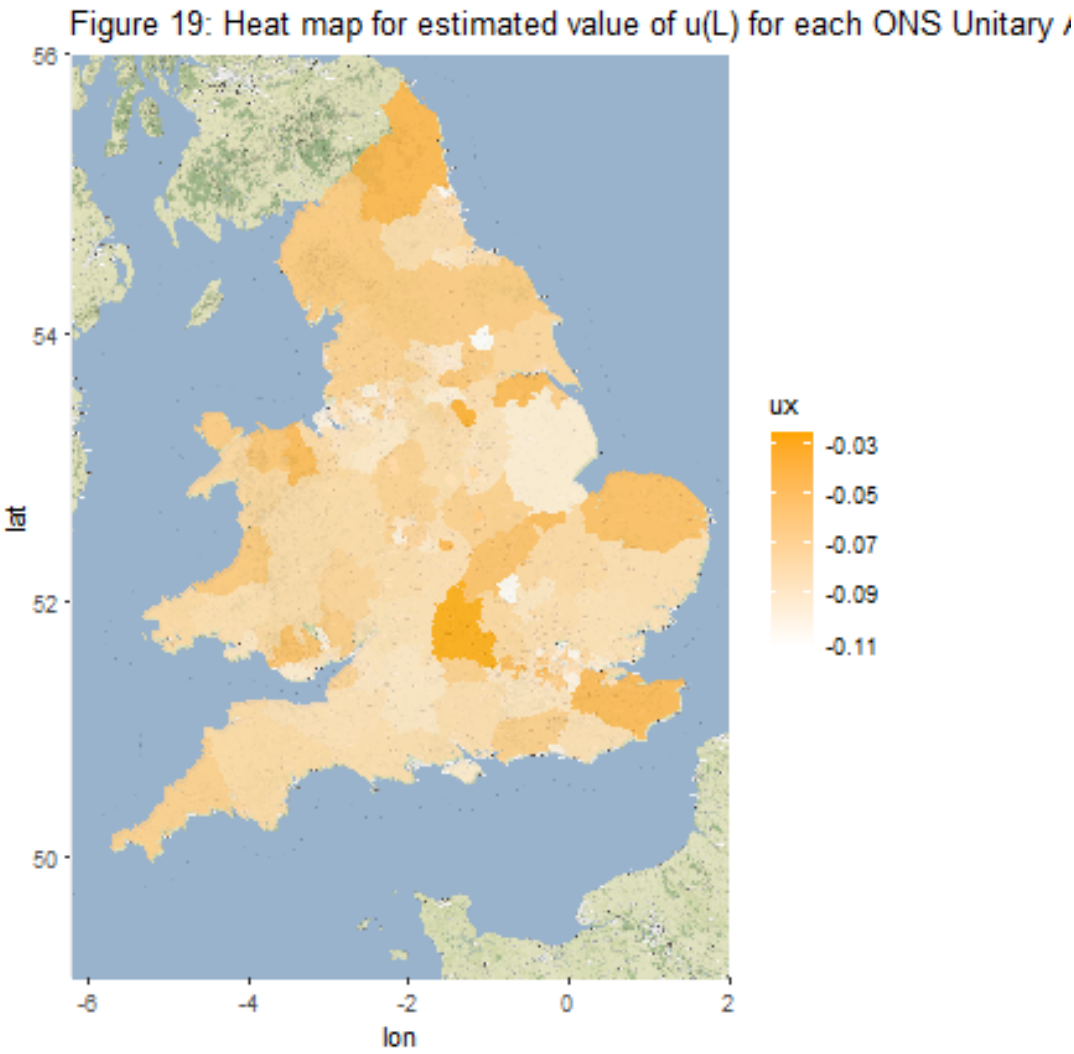


The map above visualises the α_L value for each location. The map plots for both the estimate of α_L and m_L suggest a very strong correlation - which is as expected as α_L is a function of m_L. Figure 17 also provides an initial indication to the variety within the model.

The process for estimating K_t and β_L is the method as the original Lee-Carter model too; the matrix $B_L K_t$ can be factorised into three matrices, $u(L)$, $v(t)$ and S_1, through SVD.



$v(t)$ follows a general upwards trend from January to September. Then dips to approximately 0 in October and slightly increases each month until December. The clear trend shown by the $v(t)$ curve indicates that there $log(m_L)$ is subject to monthly seasonality. This is not an issue with yearly data but will restrict the forecasting potential of this model.



$u(L)$ represents the change in $log(m_{Lt})$ in any given month for each location. The heat map for the $u(L)$ value in each location shows little correlation with the previous heat maps for α and m_L.

3.4 Fitting the model

The estimated values can now be used to calculate fitted values for $log(m_{Lt})$ using the formula:

$$log(m_{Lt}) = \alpha_L + S_1 u(L) v(t)$$

This allows comparison of the actual and fitted values for $log(m_{Lt})$:

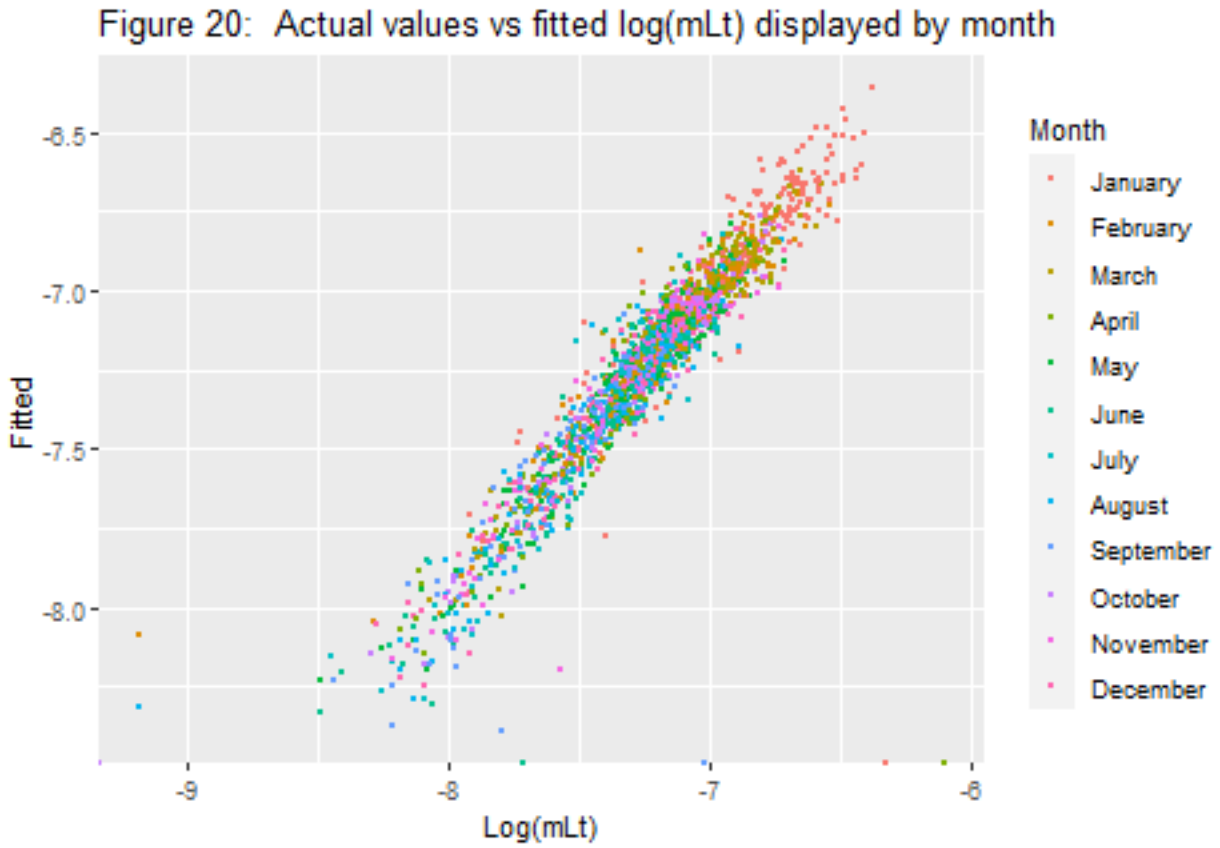
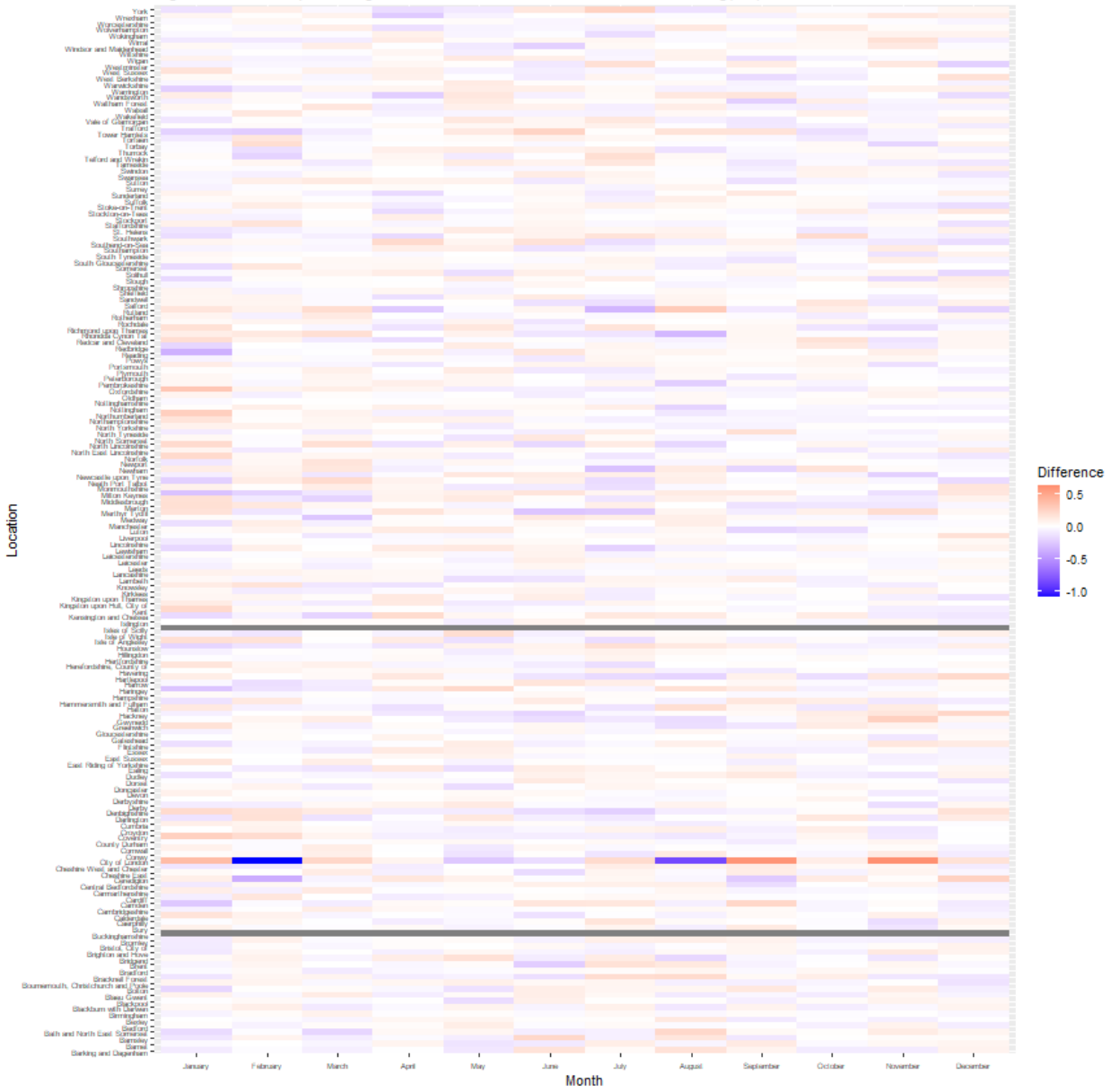


