Mapping the uncertain future of longevity: an ensemble approach for forecasting mortality

M Hancock

PhD

2021

Mapping the uncertain future of longevity: an ensemble approach for forecasting mortality

Mark Hancock

A thesis submitted in partial fulfilment of the requirements of the University of Northumbria at Newcastle for the degree of Doctor of Philosophy

Research undertaken in the Faculty of Engineering and Environment

July 2021

Dedicated to

This thesis is dedicated to Alan Hancock.

Your unwavering support allowed me to feel I could achieve anything.

Abstract

Life expectancy has been on the rise in most countries due to the continuous development in healthcare over the past century. This is positive. However, with the rising average age of humans, current plans for pensions and healthcare may need to be revised to remain affordable for a country. Therefore to inform appropriate changes to these plans, mortality forecasts need to be reliable with various sources of uncertainty incorporated. The goal of this project is to develop a model ensemble approach, through Bayesian model averaging (BMA; Hoeting et al. 1999) to forecast human mortality. This approach forecasts mortality by probabilistically combining a suite of forecasting methods, a feature that aims to incorporate model uncertainty. This source of uncertainty is ignored in a conventional single-model forecasting approach.

Applications of this method are done in two settings. The first involves the use of registry based, area-level data from the Human Mortality Database. Here, a number of unique poisson based models can be combined through a model selection weighting technique. Results will be shown comparing the single model approaches, a two-stage model averaging approach (Kontis et al, 2017) and this model averaging technique in a cross validation setting.

The second application involves the use of individual level, under-five mortality, data from the Demographic and Health Surveys. DHS data are collected from developing countries through surveys and are available for a limited numbers of years. The limited survey data pose a number of challenges and allow for different interpretations to modelling. Recently, this has been done in the context of using Poisson or Bernoulli based models (Mejia-Guevara et al., 2019; Wakefield et al., 2018). Here, a different view of the data is shown through using a novel survival analysis approach. The model selection approach to estimate model averaging weights is then extended to the survival setting, with a view to incorporating model uncertainty in forecasts.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my principle supervisor Dr Guangquan Li for his advice, knowledge and most of all patience throughout the last three years. Your willingness to help throughout the entirety of the project has been gratefully appreciated. I would also like to extend my gratitude to Dr Pete Philipson and Dr Paul Oman for the very informative meetings, comments and suggestions. I also acknowledge the PhD studentship from Northumbria University.

I would like to thank my mam and dad for all the love and support they have given me over the years. Clare and Andrew, thank you for the much needed distractions and takeaways. And (of course!) I would like to thank Nala and Penny for their constant companionship.

I would also like to thank my fellow PGR students for the support and encouragement. In particular David and Adam for always being up for a chat - or a game of pool! Finally, I'd like to thank all of my friends for providing much needed distractions.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the Faculty Ethics Committee on 27th July 2018.

I declare that the Word Count of this Thesis is 27385 words (excluding bibliography)

Name: Mark Hancock

Signature:

Date: 27/12/2020

Chapter 1

Introduction

Across the globe most countries attempt to improve the life expectancies of their populations. This evolution can differ depending on the country but the importance of accurate mortality forecasts remains the same for all. Such forecasts are used in a variety of fields, such as pension planning, insurance schemes and healthcare. Life expectancy and mortality estimates are also used to assess the quality of life and thus are part of the definition defining the Human Development Index introduced by the United Nations (UN). Being able to capture the continuing trend of rising life expectancies is also crucial to policy makers. For example, in the case of retirement planning, with the average life expectancy increasing, more funds will be recommended for a comfortable retirement. With this in mind, being able to forecast mortality accurately whilst maintaining appropriate levels of uncertainty is vital.

A number of challenges are faced when forecasting mortality. These challenges include capturing the key features of mortality, having appropriate analytical tools to model different data types and incorporating different sources of uncertainty. Here these main challenges are outlined showing some methods used to combat them.

1.1 Key features of mortality

The challenge of forecasting mortality rates and life expectancy has been attempted for centuries - with one of the first models shown as the Gompertz law of mortality (Gompertz, 1825). The commonality across mortality models, both old and new, shows the need to account for both age and time trends. These are the two dimensions captured in the popular mortality forecasting model introduced by Lee and Carter (1992). The Lee-Carter model is well known in mortality forecasting and is widely used when modelling both all-cause and cause specific mortality (Shair et al., 2017; Basnayake and Nawarathna, 2017; Tuljapurkar et al., 2000). Figure 1.1 shows the effects of age on mortality for data from the United Kingdom. In this case, the log mortality rates for a given age group, using groups of age 0, ages 1-4, then using five year increments until 85+, are plotted in Figure 1.1 for both males and females. The general pattern shows a tick shape, with some notable features. Firstly the initial spike shows children are most at risk from the moment they are born, then as they get older the mortality rate shows a strong decline. This decline halts and starts to rise again as the teenage years are reached, at which point a smaller second spike can be found known as the "accident hump". This is more prominent in males and represents the increase in accidental causes of death, such as motor vehicle accidents (Remund et al., 2017). After the accident hump the mortality then starts to gradually rise with age. The age profiles are also shown for a set of selected years. The decline of mortality over time can be found clearly with each chosen year having a lower log mortality rate across ages, whilst the underlying age pattern remains largely unchanged.

Figure 1.1: Exploratory plots of the effect of age on mortality in the UK for selected years