Figure 20 shows the fitted $log(m_{Lt})$ values against the actual values for $log(m_{Lt})$. There are visible clusters of months with the earlier months (January to March) taking higher values for both fitted and actual $log(m_{Lt})$ and the later months (October to December) taking lower values. Also noticable is a higher variation within the later months than the earlier months. This indicates that the model is better at fitting in these earlier months. Furthermore, figure 20 reinforces the issue of seasonality present in using time t as periods of months as there is seasonal differences in the values of $log(m_{Lt})$ and the variation of fits.

Figure 21: Heat map showing the difference in Actual and Fitted values of log(mx)



The difference between the actual and estimated values for $log(m_{Lt})$ are visualised in figure 21. The City of London shows significantly higher differences between the actual and fitted $log(m_{Lt})$ values. A clear issue present that limits analysis the difference in fits is the large number of UAs being analysed and the lack of clear relationships that can be analysed from using location instead of age in the model.

The average value for the residuals is NA which suggests that, although the model has its flaws, it can produce an acceptable fit for $log(m_{Lt})$.

The large number of unitary authorities in the UK creates the idea that there are relatively low population sizes in each UA. Which may potentially be decreasing the accuracy of the model. A large defining factor of each UA is the geographical location of the population, which does not have a large impact on $log(m_{Lt})$ as shown previously. Furthermore, a more significant feature of UAs, regards to $log(m_{Lt})$, may need to be considered in order to produce a more accurate model.

The data observed is for a one year period and there is clear seasonality present in monthly data. This indicated that a Monte-Carlo simulation cannot be computed as each month/ season would need to be

considered seperately and there is only one observation for each month in each location. This limits the conclusions that can be drawn surrounding this adaptation of the model as a forecast for $log(m_{Lt})$ cannot be assessed.

This adaptation of the Lee-Carter model faces various issues as a result of the lack of relationship between each location and lack of geographical relationships with $log(m_{Lt})$ as well as lack of support from the data.

There are a lack of strong conclusions that can be drawn in this section. However, the aim of this section has been met as the Lee-Carter model has been adapted succesfully to include local area rather than age. The selection of the aim for this section was motivated by the purpose of understanding the impacts of incorporating local area into the Lee-Carter model rather than constructing a 'perfect model' for mortality. Furthermore, the weaknesses recognised in the modelling process, analysis and data can be accounted for in the next section.

4 Constructing an extension to the Lee-Carter model

The final aim of the project is to propose, test and apply an extension to the Lee-Carter model to consider all age, time and local areas. The previous two aims looked at modelling age and local area seperately against time. The ONS defined unitary authorities have been considered as the local areas of the United Kingdom throughout this project as they provide a fairly accurate representation of the geographical seperations of different communities within the UK. In order to extend the Lee-Carter model to include local areas, similarities between each UA need to be considered before modelling.

4.1 Identifying clusters of local areas

Cluster identification poses three main questions:

1) Why is cluster identification necessary?

2) What characteristics should clusters be identified on?

3) How to identify clusters?

4.1.1 Why is cluster identification necessary

Cluster identification is necessary in this scenario because there are 170 different unitary authorities, each with varying population sizes, that are contained in the data. Unitary authorities provide a strong indication of geographical seperations between different communities. The geography between these communities is not an aspect that is going to be considered in this extension of the Lee-Carter model. Instead, local areas which have similar characteristics within their populations should be considered in the same population.

Additionally, modelling each of the 170 UAs in the UK as seperate populations runs the risk of overfitting as there are relatively small population sizes within each UA. Furthermore, to produce an accurate stochastic model to forecast and predict mortality, the number of local areas considered should be grouped which would consequently increase the sample size for each of the groups considered.

4.1.2 What characteristics should clusters be identified on?

The extension to the Lee-Carter model is anticipated to consider both local area and age. This implies that local areas with similar age distributions should be considered when grouping local areas. Therefore, a variable that best describes the age distribution within each local area should be considered as the variable that clusters are identified on.

The variational coefficient measures the spread of the standard deviation in relation to the mean and can be represented with the formula:

$$VC_{xL} = \frac{\sigma_{xL}}{E(xL)}$$

where VC_{xL} represents the variational coefficient of age in each location, σ_{xL} is the standard deviation of age in each location, and $E(xL)$ is the mean/ expected age of the population in each location.

By taking the variational coefficient of age in each location, a representation of the age distribution for each location can be acknowledged. Using the variational coefficient of the age within each UA allows recognition of which UAs have similar age distributions (clusters). The purpose of this approach is to enable identification of UAs with similar age distributions to allow for a bigger sample size when modelling and forecasting central death rate which will result in smaller confidence intervals (more certainty of forecast).

There was various other variables taken into account in the cluster identification process, for example: clustering based on region (North, South, South East, etc) or even mortality rate. However, clustering on age distribution seemed the most relevant approach because the original Lee-Carter model specifically takes age into account. By applying age distribution based groupings will present groups of UAs with similar age distributions to be considered as a singular populations. Additionally, each group will have an age distribution that can be identified clearly and therefore allows understanding of the variety in age distributions of UAs in the UK.
How to identify clusters? (classification tree)

A classification tree is a non-stochastic method for identifying groups through recursive partioning.[2] Classification trees are a form of CART (classification and regression tree) modelling and can be used in various contexts. In this scenario, a classification tree will be used to approximate decisions based on 'variable X' that splits the data into clusters that can be identified at a 95% significance level. The 'Classification Tree Example' describes the general premise of how the classification tree splits the data into groups.

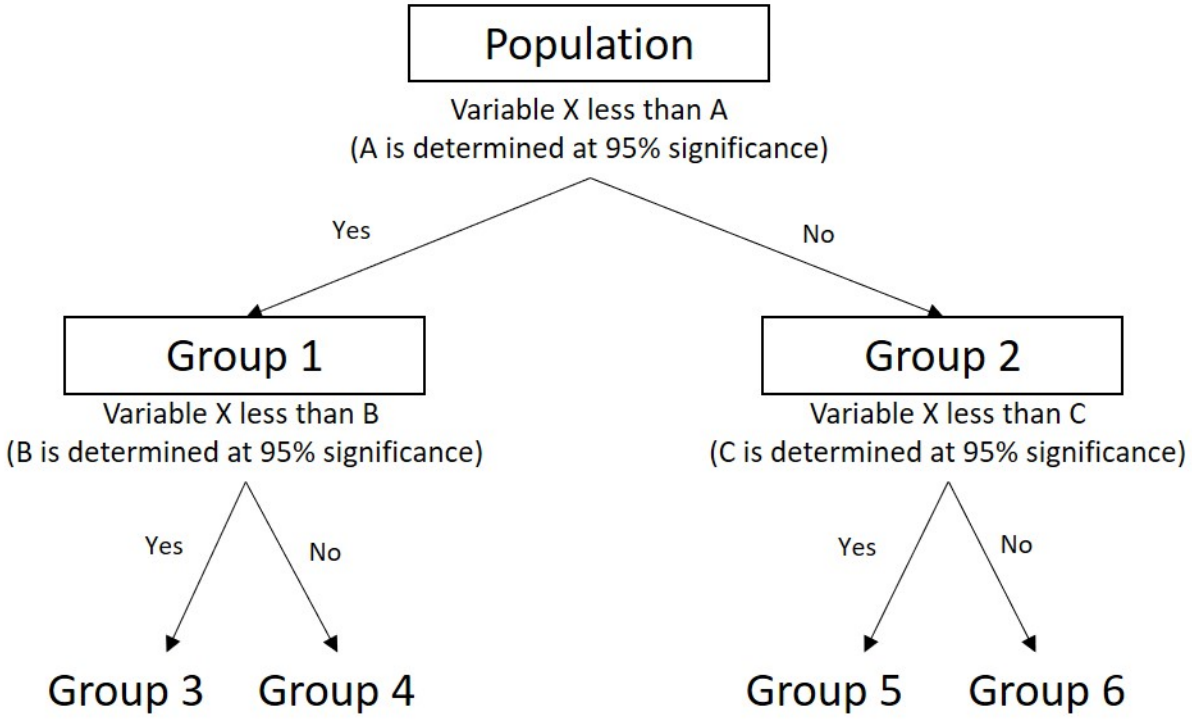


Figure 1: Classification Tree Example

The reasoning behind using a classification tree was based on the understanding that classification trees have the capability of interpreting both numerical and categorical variables. Classification trees also provide an easy to understand visualisation as an output. However, a disadvantage of this approach which should be noted is that small variations within the data can have large impacts on the results, therefore small population changes will require a new tree to be produced.[3]

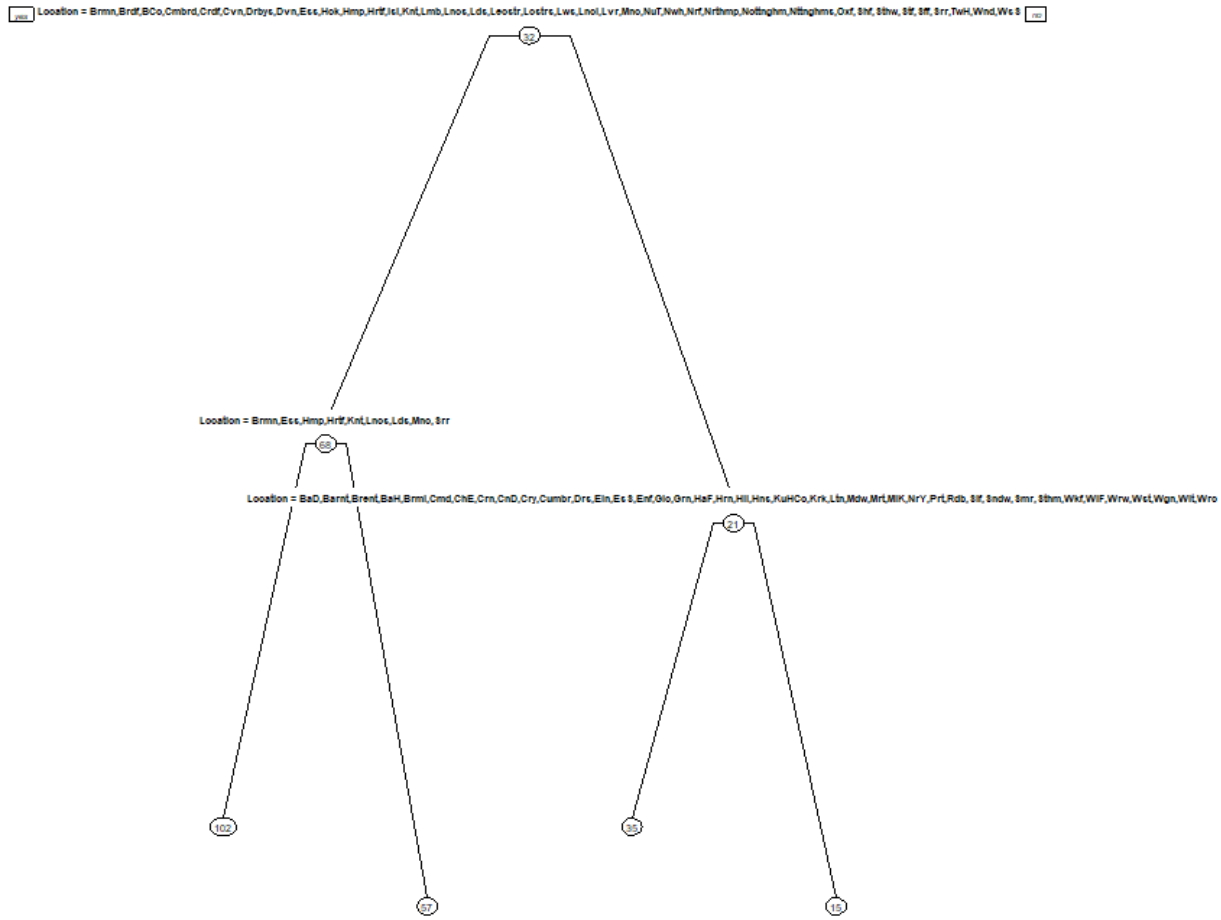
4.2 Applying a Classification Tree

The aim of applying a classification tree to the data is to identify clusters of UAs with similar age distributions by using the variational coefficient of age within each location.

The classification tree approach also enables identification of the group in which a 'new' unitary authority would belong in based upon the variational coefficient (standard deviation divided by mean age) of the proposed UA.

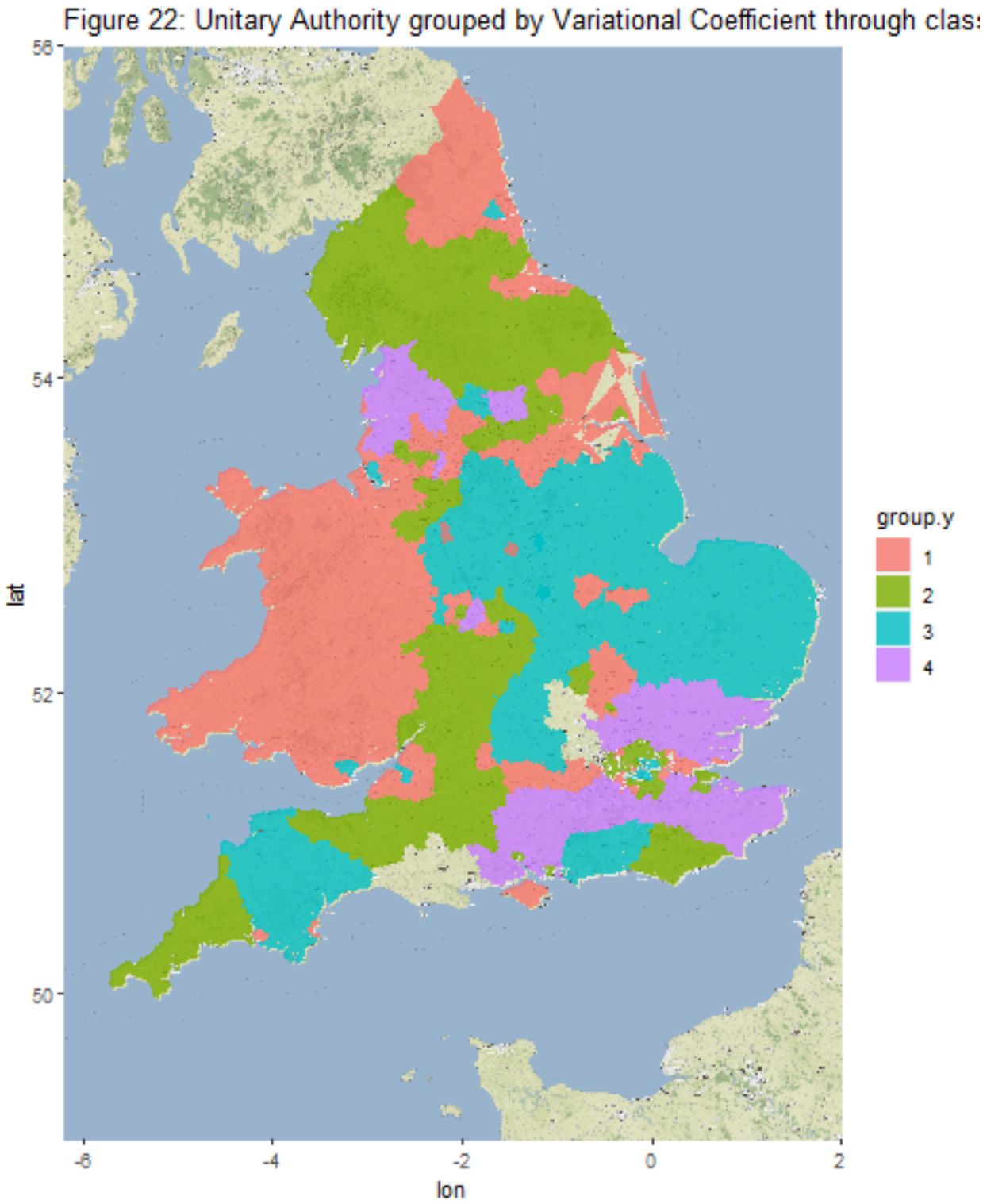
The dataset has to be optimised to produce a classification tree through formatting the dataset with onehot encoding, this means that each location has to be considered as a individual response variable with 0 or 1 values.

Applying a classification tree which regresses variational coefficient on unitary authority with a 95% significance level, the following tree is obtained:



Four individual clusters can be identified from the classification tree. This indicates that there are four clusters of UAs that have significantly different age distributions (represented by variational coefficient) at a 95% significance level. Furthermore, this will enable grouping of the 171 unitary authorities into four seperate populations.

These clusters can be visualised geographically using a map:



The map indicates which Unitary Authorities belong into which cluster (group). As the classification tree is based on variational coefficient we can understand which areas have similar age distributions on a geographical level. The groups are ascending from lowest variational coefficient to highest. Interestingly, most of Wales (apart from the Vale of Glamorgan) and surrounding areas fall into the same group. England shows a wider variety of clusters. The majority of the East Midlands as well as Central London fall into group 3. Many Outer London boroughs fall into group 3, whilst most of the UAs that surround London fall into group 4. Although, there does seem to be a strong spread of these age distributions accross England as all four groups are present in English UAs.

The next stage in the process is to construct an extension to the Lee-Carter model that considers these clusters of locations with similar age distributions that have been calculated using the classification tree. Before constructing a model, an analysis of these clusters is needed to understand the affect of clustering on the data and additionally assist the decision making process on how to adjust the Lee-Carter model to account for both age, location and time.

4.3 Analysis of groups

Visualised below are the groupings from the classification tree.