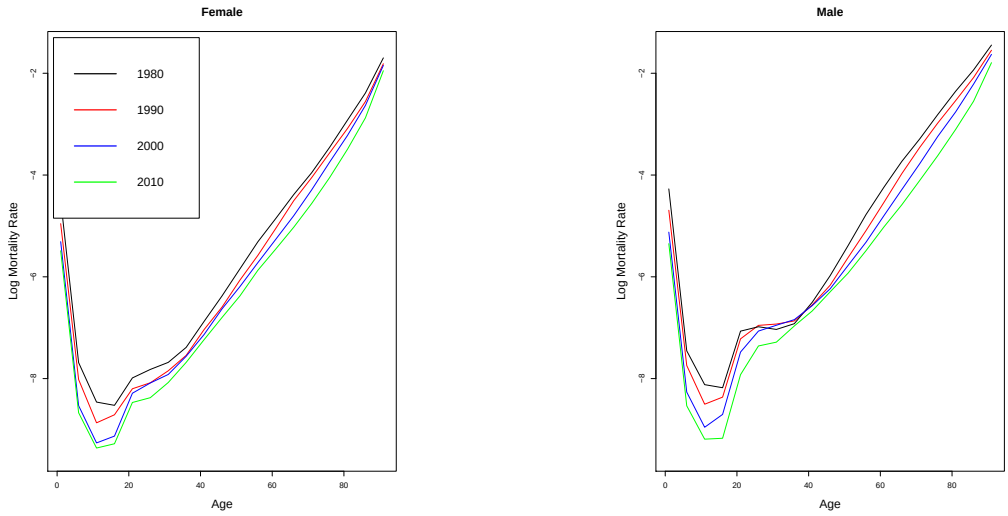
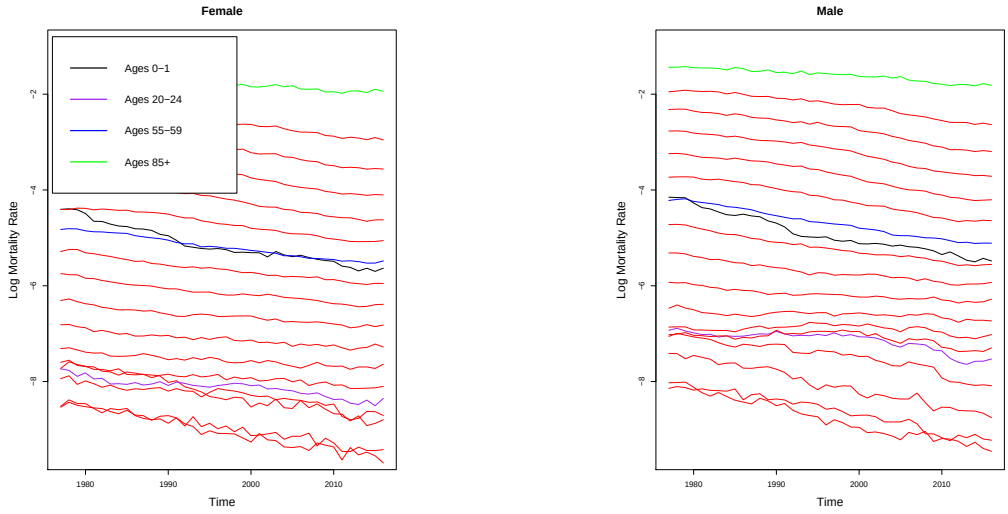


Figure 1.2 shows the effect of time on each of the age groups with more detail. Here the log mortality rates are plotted over time between 1977-2016 using the same 19 age groups. The constant attempts to reduce mortality can be found as all age groups show a decreasing mortality rate over time. A closer look shows the improvement to mortality is not the same throughout each age group, with some notable groups highlighted. Ages 0-1 and 55-59 show clearly that the mortality of these two age groups are improving at different rates over time. In fact the improvements shown by the 0-1 age group overtake that of the 55-59 group. Aside from the difference in rates of decrease, it can be found that different age groups show different patterns of decrease over time. For example the older age groups tend to show more linear decreases of (log) mortality over time. This is not the case for all ages however, as younger ages can show more non-linear time trends. This is shown by the 20-24 age group, suggesting different models may be needed to capture the differing time trends across age groups. The

introduction of a cohort effect or using piecewise linear spline models are examples of statistical tools used to model non-linear time trends in mortality. These models will be used in Chapter 2 and more detail on them are shown there.

Figure 1.2: Exploratory plots of the effect of time on mortality in the UK.



1.2 Data

There are many data sources available for modelling mortality, allowing access to both individual and area level data. In this study area level data is assumed to be at the country level. For all-cause mortality, country level data consists of the total number of deaths and total population for the given country. Examples of this are shown as datasets from the Human Mortality Database (HMD; University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany) 2000) or the Global Burden of Disease project (Murray

et al., 2001), both of which use vital statistics (where available) to provide accurate death counts, as well as official population estimates. The use of the registry based data allows for accurate mortality rate estimates for all ages. However such data is not available for all countries or regions, particularly in developing countries. To combat this, individual level survey data have been collected. The survey data, such as the Demographic and Health Surveys (DHS; ICF International 1984), allows for individual observations to be recorded. The more detailed information available allows for potential covariates that would be more difficult to track at the area level, and can also be aggregated to the area level. However, a challenge of using the survey data is that not all households can be surveyed for the programme to remain affordable. Therefore observations are not available over the entire population. Having data only on samples of the population, as opposed to the entire population, gives rise to some methodological challenges, including the incorporation of survey weights. The focus of such data involves the estimating and forecasting of under-five mortality. Commonly forecasting using the country level data, the death counts are assumed to follow a Poisson distribution (Bennett et al., 2015; Wiśniowski et al., 2015). For individual level data there are two options for forecasting under-five mortality. One involves aggregating the individual level binary outcome data (dead or alive by age five) and using a Poisson model (Mejía-Guevara et al., 2019). Another approach directly models the individual level binary outcome using a Bernoulli model (Mercer et al., 2015). A contribution of this project is the development of a novel survival approach to forecasting under-five mortality using this individual level data.

1.3 Uncertainty

As mentioned previously, the importance of not only accurate forecasts but forecasts which are able to provide appropriate levels of uncertainty are crucial. Here three main sources of uncertainty are explained. Existing strategies used to incorporate them are discussed.

Parameter uncertainty

When using a forecasting model, the forecasts produced are calculated through complicated functions of model parameters. Therefore the uncertainty found in the parameter estimates must be fully acknowledged in the resulting forecast. When using a traditional frequentist approach these parameters are assumed to be fixed and unknown, resulting in each parameter being represented by a single value. Alternatively, each parameter can be assumed to have its own distribution allowing for the uncertainty around the parameter value to be captured. This is the case in Bayesian modelling as rather than treating parameters as fixed and unknown, they are considered random variables - assuming each has its own probability distribution. When forecasting, samples from these probability distributions are then combined allowing for the propagation of parameter uncertainty to be acknowledged in the forecast. Options are also available to incorporate parameter uncertainty whilst using the frequentist framework, such as bootstrapping methods (Efron, 1992).