Group 1:

Group 2:

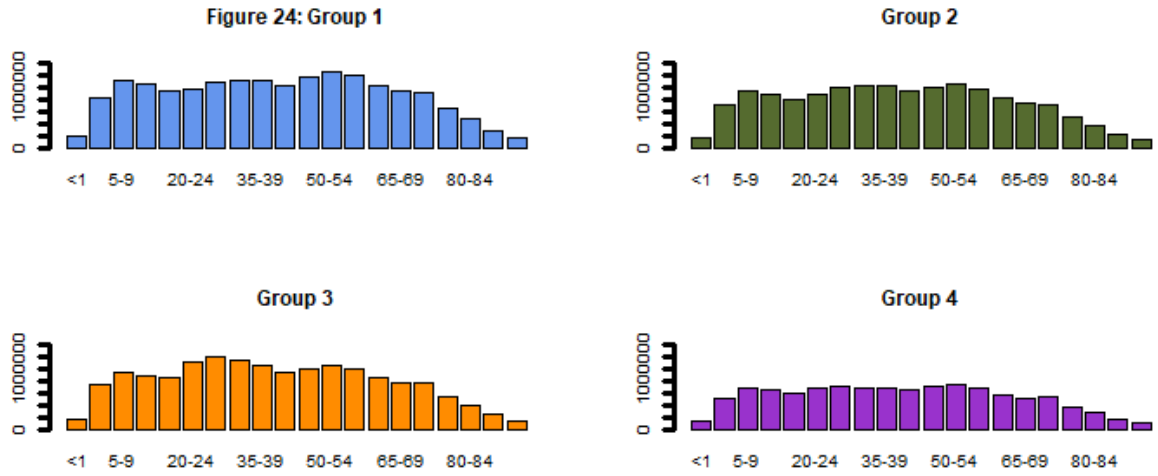
Group 3:

Group 4:

Number of Unitary Authorities classified within each group:

Group	n
1	93
2	41
3	29
4	9

Looking at the names of the UAs in each group, there are some interesting observations that can be made. Group 3 and 4 are the smallest groups (29 and 9 UAs respectively) which contain some of the biggest cities and counties in the UK (Liverpool, Cardiff and Bristol in Group 3, and Manchester, Birmingham and Leeds in Group 4) which have not been split into multiple UAs. Group 3 and 4 also both consist mainly of UAs that contain universities (high student populations). In Group 2, which contains 41 UAs, there is a large number of the London UAs that can be noticed, like Greenwich, Barking, Westminster and Enfield. Group 1 is the largest cluster of UAs and contains 93 different unitary authorities. There is not an obvious relationship between the UAs that can be noticed within Group 1 which suggests that this group represents the rest of the UK (that are not contained in group 2, 3 or 4) and the differences in age distributions for these UAs are not distinguishable at a 95% significance level.



Comparing the age group distributions for the four groups, they all follow a similar shape with a few noticable differences. The first age group contains only one age group (this is due to newborn mortality being a seperate category itself because live-births are usually more common than infant mortality and so should be categorised seperately) and so has a significantly lower frequency. Group 1 UAs contains a higher proportion of the population within the 50-60 year old age bracket, group 2 seems to have a slightly higher proportion of the population between 25-40, the distribution for group 3 has a younger peak than the other groups as it peaks between ages 20-34 and group 4 shows a relatively flat distribution. As mortality is generally higher within older age groups, the expectation would be that group 1 should have the highest mortality rates and group 3 and 4 should have the least.

4.4 Choice of model

4.4.1 Proposals

As there have been four clusters of UAs with different age distributions identified through the classification tree process, they can be accounted for in the modelling process. There are a various potential methods to incorporate the results of the classification tree into Lee-Carter model. However, there are two proposals for models that can be considered as relevant to the aim of this section and the purpose of the Lee-Carter model:

1) Consider the four clusters as an ordered variable (with group 4 representing the youngest age sample ascending to group 1 as the oldest age sample) and take this as the x (age variable) in the standard Lee-Carter model. This is similar process to the model constructed previously which adjusted the Lee-Carter to include location instead of age; in this situation the group would be replacing age:

$$log(m_{gt}) = \alpha_g + \beta_g K_t + \epsilon_{gt}$$

This approach considers all age, location and year. Age and location are represented in the group g variable.

The creation of this approach has been carefully considered to combat some of the main issues present in the adaptation of the Lee-Carter model in the previous section: UAs have been grouped; meaning local area and age can be considered as a categorical variable with an order, there are large populations in each group and there are only 4 groups which allows for groups to be easily compared.

2) Produce an individual Lee-Carter model for each of the four clusters and combine these to produce an estimate for the entire population. Each individiual group will be constructed using the original Lee-Carter model:

$$log(m_{gxt}) = \alpha_{gx} + \beta_{gx}K_{gt} + \epsilon_{gxt}$$

for $g \in (1, 2, 3, 4)$

Combining the four matrices of $log(m_{gxt})$, a model for the entire (national) population can be formulated:

$$log(m_{xt}) = log(\frac{\sum_{g=1}^{4} P_{gx}(exp(log(m_{gxt})))}{sum(P_{1x}) + sum(P_{2x}) + sum(P_{3x}) + sum(P_{4x})})$$

where P is a population matrix of g and x

This model produces a seperate Lee-Carter model for each of the groups, producing four matrices of $log(m_{gxt})$ (for each group). The exponent of is then $log(m_{gxt})$ taken to achieve m_{gxt}. Each m_{gxt} is then multiplied by the population at each age in each group P_{gx} which obtains the total number of deaths for each group g in each age x at time t. These four matrices are all then added together to produce the total number of deaths at each age x at time t for the entire UK. We then divide this matrix by the national population at each x which results in matrix m_{xt}. The logarithm can then be taken to get $log(m_{xt})$.

4.4.2 Decision

The model that will be used is the second model.

The main reasonings behind this decision:

- The results will be in the form $log(m_{xt})$ rather than $log(m_{gt})$ which means that it can be compared to the original Lee-Carter model.
- The model considers each age, location and time seperately whilst the first model considers group g as a representation of age and location.
- There are only four groups considered in model 1 (with no x variable), which may result in lack of trend and result in issues in the estimation of α and β.

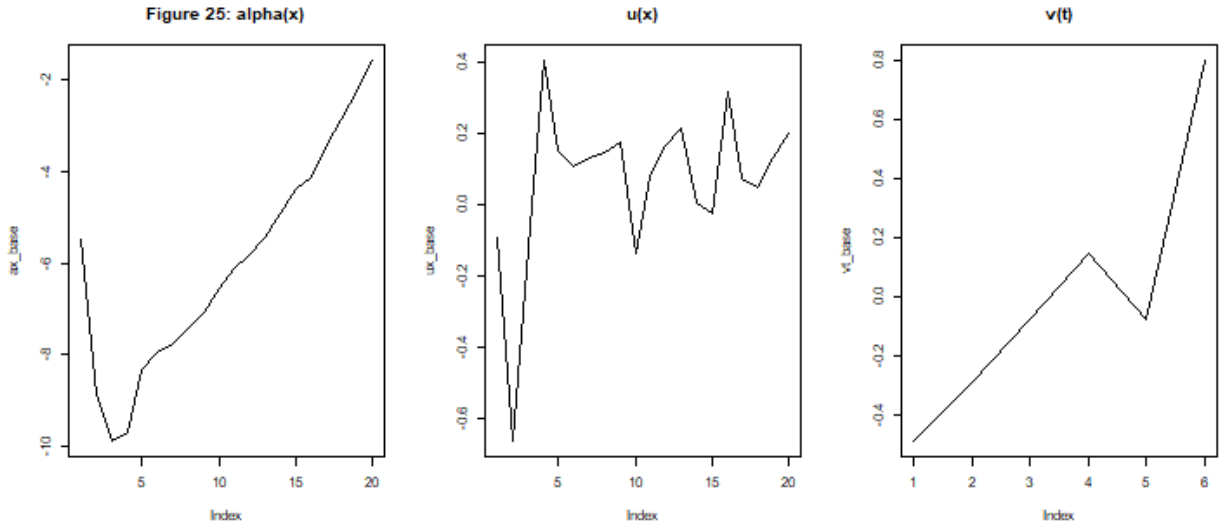
4.4.3 Data

Going forward with the model, the data being used has been obtained from https://www.nomisweb.co.uk/ which is an ONS affliated site that allows users to query from their datasets on specific conditions. The data that has queried on is the 'Mortality Statistics' dataset with ONS Unitary Authority, year (2013-2019), age group and deaths as the query selections. Unfortunately, there are only 7 years of mortality statistics available in the dataset, this should present no issue in constructing an extension to the Lee-Carter model. However, this will naturally result in a decrease in the accuracy of the stochastic model and limit both forecasting and comparison capabilities. This dataset was chosen as it includes year, rather than month, to avoid seasonality issues enountered in this previous section.

4.5 Model for comparison

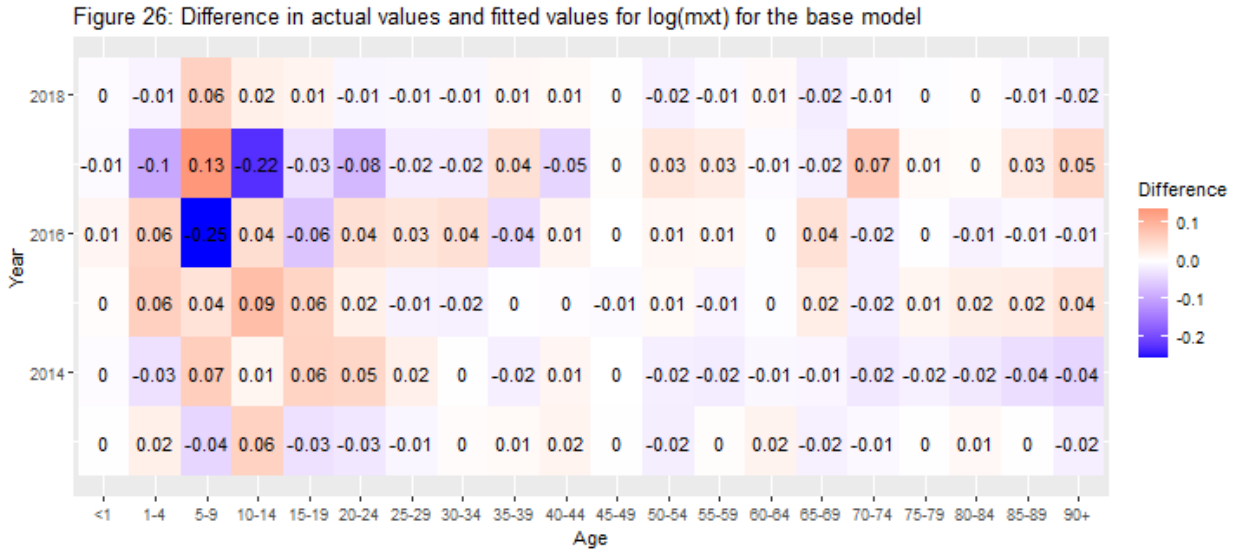
Model comparison is a critical process in the evaluative process of a creating a model. The original Lee-Carter model can be used as a benchmark to compare the proposed stochastic model with. However, the Lee-Carter model constructed in section 2 was created using actuarial life table data. In order to maintain consistency and fairly compare the models, a Lee-Carter model needs to be constructed with the ONS mortality data.