Data uncertainty

Data uncertainty refers to the uncertainty behind the observed values used to model from. An example as to where this uncertainty can arise is the data collection method used. For example when using the DHS (survey based) data, some accuracy of the data may be lost

due to the respondent not fully recalling the event. Uncertainty can also be found in registry based datasets as the populations are estimated using censuses or official estimates, giving rise to uncertainty around the true totals. These issues in data uncertainty are commonly tackled through using probability distributions (or likelihoods), such as Poisson or Normal distributions, to account for errors with the recorded data. Another issue that can create data uncertainty is the presence of missing data. Handling of missing data can be challenging depending on the type of missing data. For example if the data are assumed to be missing completely at random, the observations can simply be removed from the dataset and not be expected to bias the parameter estimates. Alternatively if it is expected the missing data are not missing at random, parameter estimates may change, meaning missing data will need to be estimated. For a further review on types, and methods to deal with missing data the reader is referred to Kang (2013).

Model uncertainty

When forecasting, it is common to compare a number of competing models which place different assumptions over the data. When comparing the models a conventional aim is to find the best model, and then use this model to provide forecasts. However the use of this single model approach neglects the information available through the other models, and therefore the uncertainty associated with the final forecasts may not be fully representational of the data being used. A common strategy to incorporate model uncertainty into forecasts is the use of model averaging. When using model averaging, forecasts from a (competing) group of models are probabilistically combined to create the forecasting distribution. The use of model averaging is very popular in statistical literature and has been applied in a variety of fields such as weather forecasting (Raftery et al., 2005; Sloughter et al., 2010), economics (Wright,

2008; Jacobson and Karlsson, 2004) and astrophysics (Parkinson and Liddle, 2013). Although it is commonly found that model averaged forecasts outperform forecasts from single model approaches, given the unpredictability of forecasting into the future it is possible for single model forecasts to perform well. Therefore the use of model averaging does not guarantee superior forecasts, however the incorporation of model uncertainty can allow for forecasts to perform more reliably. The use of model averaging has been used in practice for mortality forecasting. Kontis et al. (2017) used a Bayesian model averaging approach to forecasting life expectancy. In the study, a group of 21 models are combined probabilistically to forecast life expectancy for 35 industrialised countries. One goal of this project can be viewed to be inspired by the work of Kontis et al. (2017), in which an aim is to improve the methods used to estimate model weights. In other applications of Bayesian model averaging in the mortality forecasting setting, Kontis et al. (2020) show the impact of COVID-19 in 21 industrialised countries. Fang et al. (2016) evaluate associations between air pollution and respiratory mortality. Whilst Benchimol et al. (2016) and Shang and Haberman (2018) use Bayesian model averaging to forecast mortality for older ages (60+) in Spain and Japan respectively.

The use of the Bayesian framework allows for the incorporation these three sources of uncertainty to be represented in the final forecasts. These forecasts depend on the parameter estimates for each model and, when using model averaging, the weight assigned to each model. Therefore incorporating uncertainty within the joint posterior distribution of the parameters and models is important. Specifically, the joint posterior can be decomposed into three components

$$\begin{aligned}\Pr(\textit{model, parameters} \mid \textit{data}) \propto \Pr(\textit{data} \mid \textit{model, parameters}) \\ \times \Pr(\textit{model} \mid \textit{parameters}) \\ \times \Pr(\textit{parameters})\end{aligned}$$

Under this decomposition, each component addresses one of the three sources of uncertainty separately, but at the same time, they are formally combined, allowing us to learn about the models and the parameters of these models. Specifically, $\Pr(\textit{data} \mid \textit{model, parameters})$ is the likelihood to deal with data uncertainty. $\Pr(\textit{model} \mid \textit{parameters})$ is a way to deal with model uncertainty as it recognises that there is a probability associated with a model. When combined with data, this model probability will be updated to reflect how each model is supported by the data. $\Pr(\textit{parameters})$ deals with parameter uncertainty. Associating each parameter with a probability distribution (through specifying a prior distribution) acknowledges that each parameter is unknown and thus has uncertainty. The advantage of the Bayesian framework is not only a natural way to deal with the three main sources of uncertainty, but also a joint incorporation of the uncertainty. In other words, model parameters and model weights can be estimated with all sources of uncertainty accounted for. These estimates are then used to produce forecasts, allowing for the uncertainty to be represented in forecasts.

1.4 Aims and Objectives

The first goal of this project is to offer an improvement to the existing life expectancy forecasting approach shown by Kontis et al. (2017). This alternative approach, here called the model selection method, shows promising results when compared to the two-staged approach

used by Kontis et al. (2017). The model selection method also has contrasting features to the two-staged approach, that can improve forecasting ability. For example, the model selection approach offers more flexibility when specifying model weights. This added flexibility can allow for model weights to be estimated at the age group level - a feature which is not possible when using the two-staged approach. Using the UK data, it was shown in Figure 1.2 that different ages show different mortality patterns over time. Different models can be used to capture these different trends. Therefore using age group specific weighting, different combinations of models can be used at each age group leading to improved age group and life expectancy forecasts.

The next goal is to transition into the challenge of forecasting using the individual level survey data. As mentioned previously, current methods involve aggregating the data (Mejía-Guevara et al., 2019) or using a Bernoulli model (Mercer et al., 2015). Here a different angle to viewing the survey data is shown through a novel survival analysis approach. In this approach survival models are used to capture the change in mortality across birth cohorts, with view for forecasting for future cohorts. The method is then compared against a data aggregation approach, in which different variations of the Lee-Carter model are used.

A third goal is improve the survival analysis framework for forecasting under-five mortality. This improvement is with view to include model uncertainty into the forecasts, by showing how model weights can be obtained through a model selection styled approach.

1.4.1 Statement of original contribution

- An improved BMA method for forecasting life expectancy using a model selection approach
- A novel approach to forecasting under-five mortality using a survival analysis approach.
- An extension to the survival analysis framework to show how BMA weights can be obtained, with a view to incorporate model uncertainty in the forecasts.

1.5 Thesis Structure

In Chapter 2 the model selection Bayesian Model Averaging (BMA) technique to forecast life expectancy is shown and compared in a cross validation setting against the approach used by Kontis et al. (2017). Chapter 3 shows a more detailed look into the features of both BMA methods. In Chapter 4 a novel approach to forecasting under-five mortality using the DHS data is shown through using survival analysis. Chapter 5 then combines this survival approach with the BMA method, showing how model weights can be estimated. Finally, Chapter 6 concludes the research as well as outlines some possible extensions that can be made.

Chapter 2

A Bayesian Model Averaging approach to mortality forecasting

Providing mortality forecasts that are accurate not only in terms of point estimates but also uncertainty is vital to help allocate resources such as pension funds or medical resources. A common practice is to rely on using a single model to forecast (Shair et al., 2017). In doing so a key area of uncertainty, model uncertainty, is neglected. To incorporate model uncertainty Bayesian Model Averaging (BMA; Hoeting et al. 1999) is used. A key issue when using not only Bayesian model averaging, but model averaging in general, is how model weights are obtained. In this chapter a BMA approach to mortality forecasting is shown through using a model selection method to obtain weights. The forecasting ability of this method is then compared against the BMA approach shown by Kontis et al. (2017) for both male and female data over three separate time periods.

2.1 Introduction

Life expectancy has been on the rise in most countries due to the continuous development in healthcare over the past century. This is positive. However, with the rising average age of humans, current plans for pensions and healthcare may need to be revised to remain affordable for a country. Therefore to inform appropriate changes to these plans, mortality forecasts need to be reliable with various sources of uncertainty incorporated.

The issue of forecasting mortality rates and life expectancy has been attempted for centuries - with one of the first models shown as the Gompertz law of mortality (Gompertz, 1825). More recently the most used model for mortality forecasting is found to be the Lee-Carter model (Lee and Carter, 1992). The Lee-Carter model consists of an age specific intercept and an age specific gradient with disturbances. This model is able to capture two important effects in mortality namely the age and time effects - and is similar in principle to other models used in this chapter. Despite being quite a simple model, the Lee-Carter model, and its variations, performs well in mortality forecasting for developed countries (Booth and Tickle, 2008). A recent example of this is shown by Shair et al. (2017) in which various models (including the Lee-Carter model) are compared against each other but each independent model is used for forecasting. Although when using a single model the forecast performance can be good, mortality patterns are assumed to be captured only through this model, neglecting forecasts other models can provide.

An alternative to the single model approach can be found through the use of model averaging. Using model averaging, a group of forecasting models can be combined using weights that represent the performance of each model. In the Bayesian paradigm Hoeting et al. (1999) show a Bayesian Model Averaging (BMA) setting in which the BMA forecasts can be represented as

$$\Pr(\Delta|D) = \sum_{j=1}^{k} \Pr(\Delta|M = j, D)\Pr(M = j|D)$$

where Δ is the quantity of interest, in this case a set of mortality rates forecasted across different age groups. The observed data are D and the number of models included is k. Under this notation $\Pr(\Delta|D)$ is the model averaging forecast distribution, given the observed data averaged across different models probabilistically. This can be found as a combination of the individual model forecasting distribution $\Pr(\Delta|M = j, D)$ and the posterior probability of the model $\Pr(M = j|D)$, which acts as the weight applied to the given model. Here it is easy to find the forecasting distribution for a given model, however the model weights can be difficult to compute, leading to different approaches to determine weights.

In Bayesian (and frequentist) model averaging some methods to obtain model weights involve the use of various Information criteria, such as the commonly used Akaike Information Criterion (AIC; Akaike 1974) or alternatively the Bayesian Information Criterion (BIC; Schwarz et al. 1978). These criterion are used to calculate model weights as they can both be used to compare model performance. Model weights would be calculated through normalising $exp(-\frac{1}{2}\text{IC})$ across all models, where IC is the chosen information criterion. Other methods to assign model weights have also been used such as through measuring the model performance in cross validation

settings, most commonly using leave-one-out cross validation (Hansen and Racine, 2012; Yao et al., 2018). Leave-one-out cross validation involves hiding a single observation from each model, to then compare the observed value against the expected. This can then be repeated, by hiding a different observation, allowing for more information about model performance.

In the mortality forecasting setting a number of studies incorporate model uncertainty in forecasts. Kontis et al. (2020) used a BMA approach to measure the impact of COVID-19 in 21 industrialised countries. In the study, models were used to estimate the expected mortality as if the pandemic had not occurred. When combining models for BMA equal weights were then assigned to each model, assuming each model is equally likely to produce the best estimates. In another study, Shang and Haberman (2018) forecast mortality for Japanese men and women aged 60 to 99. In the study a trimming approach to BMA is used in which a large number of models are considered (17 in their application). The forecasting ability of these models are compared, in which the best performing models are placed in a superior model set. BMA forecasts are then produced by applying equal weights to each of the superior models. Fang et al. (2016) used BMA to find associations between air pollution and respiratory mortality. In the study model weights were calculated through using the BIC in the same manner as mentioned above.

Kontis et al. (2017) show a study in which an ensemble of 21 forecasting models are combined, through BMA, to produce life expectancy forecasts. To obtain model weights they show a two-staged cross validation approach in which a number of observed years are hidden from the models, to then be used to compare the forecasting performance of each. Although this method compares the forecasting ability of each model, rather than how well the models fit

the data, this usage of cross validation is found to be computationally costly due to the need to fit models twice, with one using the set of training data and again using the entire dataset. A secondary issue can also arise here when dealing with very limited data as the requirement to hide some data from each model may no longer be feasible when the number of observation times are already small. In this study, an alternative method to obtain model weights is shown, through a model selection technique. The forecasts are then compared against that of the two-staged approach shown by Kontis et al. (2017). A key advantage of the model selection technique is that model weights can be estimated through a single fitting of each model, negating the need to split the data in any way.