Using the same approach in section 3, values for α_x, $u(x)$ ($u(x) = \beta_x$) and $v(t)$ ($v(t) = \frac{K_t}{S_1}$) can be estimated:



It can be observed that α follows a similar trend as observed in the previous models; there is a high levels of infant mortality which is followed by a sharp decrease, α then increases gradually as age increases. $u(x)$ shows an interesting trend, there does not seem to be any significant observations from the relationship of age x and $u(x)$. $v(t)$ shows a constant increase from $t = 1$ to $t = 6$ with a sharp decrease at $t = 5$ (year 2017). The graph for $v(t)$ highlights the lack of observed years present in the data, which reinforces that there will be a decrease in the reliability of forecasting for $v(t)$.

Furthermore, these estimated values can then be used to fit the Lee-Carter model to the data. This then allows the difference between the fitted and actual values of the model to be observed:



The Lee-Carter model seems to fit fairly accurately in the older age groups (for ages older tha 30) as there is relatively low differences between the actual and fitted values. However, there is a lot of variability within the younger age groups. It can be seen that the model significantly both under and over estimates $log(m_{xt})$ in the younger age groups, especially within years 2016 and 2017.

4.6 Model production

Now that the comparison model has been created, the process of producing the extension of the Lee-Carter model can proceed.

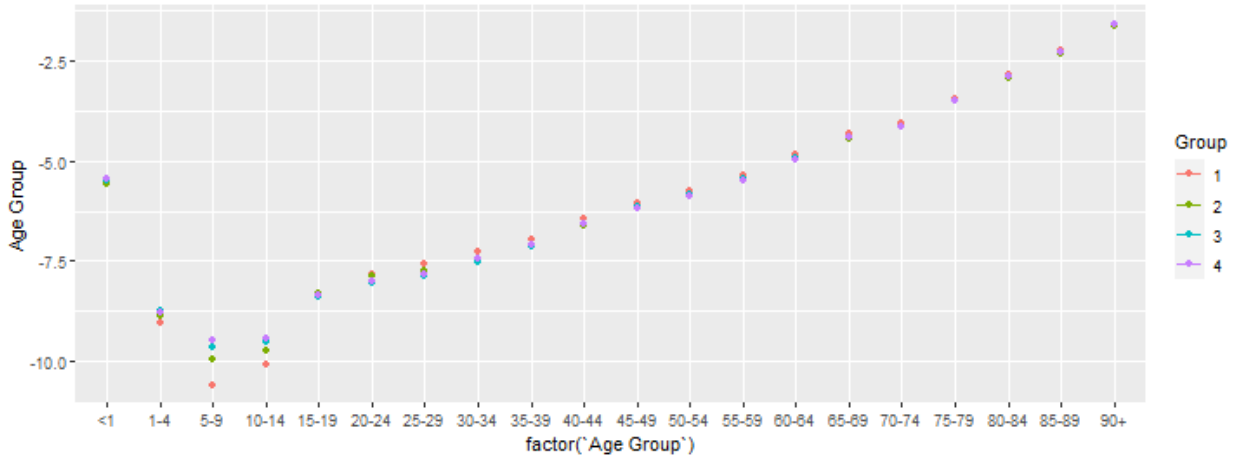
The first step in this process is to split the data into four datasets representing each group identified with the classification tree previously. Each group can be considered as an individual population meaning that a Lee-Carter model can be created for each group. To fit a Lee-Carter model, values for α, β and K_t need to be estimated. α in this model represents the force of mortality in each age group x for each group g:

$$\alpha_{gx} = \frac{\sum_{t=1}^{6} log(m_{gxt})}{6}$$

where $log(m_{g,x,t})$ is the logarithm of central death rate in group g at age x in year t. As there are 6 observed values for t (2013 to 2018), t takes values from 1 to 6.

Looking at the estimated values for α_{gx}, it is noticable that the shape for each group's α curve is similar to the shape of the curve of the α in the comparison model. The main variation between values for α lie within the younger age groups. The trough for the curve of α is present in the 5-9 years old age group and the value of the trough differs in each group; group 1 shows the lowest trough, followed by group 2, then group 3 and group 4 which show a relatively small difference. After age group 15-19, the α values stay fairly closely together. This suggests that the force of mortality in groups 3 and 4 and that these two populations might be very similar.

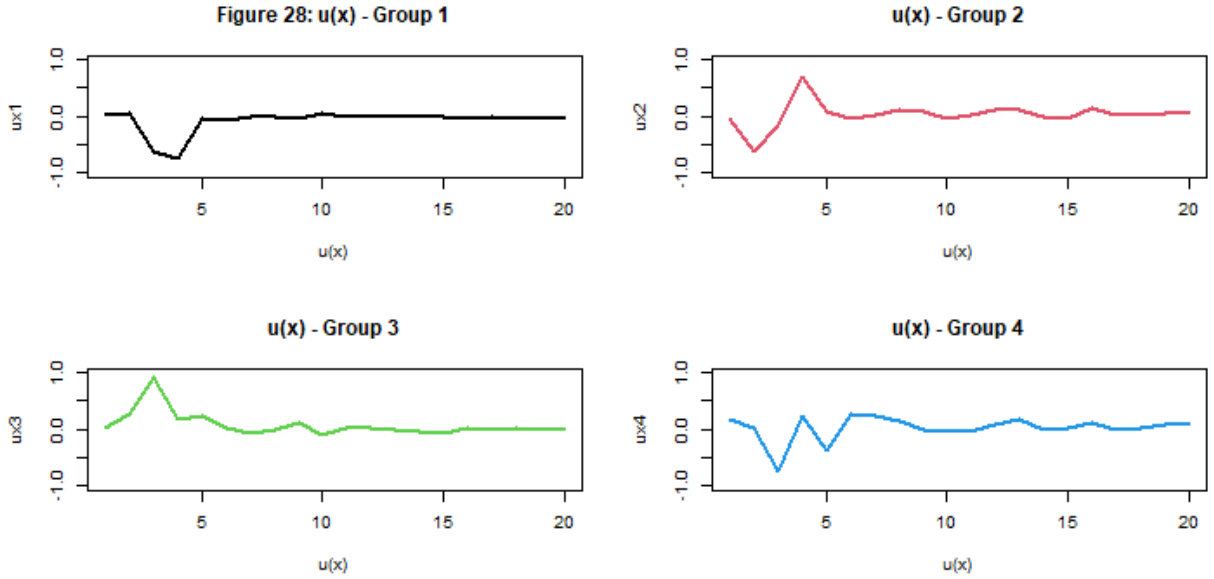
Figure 27: Estimated alpha values for all groups



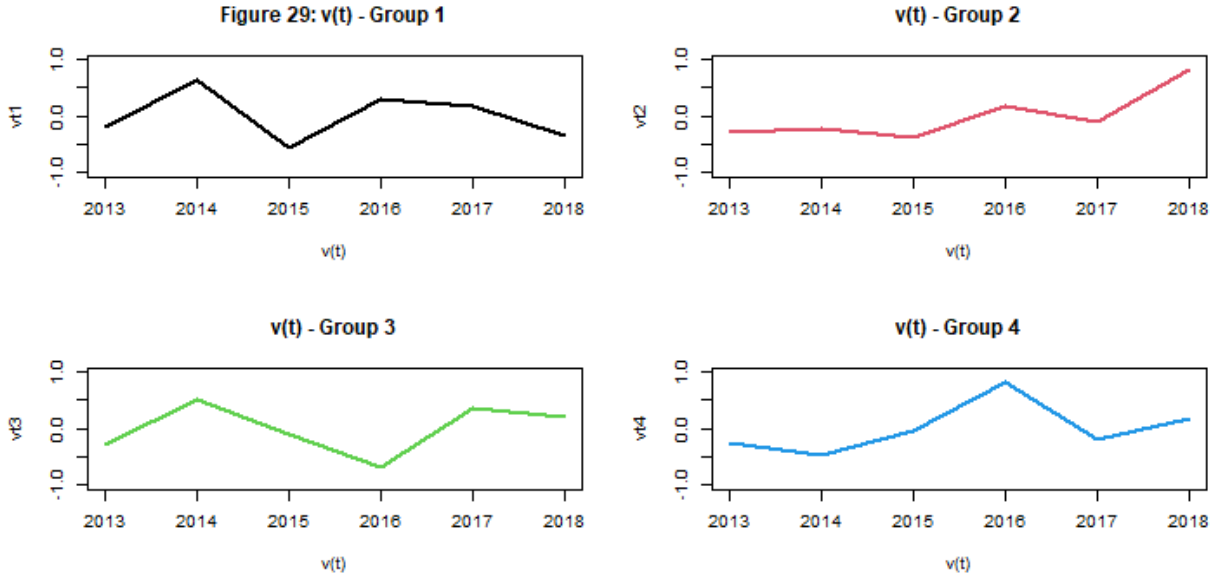
The next stage in the process of producing the model is estimating values for β and K. Again, this process is the same for as the original Lee-Carter model (using singular value decomposition). However, seperate estimates of β and K need to be obtained for each group.

$$\beta_{gx}K_{gx} = log(m_{gxt}) - \alpha_{gx} = S_{g1}u(g,x)v(g,t)$$

where S_{g1} is the first value obtained from diagonal matrix produced from SVD for each group. $u(g,x)$ and $v(g,t)$ are the first column vectors from the u and v matrices produced by SVD for each group.