2.2 Model Averaging Weights

2.2.1 Model Selection weighting system

To show a case for the model selection weighting system, a general all cause mortality forecasting setting is used. Consider the modelling framework

$$y_{at} \sim Poisson(n_{at}\mu_{at})$$

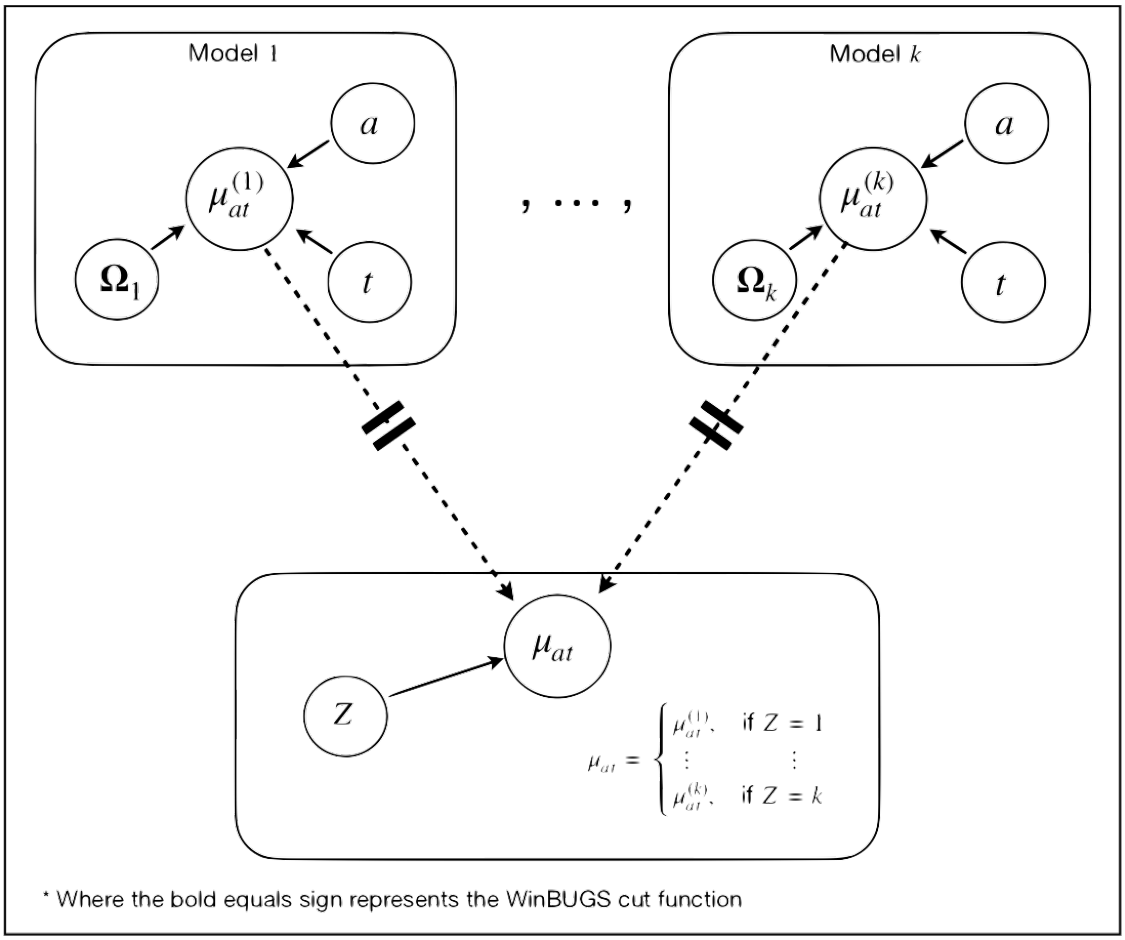
where y_{at} is the number of deaths in year t at age group a and n_{at} is the corresponding population size. In this form, μ_{at} represents the mortality rate in which a model structure can

be placed on. For a model averaging ensemble, a set of k independent forecasting models are used to capture different trends in the mortality rate. Such as

$$\begin{aligned} log(\mu_{at}^{(1)}) &= f_1(a,t,\mathbf{\Omega}_1), && \text{Model } 1 \\ \vdots \qquad & \qquad \vdots && \qquad \vdots \\ log(\mu_{at}^{(k)}) &= f_k(a,t,\mathbf{\Omega}_k), && \text{Model } k \end{aligned}$$

where $f_j(a,t,\mathbf{\Omega}_j)$ represent different functions involving age (a) and time (t) with the vector of parameters being $\mathbf{\Omega}_j$ for $j = 1, ..., k$. A model selection parameter Z can then be introduced to determine which model fits the data best, therefore with Z selecting a model at the iterative level, the probability of a model j being selected $\Pr(Z = j|\text{data})$ can be found. The number of times a certain model has been selected can then be proportional to its model weight. Prior specifications are made to all model parameters as well as the selection parameter Z. In this study an equal prior weight is assigned to each model $Z \sim Categorical(\frac{1}{k}, ..., \frac{1}{k})$ assuming all models are equally likely to be selected. This process is illustrated in Figure 2.1.

Figure 2.1: A visual representation of the model selection weighting procedure at the iterative level



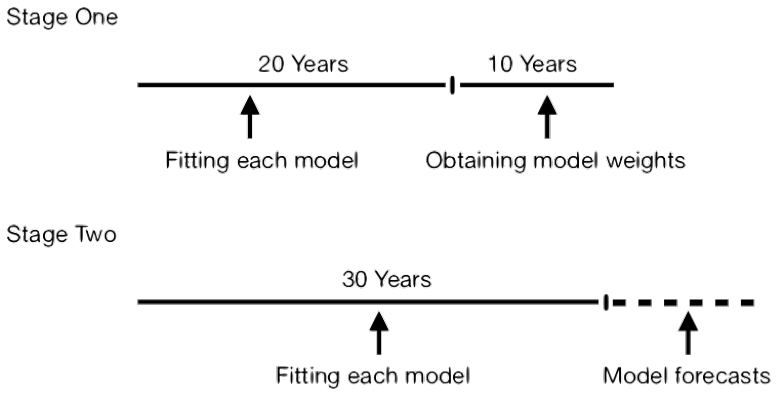
2.2.2 Alternative Two-Stage approach

As a comparison to the model selection technique for defining model weights, the formulation shown by Kontis et al. (2017) is used. Kontis et al. (2017) use Bayesian Model Averaging to forecast the life expectancy of males and females in 35 countries. The weighting method used involves a two-staged approach in which a proportion of the data is hidden from the models for the weights to be calculated in a cross validation period. The models are then used with the full amount of data in which model weights are applied to these forecasts to form the model averaging forecast. To calculate model weights, the mortality rate forecasts are combined to show the projected life expectancy from birth then the mean model forecasts are compared through their projection bias. The projection bias is defined as the amount, on average, the expected forecast of life expectancy deviates from the observed. The absolute value of the projection bias is then taken and normalised to form each models weight. Therefore the model weight w_j for the j^{th} model is shown by

$$w_j = \frac{exp(-|\text{Projection Bias}_j|))}{\sum_{i=1}^{k} exp(-|\text{Projection Bias}_i|))}$$

Figure 2.2 shows a graphical representation of the application of the two-stage approach where thirty years of data are used. In the first stage, due to a cross validation section to measure forecasting performance, the full dataset needs to be split into two sections. The first split of the data, here twenty years, is used to fit each model. The models are then to forecast for ten years, in which the withheld data are used to compare forecast performance - leading to calculating model weights. In the second stage each model is refitted using the full thirty years of data. The model weights, acquired during the first stage, can then be applied to the forecasts made.

Figure 2.2: Showing the structure of the two-stage approach



Here the life expectancy transformation is used to combine the age specific death rates to a single number forecast. The life table methods used are shown in Preston et al. (2000).

When using both the model selection and two-staged weighting methods each model is fitted using Markov Chain Monte Carlo (MCMC) software WinBUGS (Lunn et al., 2000). MCMC uses a simulation based method in order to approximate posterior distributions for a given parameter. After the chosen model and prior distributions have been specified an MCMC method uses a Markov chain, meaning that every sample is dependent on the previous sample, starting from user specified initial values. The underlying distribution of the Markov chain is the posterior distribution of interest therefore after running the chain past convergence, and discarding the burn in period, the resulting samples can be used to find the relevant posterior summaries. For example the posterior mean of a given parameter can be found as the mean of the samples from the chain, and the credible interval can be found by taking the relevant quantiles. Here, due to the Bayesian framework the credible interval is used rather than a confidence interval. The confidence interval used in a frequentist approach which views each

parameter as fixed, therefore the confidence interval captures the uncertainty in the interval given. For example if a 95% confidence interval is provided then there is a 95% chance that the interval covers the true value of the parameter. Alternatively the Bayesian framework assumes each parameter to be random variables, with each having its own probability distribution. Therefore the credible interval expresses the uncertainty in the parameter whilst the interval remains fixed. This uncertainty in the posterior samples of the parameters is then incorporated into the final forecasts of the models.

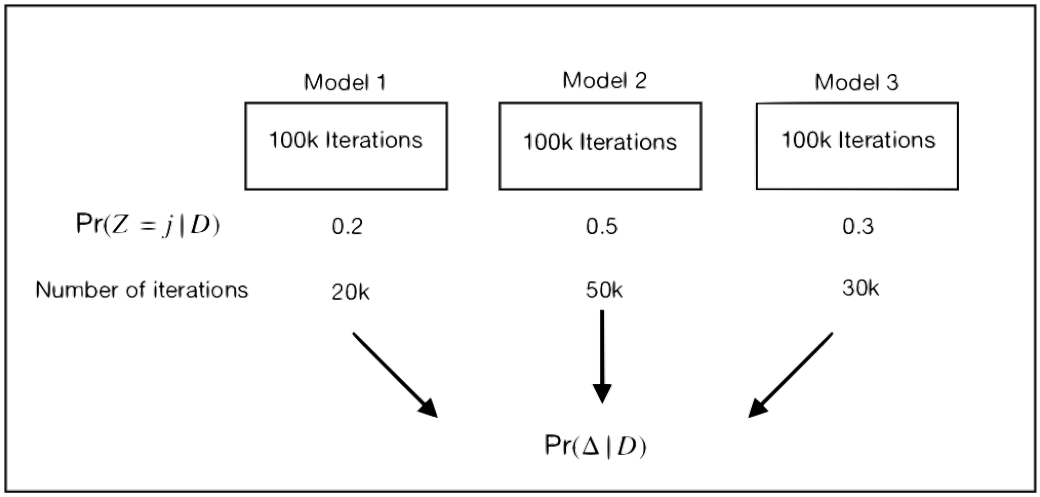
A better posterior approximation will be found when running the chain for longer, however the more iterations that are used the longer the model will take to run. In this case two chains were used to ensure convergence to the same parameter estimates, with the posterior summaries using both chains, and each chain was ran for 100,000 iterations after accounting for burn in.

When using the model selection approach, each model is fitted independently using the entire dataset. The WinBUGS cut function is used to allow for each model to be estimated independently from not only each other but also the overarching model selection parameter (Plummer, 2015). This allows all of the estimation to be done through using a single model script. To check the cut function does as intended, each model was fitted fully independently from each other (each with its own model script) and it was found that parameter estimates were similar to those when the cut function was used. If the issue arises that the cut function alters the estimation of the models, each model can be fitted independently and the model outputs can then be fed into a separate model selection parameter, as done in Chapter 5.

2.3 Producing forecasts using the model selection approach

After model weights have been found the next step of model averaging is how the weights are used to combine each model forecast. Here each forecasting model is combined through a sample of its forecasting posterior distribution proportional to its assigned weight. This is illustrated in Figure 2.3 in which 100,000 iterations of a forecast are stored from three models. The weight for model j for $j = 1, ..., k$ is $\Pr(Z = j|D)$ and $\Pr(\Delta|D)$ is the model averaging forecasting distribution. A random sample of the iterations, with a sample size proportional to the model weight, are then combined to use as the model averaging projection. This method of applying weights is also used by Kontis et al. (2017).

Figure 2.3: Illustrating how model iterations are used along with model weights to form the model averaging forecasts



2.4 Application

2.4.1 Human Mortality Database

The data being used is obtained from the Human Mortality Database (HMD; University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany) 2000). Data from the HMD are publicly available and the HMD holds over 40 years of data for over 40 countries. The data available are area level and registry based, containing information such as the age-specific number of deaths and populations of a given country.

The dataset being used for our group of models is from the UK.[1] To account for the different relationships between age and mortality, ages at death are split into 19 age groups starting from age group 0 then age groups 1-4, 5-9, 10-14 and so on until the last age group of 85 and above. When modelling life expectancy it is common to separate genders as it has been found that females usually experience a higher life expectancy than males (Mathers et al., 2001).