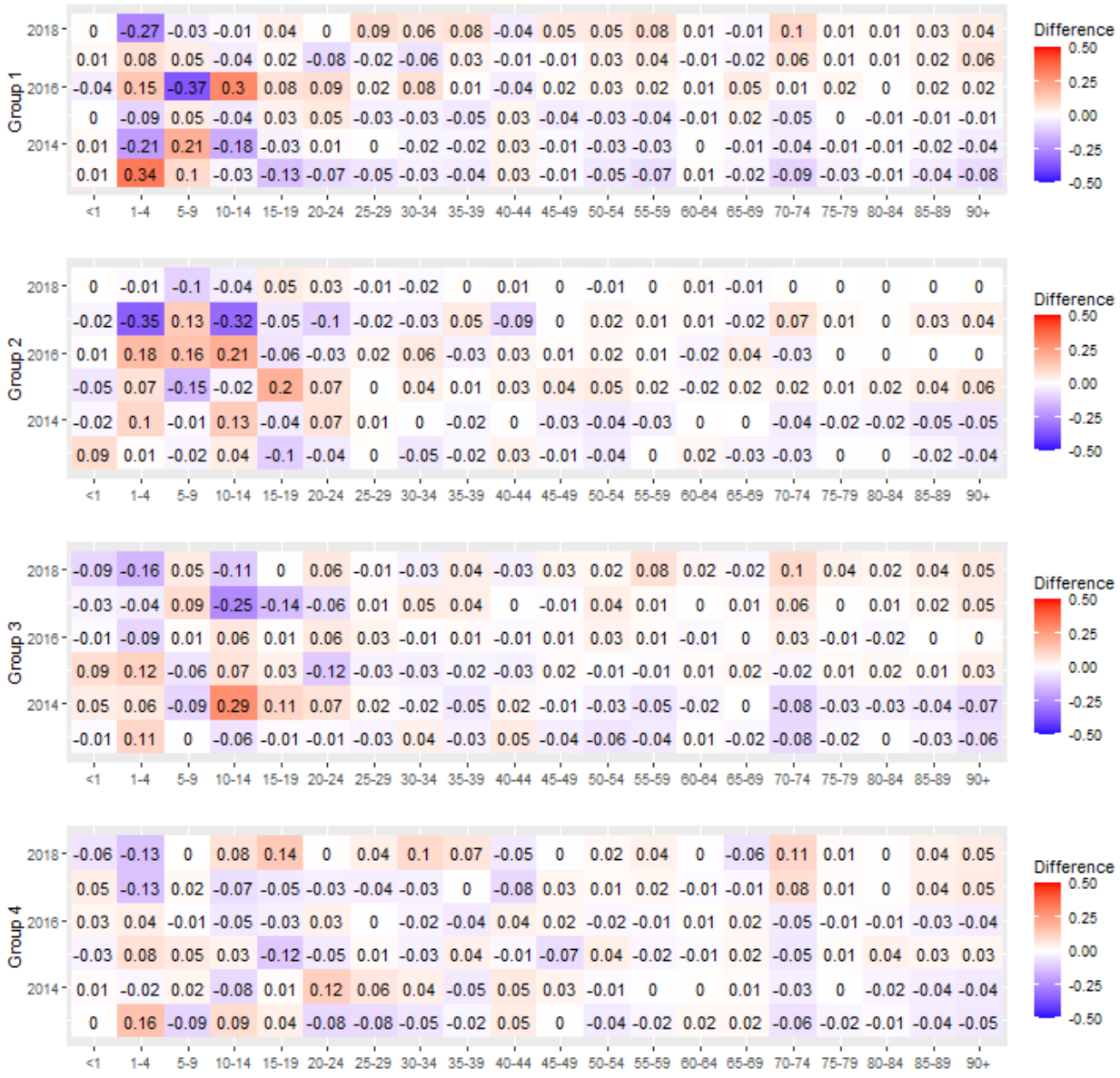


The $u(x)$ value represents β_x which is the change in $log(m_{xt})$ in each age group in any given year. Unlike the estimates for α_x, the $u(x)$ estimates are very different for each group. The curve for $u(x)$ differs for all groups in the earlier age groups. For the older age groups the value for $u(x)$ stays around approximately 0 for all groups. The fluctuations of $u(x)$ between each group signifies that there is a significant difference in mortality within the younger age groups for each cluster of age distributions. This also supports that the reasoning behind grouping the local areas based on their age distributions as mortality seems to effect each population cluster differently.



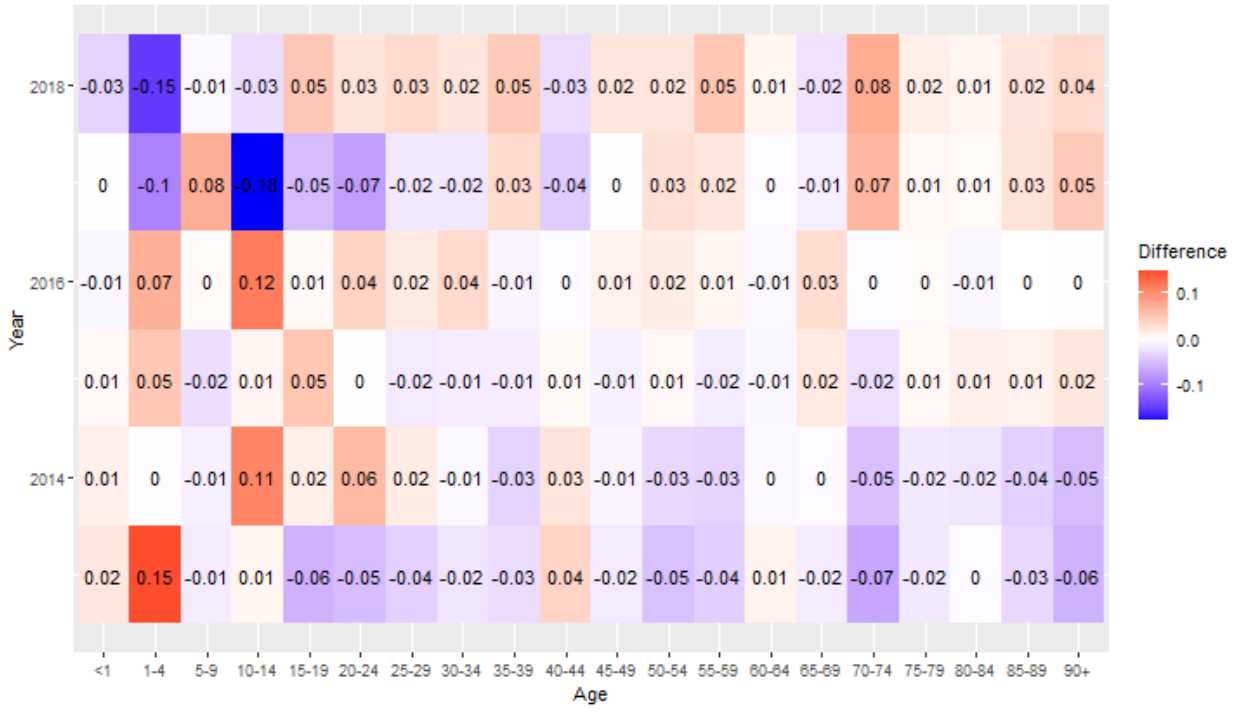
The curves for $v(t)$ show high variation between each different group. A clear limitation in the estimates for $v(t)$ is the lack of years observed due to there only being 6. This does not allow us to gain a strong representation of the change in $log(m_x)$ for a change in t. Observing the 6 years available, it is evident that $v(t)$ follows a general upwards trend with visible peaks in 2014 for groups 1 and 3, and 2016 for groups 2 and 4.

Figure 30: Difference in actual and fitted values for log(gmx) by group, age and year



Similar to the comparison model, the residual values for each of the groups are generally very low for the older ages and there is higher variation within the younger age groups. Groups 3 and 4 have lower residual values compared to the other groups and the comparison model which suggests that the smaller sample sizes of UAs have closely related age distributions, suggesting that the Lee-Carter model is able to fit better for these groups as there is less volatility within each group. There are no signifcant comparisons that can be observed in the difference of residuals of the new model (extention) and Lee-Carter model (comparison).

Figure 31: Difference in actual values and fitted values for log(mx) for the combined model



4.7 Evaluating the model

	Lee-Carter Model	New Model
Min.	-0.2534	-0.1781
1st Qu.	-0.0156	-0.0180
Median	-0.0012	-0.0003
Mean	0.0000	0.0010
3rd Qu.	0.0167	0.0234
Max.	0.1347	0.1513
SD	0.0445	0.0432

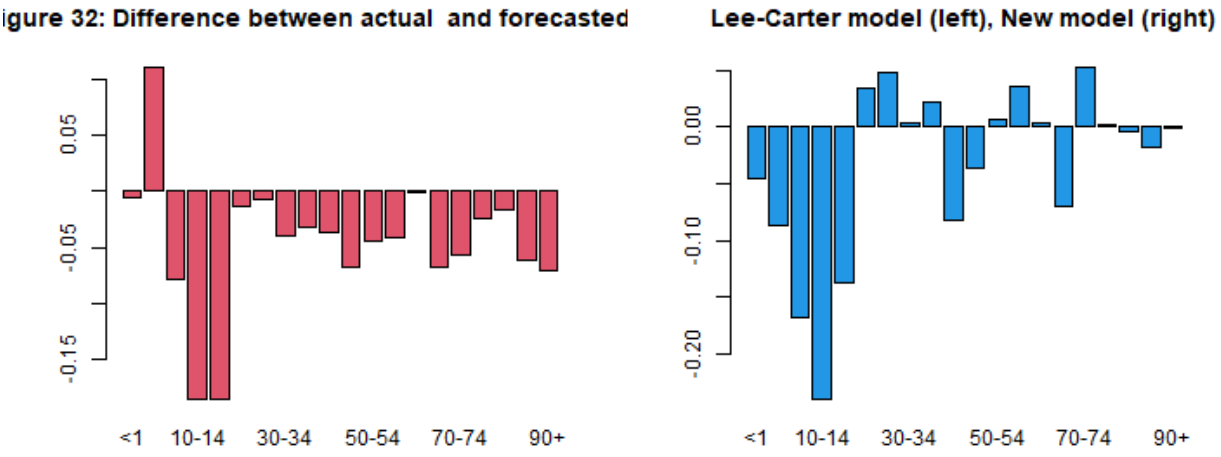
Interestingly, the summary statistics for the new model and the original Lee-Carter model suggest that they perform very similarly. By design, the residuals for the Lee-Carter model have a mean of 0, whilst the new model has a slightly higher mean value of 0.001. The standard deviation of the new model is lower than the Lee-Carter model. Although, this difference is very small. Observing the statistics implies that the model has the potential for being an accurate method to model and predict $log(m_{xt})$. However, there is no sufficient evidence that the new model is better/ worst than the original Lee-Carter model.