To test this assumption for the UK data, a Poisson model is used with gender as a covariate. Here males were used as the reference group and the UK data used is for years 1947-2016. The output of the test is shown in Table 2.1 and confirms that the effect of separating genders is highly significant as zero is not contained in the parameters credible interval. This shows, as expected, that females experience a lower mortality rate to that of males. Therefore in this application males and females are separated.

[1] Data accessed 12th June 2018

Table 2.1: Output showing the significance gender has as an effect on mortality for UK data

Covariate	Mean	Std. Deviation	2.5%	97.5%
Intercept	-4.482	2.123e-4	-4.483	-4.482
Gender	-0.04786	2.999e-4	-0.04845	-0.04728

2.4.2 Models

2.4.2.1 Age-Time Model

In this application five models are used to for the model ensemble. The first of the five is a linear age-time model taking the form

$$y_{at} \sim Poisson(n_{at}\mu_{at})$$

where y_{at} and n_{at} represents the mortality counts and populations respectively for each age group a at year t. A log-linear trend is placed on the mortality rates μ_{at}

$$log(\mu_{at}) = \alpha_0 + \alpha_a + (\beta_0 + \beta_a)(t-1) + \epsilon_{at} \tag{2.1}$$

with priors

$$\begin{aligned} \alpha_0 &\sim N(0, 10000) & \beta_0 &\sim N(0, 10000) \\ \alpha_{1:19} &\sim RW_1(\mathbf{W}_a, \sigma_\alpha^2) & \beta_{1:19} &\sim RW_1(\mathbf{W}_a, \sigma_\beta^2) \\ \frac{1}{\sigma_\alpha^2} &\sim Gamma(0.001, 0.001) & \frac{1}{\sigma_\beta^2} &\sim Gamma(0.001, 0.001) \\ \epsilon_{at} &\sim N(0, \sigma_\epsilon^2) & \frac{1}{\sigma_\epsilon^2} &\sim Gamma(0.001, 0.001) \end{aligned}$$

A common intercept and gradient are used in the form of α_0 and β_0. Any deviations found using the different age groups can then be found through α_a and β_a. An overdispersion term ϵ_{at} is included to account for any extra data variability unaccounted for through the age specific trends and is applied through using hierarchical centring (Gelfand et al., 1995). Vague priors are used on the overall intercepts and the precision of the overdispersion term, allowing the parameters to be estimated solely from the data. It is understood that the Gamma prior placed on the precision for the overdispersion term can be viewed as slightly informative when using a small number of observations (Gelman et al., 2006). To check the sensitivity of this prior a strictly positive uniform prior was placed on the variance of the overdispersion term and was found to produce the same parameter estimates after transforming the output the the precision. Due to the similarities between neighbouring age groups, as shown in Figure 1.1, a random walk of order 1 (RW_1) is used as a prior. Taking α_a as an example, the RW_1 structure is shown as

$$\alpha_a | \boldsymbol{\alpha}_{\{-a\}}, \sigma_\alpha^2 \sim \begin{cases} N(\alpha_{a+1}, \sigma_\alpha^2) & \text{for } a = 1 \\ N(\frac{\alpha_{a-1}+\alpha_{a+1}}{2}, \frac{\sigma_\alpha^2}{2}) & \text{for } a = 2, ..., 18 \\ N(\alpha_{a-1}, \sigma_\alpha^2) & \text{for } a = 19 \end{cases} \tag{2.2}$$

where $\boldsymbol{\alpha}_{\{-a\}}$ is the set of age specific parameters excluding α_a. Equation 2.2 can be expressed in the more concise form

$$\alpha_a|\boldsymbol{\alpha}_{\{-a\}}, \sigma_\alpha^2, \mathbf{W}_\alpha \sim N\left(\frac{\sum_{i\in\Delta a}\alpha_i}{m_a}, \frac{\sigma_\alpha^2}{m_a}\right) \tag{2.3}$$

where Δa is the set of neighbours to age group a, m_a is the number of neighbours age group a has and $\mathbf{W}_\alpha$ is a weights matrix defining which age groups are neighboured. As an example of the weights matrix, given the first five age groups $\mathbf{W}_a$ is shown as

$$\mathbf{W}_a = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Therefore by using this structure for the full set of 19 age groups, both the age specific model intercept α_a and gradient β_a assume the effects of each neighbouring age groups are similar.

2.4.2.2 Age-Period-Cohort Model

The second model used is an Age-Period-Cohort (APC) model, first shown by Clayton and Schifflers (1987). The APC model used here is similar to the Age-Time model with the addition of one parameter dimension. The added parameters are used to capture any effect the birth cohort of the person may have on their mortality. As this application is using data from the United Kingdom, this component is to capture differences between birth cohorts

such as anyone born in the 'golden' cohort. The golden cohort relates to a significant life expectancy increase for anybody that was born between 1925 and 1934 which is expected to be due to environmental conditions (such as rationing at an early age) during the First World War and Depression (Goldring et al., 2011). There are however known issues when using an APC model, as documented by O'Brien (2011), in which due to the age, time and birth cohort being dependent on each other the sole effects of each are unidentifiable. Although this creates a problem when analysing the effects of the model it does not interfere with the overarching mortality rate, and therefore is not an issue with this application as it is only focused on the mortality rate projections. The form of this APC model being used is shown as

$$y_{at} \sim Poisson(n_{at}\mu_{at})$$

$$log(\mu_{at}) = \alpha_0 + \alpha_a + (\beta_0 + \beta_a + \gamma_c)(t-1) + \epsilon_{at}, \quad \text{where } c = T - a \tag{2.4}$$

$$\begin{aligned}
\alpha_0 &\sim N(0, 10000) & \beta_0 &\sim N(0, 10000) \\
\alpha_{1:19} &\sim RW_1(\mathbf{W}_a, \sigma_\alpha^2) & \beta_{1:19} &\sim RW_1(\mathbf{W}_a, \sigma_\beta^2) \\
\frac{1}{\sigma_\alpha^2} &\sim Gamma(0.001, 0.001) & \frac{1}{\sigma_\beta^2} &\sim Gamma(0.001, 0.001) \\
\gamma_{1:129} &\sim RW_2(\mathbf{W}_c, \sigma_c^2) & \frac{1}{\sigma_c^2} &\sim Gamma(0.001, 0.001) \\
\epsilon_{at} &\sim N(0, \sigma_\epsilon^2) & \frac{1}{\sigma_\epsilon^2} &\sim Gamma(0.001, 0.001)
\end{aligned}$$

where T represents the calendar year. The APC model is structured the same as the Age-Time model with the addition of the set of cohort parameters γ_c. A random walk smoothing prior is placed on the cohort effects in the form of a random walk of order 2 (RW_2). The RW_2 prior

holds a stronger assumption than the RW_1 in which the effects of four neighbouring cohorts are assumed to be similar, rather than two. This stronger prior is placed due to each cohort being defined as $c = T - a$ and with the ages being in age groups, the exact birth year of each person is unknown - leading to overlapping birth years in each cohort group c.