4.8 Markov-Chain Monte-Carlo simulation to forecast 2019 mortality

Purposefully withheld from the dataset used to produce the model and the comparison model is the mortality statistics from the year 2019. The reasoning behind this decision is to be able to use a Markov-Chain Monte Carlo simulation on both models (trained on data from 2013-2018) to predict the mortality rate in 2019 (data that the models have not been trained on). Although, this is only one additional year of data, the aim is to be able to understand the difference in the forecasting accuracy of the two models.

Assuming the difference in each value for $v(t)$, $v(t)-v(t-1)$, is a normally distributed random variable with mean μ and standard deviation σ. A simulation of future values of $v(t)$ can be produced by drawing from

the distribution of the difference in $v(t)$. Drawing from the distribution 10,000 times will produce 10,000 predicted values for the difference between 2018 and 2019's value for $v(t)$. Then by adding the value for $v(t)$ in 2018, estimated values for $v(t)$ are obtained. Taking the median value of this vector of 10,000 estimates a singular estimate for $v(t)$ is obtained. Taking the 5th and 95th percentile will also provide a 90% significance level interval for $v(t)$.



The graph shows the difference between the actual and forecasted values for $log(m_x)$ for original Lee-Carter model (left) and new model (right) in each age group. The original Lee-Carter model under fits for almost all age groupss, whilst the new model shows a relatively even spread of both under fitting and over fitting.

Observing the summary statistics for these residual values:

	Lee-Carter Model	New Model
	Min. :-0.18523	-0.239707303748061
	1st Qu.:-0.06721	-0.0725796251014272
	Median :-0.03993	-0.00297540966434451
	Mean :-0.04607	-0.0339027072345162
	3rd Qu.:-0.01562	0.0109072252950269
	Max. : 0.11111	0.0530297012883212
SD	0.0624	0.0768

The mean of the difference between the actual and forecasted values for the year 2019 is lower for the new model than the Lee-Carter model. the standard deviation of the difference shows the opposite, the Lee-Carter model has a slightly lower standard deviation. The difference between both these statistics is fairly negligible and again does not provide sufficient evidence that either model is a better predictor for mortality.

4.9 Final Remarks

Although the comparison between the models could not clarify to whether the extension improved the fitting of the Lee-Carter model, the aim which to propose, produce and apply an extension to the Lee-Carter model that considers age, local area and time has been met. furthermore, the results of this analysis encourage future work on this project to declare whether or not the extension produced can be considered a better model for mortality than the Lee-Carter model.

5 Summary, Limitations and Future Work

5.1 Overview

During the initial stages of the project, three clear aims were identified: 1) Understanding and applying the Lee-Carter model to UK mortality data, 2) Adapting the Lee-Carter model to incorporate local areas within the UK rather than age groups, and 3) Proposing, testing and applying an extension to the Lee-Carter model to consider all age, time and local area.

The first aim of this project was to uderstand and apply the Lee-Carter model to UK mortality data. Through taking UK mortality life table data, variables α, β and K of the Lee-Carter model could be estimated. The estimation of these variables then enabled estimation of $log(m_{xt})$ in the Lee-Carter model. Through this process, an understanding of the Lee-Carter model was gained. Fitting of the Lee-Carter model to the UK mortality data also allowed a Monte-Carlo Markov Chain simulation of $log(m_{xt})$ to be produced, this created an idea of the trend that mortality in the UK is likely to follow over the next 20 years. This also provided insight to the forecasting capabilities of the Lee-Carter model.

The second aim of the project was to adapt the Lee-Carter model to include local areas within the UK rather than age. This aim was achieved through usage of ONS mortality data in 2018 and considering ONS Unitary Authorities as 'local areas'. By susbstituting unitary authorities into the Lee-Carter model instead of age groups the second aim of the project was met. By fitting this model allowed recognition of the differences levels of mortality in different UK local communities.

The final aim of the project was to propose, test and apply an extension to the Lee-Carter model to consider both local area and age groups. Using a classification tree approach allowed clusters of unitary authorities with similar age distributions to be identified. Taking these clusters into account, two different models were proposed. However, there was a clear stronger choice of model that allowed forecasting capabilities and more flexiblity in terms of comparisons (between the original Lee-Carter model). Fitting this model and forecasting for UK mortality in 2019 was then compared to the fitting and forecasted results of the original Lee-Carter model. These results were then compared against the actual values for mortality in 2019. No significant evidence was found that the new proposed model was a better/ worst predictor of mortality than the original Lee-Carter model. Comparing the models indicated that the new model has the potential to be a considered a strong model for mortality through the similarities with the Lee-Carter model's residual and forecasted values. The conclusion that was made is that this would need to be tested further with more data to understand whether or not the extension can be considered better than the Lee-Carter model.

After reviewing each aim of the project, it is clear that they have all been met. However, there are limitations that the models/ data/ methods that can be observed.

5.2 Limitations

Adapting the Lee-Carter model to incorporate local area rather than age for the second aim posed various limitations (mentioned prior). These limitations were combatted and considered in the final section. Discussed below are the obstacles that were encountered in the final model which could not avoided/ cannot be avoided in future.

An obvious limitation in the final section of the project is the lack of data that could be used to apply the comparison and new model. The ONS data provided only 7 years (2013-2019); this was due to limitations in the availability of UK mortality datasets that provided all year, age and unitary authority. This limited the ability to analyse and compare the new model with the comparison model. This could be avoided by using data from a different country which have age and local area mortality datasets which cover a longer period, for example: US mortality by US state.

Another limitation in the extension of the Lee-Carter model that was produced is the classification tree being based on 2018 population data. The clusters identified are only valid for a range of years as populations as populations within local areas constantly change and after a certain period of time, the 2018 population data would no longer be relevent to the data being used. This could be potentially evaded with a dataset that contains more years of data and basing the classification tree on a variable that represents the change in the age distribution over time.

5.3 Future Work

The final model produced presents many potential opportunities for projects in the future.

If further years of UK mortality by age group and unitary authority were to become publically available then this project could be further built upon. As this would enable a more accurate analysis of comparing the extension of the Lee-Carter model with the original Lee-Carter model. This would then potentially enable a conclusion to be drawn on whether or not the extension to the model is a better predictor for mortality than the original model.

Modelling mortality in the United States with the extension of the Lee-Carter model and considering each state as a local area within the model could be a potential project for the future as the 'National Centre for Health Statistics' have a large number of datasets publically available.[6] Additionally, the USA has a range of diversity from state to state which therefore would potentially allow for interesting conclusions to be drawn.

The model could also be built upon further to not only include local area and age, but additional variables as well. For example: ethnicity, household income and sex. These are common factors that are mentioned in mortality discussions.

6 Appendix

The entirety of this report was produced using R and R Markdown.

6.0.1 Libraries

Data Manipulation & Graphs: - ggplot2 - tidyverse - reshape2 - stringr

Map Plots: - sf - rmapshaper - ggmap - rgdal - RColorBrewer

Classification Tree: - rpart - mltools - reshape2 - onehot - rpart.plot - RWeka

Word Clouds: - wordcloud2 - gtable - webshot - htmlwidgets

R Markdown: - knitr - bookdown - biblatex

6.0.2 Estimation of variables

```
# log(mxt)
deaths$lnmx = log(deaths$mx)

# alpha
ax =
  deaths %>%
  group_by(Age) %>%
  summarise(summx = sum(lnmx))

ax$ax = ax$summx / nlevels(as.factor(deaths$Year))

## k bx
Mtrx = matrix(ncol= nlevels(as.factor(deaths$Year)),
              nrow = nlevels(as.factor(deaths$Age)))
```

```
colnames(Mtrx) = levels(as.factor(deaths$Year))
rownames(Mtrx) = levels(as.factor(deaths$Age))

for(i in 1922:2018){
  for(j in 0:109){
      Mtrx[j+1,i-1921] =
        (filter(deaths, Year == i & Age == j)$lnmx[1]) - (filter(ax, Age == j)$ax)
  }
}

# SVD
svd_mtrx = svd(Mtrx)


# u, v and s
ux = svd_mtrx$u[,1]
vt = svd_mtrx$v[,1]
d = svd_mtrx$d[1]


# Fitted Log(mxt)
# log(mx) = ax + (s*ux*vt)
fitted_lnmx = matrix(ncol= nlevels(as.factor(deaths$Year)),
                     nrow = nlevels(as.factor(deaths$Age)))

colnames(fitted_lnmx) = levels(as.factor(deaths$Year))
rownames(fitted_lnmx) = levels(as.factor(deaths$Age))

for(i in 1922:2018){
  for(j in 0:109){
      fitted_lnmx[j+1,i-1921] =
        (filter(ax, Age == j)$ax) + (d*ux[j+1]*vt[i-1921])
  }
}

new_fit = melt(fitted_lnmx)
colnames(new_fit) = c("Age","Year", "Fitted")
```

6.0.3 Monte-Carlo simulation of v(t) and plots

```
runs = 100000
end_vt = vt[length(vt)]
diff_vt = diff(vt)

paths<-10000
years<-20
sample<-matrix(0,nrow=(years+1),ncol=paths)

for(i in 1:paths){
  changes <- rnorm(years,mean=mean(diff_vt),sd=sd(diff_vt))

        sample[1,i]<-end_vt
```

```
        for(j in 2:(years+1))
        {
            sample[j,i]<-sample[j-1,i]+changes[j-1]
        }
}

    sample = sample[-1,]

    par(mfrow = c(1,2))

    matplot(sample,main="Figure 9: Monte-carlo simulation for v(t)",
            xlab="Year",ylab="v(t)",type="l")

    new_sample = vt
    for(i in 1:(paths-1)){
    new_sample = cbind(new_sample,vt)
    }

    new_sample = rbind(new_sample, sample)

    matplot(1922:(2018+years),new_sample,
            main="(Attached to actual values)",xlab="Year",ylab="v(t)",type="l")
```

6.0.4 Percentiles of v(t)

```
for(i in 1:years){
  if(i ==1){
    percentiles_df = data.frame(row.names = 1:20)
    ordered_sample = data.frame(row.names = 1:paths)
    ordered_sample$sample = as.numeric(sample[i,])
    ordered_sample = ordered_sample[order(ordered_sample$sample),]

    percentiles_df$L[i] = ordered_sample[(0.05*paths)]
    percentiles_df$M[i] = ordered_sample[(0.5*paths)]
    percentiles_df$U[i] = ordered_sample[(0.95*paths)]


  } else{
    ordered_sample = data.frame(row.names = 1:paths)
    ordered_sample$sample = as.numeric(sample[i,])
    ordered_sample = ordered_sample[order(ordered_sample$sample),]

    percentiles_df$L[i] = ordered_sample[(0.05*paths)]
    percentiles_df$M[i] = ordered_sample[(0.5*paths)]
    percentiles_df$U[i] = ordered_sample[(0.95*paths)]

  }
}

percentiles_tbl = percentiles_df
```

```
rownames(percentiles_tbl) = seq(2019, 2018+years,1)
percentiles_tbl = round(percentiles_tbl,5)
colnames(percentiles_tbl) = c("5th", "Median", "95th")


percentiles_df$Year = seq(2019, 2018+years,1)


ggplot(data = percentiles_df, aes(x = Year, M))
+geom_line(color = "black", size = 1,lty=1) +geom_point() +
  geom_ribbon(aes(ymin = L, ymax = U), alpha = 0.3) +
  ggtitle("Figure 10: Median of v(t) shown with 5th and 95th percentile ribbon")

OLD_sample = data.frame(new_sample[1:(nrow(new_sample)-years),])
OLD_sample$Year = seq(2018-96, 2018,1)

ggplot(data = percentiles_df, aes(x = Year, M)) +geom_line(color = "blue", lty=2)  +
  geom_ribbon(aes(ymin = L, ymax = U), alpha = 0.2) +
  ggtitle("Figure 11: Median of v(t) forecast with 5th and 95th percentile shown with actual data") +
  geom_line(data =OLD_sample, aes(x = Year,y = new_sample),lty=2)
```

6.0.5 Map plots

```
 #import shape file data
shpla =readOGR(
  dsn="Counties_and_Unitary_Authorities__December_2017__Boundaries_UK.shp",
  verbose = FALSE)
shpla =  spTransform(shpla, CRS("+proj=longlat +datum=WGS84"))

shpla@data$id = rownames(shpla@data)
shpla.points = fortify(shpla, region = "id")
shpla= merge(shpla.points, shpla@data, by = "id")

la = merge(as.data.frame(shpla), summary_deaths,
           by.x = "ctyua17cd", by.y = "Area Code", all.x=FALSE)
map_datt <- get_stamenmap( bbox = c(left = -6.2, bottom = 49,
                                    right = 2, top = 56))

la=la[order(la$order),]

library(RColorBrewer)
###no background map##
##ggplot(data = la, aes(x =long.x, y=lat.x,group=group, fill = Deaths))
##+geom_polygon(color = "black") + scale_fill_gradient(low="white", high="red")

ggmap(map_datt) +
  geom_polygon(data = la,aes(x=long.x, y=lat.x,group=group, fill = Deaths),
               alpha =0.8, show.legend=TRUE)+ scale_fill_gradient(low="white", high="blue") +
  ggtitle("Figure 14: Heat map for number of deaths in each ONS Unitary Authority in 2018")
```

6.0.6 Classification Tree:

(Manual grouping)

```
# Classification tree:
pop_tree = rpart(data = pop_tree_data, formula = variational_coef ~ Location , control = rpart.control(

# Plot
prp(pop_tree, type=1, under = TRUE, xflip=TRUE, tweak=1, cex = 0.6, split.cex = 1)

# Group data
GROUP_POP_TREE = function(p) {

group4 = c("Barnsley","Bath and North East Somerset","Bedford","Bexley","Blackburn with Darwen","Blackp

group5 = c("Barking and Dagenham","Barnet","Brent","Brighton and Hove","Bromley","Camden","Cheshire Eas


group6 = c("Bradford","Bristol, City of","Cambridgeshire","Cardiff","Coventry","Derbyshire","Devon","Ha

group7 = c("Birmingham","Essex","Hampshire","Hertfordshire","Kent","Lancashire","Leeds","Manchester","S

RESULT = NULL

if(p %in% group4){RESULT=1}
if(p %in% group5){RESULT=2}
if(p %in% group6){RESULT=3}
if(p %in% group7){RESULT=4}

return(RESULT)
}


rf_data$group = sapply(rf_data$Location, FUN = GROUP_POP_TREE)

grp_key =
  rf_data %>%
  select(Location, group) %>% unique()

grp_key$group = as.factor(grp_key$group)
```

6.0.7 Word Clouds:

```
grp_key_wc = grp_key
grp_key_wc$group = as.numeric(grp_key_wc$group)

# 1
my_graph <- wordcloud2(filter(grp_key_wc, group ==1), size= 0.1, color = 'cornflowerblue')
saveWidget(my_graph, "tmp.html", selfcontained = F)
webshot("tmp.html", "wc1.png", delay = 5, vwidth = 1000, vheight = 300)

# 2
```

```
my_graph <- wordcloud2(filter(grp_key_wc, group ==2), size= 0.1, color = 'darkolivegreen')
saveWidget(my_graph, "tmp.html", selfcontained = F)
webshot("tmp.html", "wc2.png", delay = 5, vwidth = 1000, vheight = 300)

# 3
my_graph <- wordcloud2(filter(grp_key_wc, group ==3), size= 0.1, color = 'darkorange')
saveWidget(my_graph, "tmp.html", selfcontained = F)
webshot("tmp.html", "wc3.png", delay = 5, vwidth = 1000, vheight = 300)

# 4
my_graph <- wordcloud2(filter(grp_key_wc, group ==4), size= 0.1, color = 'darkorchid')
saveWidget(my_graph, "tmp.html", selfcontained = F)
webshot("tmp.html", "wc4.png", delay = 5, vwidth = 1000, vheight = 300)
```

6.0.8 Combining the models by group:

```
# Expontial to get mx
Mx1 = as.data.frame(exp(fit1))
Mx2 = as.data.frame(exp(fit2))
Mx3 = as.data.frame(exp(fit3))
Mx4 = as.data.frame(exp(fit4))

# Death count vals (by multiplying by population)
for(j in 1:6) {
  for(i in 1:20){
    Mx1[j,i] = Mx1[j,i] * age_dists[1, i+1]
    Mx2[j,i] = Mx2[j,i] * age_dists[2, i+1]
    Mx3[j,i] = Mx3[j,i] * age_dists[3, i+1]
    Mx4[j,i] = Mx4[j,i] * age_dists[4, i+1]
  }
}

# Number of deaths in entire population
Mx = Mx1 +Mx2 + Mx3 +Mx4

total_pop = colSums(age_dists[,2:21])
for(j in 1:6) {
  for(i in 1:20){
    Mx[j,i] = Mx[j,i] / total_pop[i]
  }
}

# log mx fitted values
fit_lnmx = log(Mx)


### Difference (residuals)
Diff_new = as.matrix(act_pop_lnmx - fit_lnmx)

Diff_new = melt(Diff_new)
colnames(Diff_new) = c("Year", "Age", "Difference")
```

References

[1] *BBC News: Where are the UK's youngest and oldest city populations?* URL: https://www.bbc.co.uk/news/uk-43316697.

[2] *CART Modelling*. URL: https://www.statmethods.net/advstats/cart.html.

[3] *Decision Trees*. URL: https://scikit-learn.org/stable/modules/tree.html.

[4] Federico Girosi and Gary King. *Understanding the Lee-Carter Mortality Forecasting Method*, p. 4. URL: https://gking.harvard.edu/files/lc.pdf.

[5] *Human Mortality Database*. URL: https://www.mortality.org.

[6] *National Centre for Health Statistics*. URL: https://www.cdc.gov/nchs/nvss/mortality_tables.htm.

[7] *Nature: Stochastic Modelling*. URL: https://www.nature.com/subjects/stochastic-modelling.

[8] *Office of National Statistics - Central Death Rate Definition*. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/methodologies/guidetocalculatingnationallifetables#:~:text=Definition%20%2D%20This%20is%20known%20as,age%20over%20the%20same%20period.

[9] *Office of National Statistics: UK Geographies*. URL: https://www.ons.gov.uk/methodology/geography/ukgeographies/administrativegeography/england#counties-non-metropolitan-districts-and-unitary-authorities.

[10] *ONS Population Estimates*. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthern

[11] *Princeton University: The Lee-Carter Model*. URL: https://data.princeton.edu/eco572/LeeCarter.pdf.

[12] *Singular Value Decomposition*. URL: https://www.cs.cmu.edu/~venkatg/teaching/CStheory-infoage/book-chapter-4.pdf.

[13] *UK Population Statistics*. URL: https://www.worldometers.info/world-population/uk-population/.