2.4.2.3 Lee-Carter Model

The third model used is an adaptation of the Lee-Carter model. The Lee-Carter model, introduced by Lee and Carter (1992), has been widely used in both all cause and cause specific mortality forecasting (Shair et al., 2017; Basnayake and Nawarathna, 2017; Tuljapurkar et al., 2000). In the Bayesian paradigm, the model is shown as

$$y_{at} \sim Poisson(n_{at}\mu_{at})$$

$$log(\mu_{at}) = \alpha_a + \beta_a\gamma_t + \epsilon_{at} \tag{2.5}$$

$$\begin{aligned}
\alpha_{1:19} &\sim N(0, 10000) & \beta_{1:19} &\sim N\left(1, \frac{1}{3}\right) \\
\gamma_t &= \gamma_{t-1} + d + \tilde{\epsilon}_\gamma, & &\text{where } \gamma_1 = 0 \\
d &\sim N(0, 10000) & & \\
\tilde{\epsilon}_\gamma &\sim N(0, \sigma_\gamma^2) & \frac{1}{\sigma_\gamma^2} &\sim Gamma(0.001, 0.001) \\
\epsilon_{at} &\sim N(0, \sigma_\epsilon^2) & \frac{1}{\sigma_\epsilon^2} &\sim Gamma(0.001, 0.001)
\end{aligned}$$

An overall age intercept is found through α_a with β_a and γ_t representing the age-time effects

respectively. Due to the multiplicative age and time effects having identifiability issues when sampling, a more strict prior is placed on β_a. This prior ensures the age effect is to remain positive, however has roughly a 5% chance to change the trend. The Lee-Carter model assumes the age effects β_a stay constant over time and therefore only forecasts the time effects γ_t. To do this a random walk with drift is used, in which the drift is estimated with a vague prior.

2.4.2.4 Piecewise Linear Spline Models

Two Piecewise Linear Spline models are also used for forecasting. They are used as an extension to the linear Age-Time model in which the most recent observed years have more of an impact when forecasting. The two models are chosen with different knot (K) positions as a 'true' knot position is unknown. The first model has a knot $\frac{2}{3}$ through the time period whilst the second knot is placed $\frac{4}{5}$ through the time period. Therefore with thirty years of in-sample data the first model uses knot $K = 20$ whilst the second uses knot $K = 24$. The form of the spline model is shown as

$$y_{at} \sim Poisson(n_{at}\mu_{at})$$

$$log(\mu_{at}) = \alpha_0 + \alpha_a + \epsilon_{at} + \begin{cases} (\beta_0^{(0)} + \beta_a^{(0)})t, & \text{if } t \leq K \\ (\beta_0^{(0)} + \beta_a^{(0)})K + (\beta_0^{(1)} + \beta_a^{(1)})(t - K), & \text{if } t > K \end{cases} \tag{2.6}$$

$$\alpha_0 \sim N(0, 10000) \qquad \beta_0^{(0)} \sim N(0, 10000) \qquad \beta_0^{(1)} \sim N(0, 10000)$$
$$\alpha_{1:19} \sim RW_1(\mathbf{W}_a, \sigma_\alpha^2) \qquad \beta_{1:19}^{(0)} \sim RW_1(\mathbf{W}_a, \sigma_{\beta^{(0)}}^2) \qquad \beta_{1:19}^{(1)} \sim RW_1(\mathbf{W}_a, \sigma_{\beta^{(1)}}^2)$$
$$\frac{1}{\sigma_\alpha^2} \sim Gamma(0.001, 0.001) \qquad \frac{1}{\sigma_{\beta^{(0)}}^2} \sim Gamma(0.001, 0.001) \qquad \frac{1}{\sigma_{\beta^{(1)}}^2} \sim Gamma(0.001, 0.001)$$
$$\epsilon_{at} \sim N(0, \sigma_\epsilon^2) \qquad \frac{1}{\sigma_\epsilon^2} \sim Gamma(0.001, 0.001)$$

Similarly to the Age-Time model and APC model each overall intercept has a vague prior whilst a random walk one smoothing prior is placed on the age effects.

Each model is fitted simultaneously to the thirty years of data using the MCMC software WinBUGS (Lunn et al., 2000) to allow for the model selection weighting system. All model fittings used 100,000 iterations (after accounting for burn in) to gather forecasts and estimate model weights.

2.4.3 Application of Model selection Weighting System

Within the application of the model selection weighting system two key specifications are applied. Firstly, in this application each model utilises age specific parameters to capture the different mortality patterns amongst age groups. With these age specific parameters, each model is able to capture different time effects across each age group. Therefore with each age group not showing the same effects of mortality across time, it is incorrect to assume one model has superior forecasts across all age groups. To counteract this, model weights are estimated at the age group level allowing for different models that capture different time trends across age groups to be selected. This age specific weighting is notably different from that

of the two-stage approach (Kontis et al., 2017) which uses one weight per model due to the life expectancy transformation. The second specification involves how each model posterior density is fed into the model selection. With each full density being fed into the model selection it is found that in-sample model fits are very similar due to the inclusion of the overdispersion terms. To capture how well the underlying model performs, the posterior model densities are used to estimate weights without the use of the overdispersion parameters - allowing for the core effects of each model to be judged. A visual representation of the differences between both the model selection and two-stage approach can be found in Figure 2.4, whilst the differences in features can be found in Table 2.2.

Figure 2.4: A visual representation of the differences between both approaches

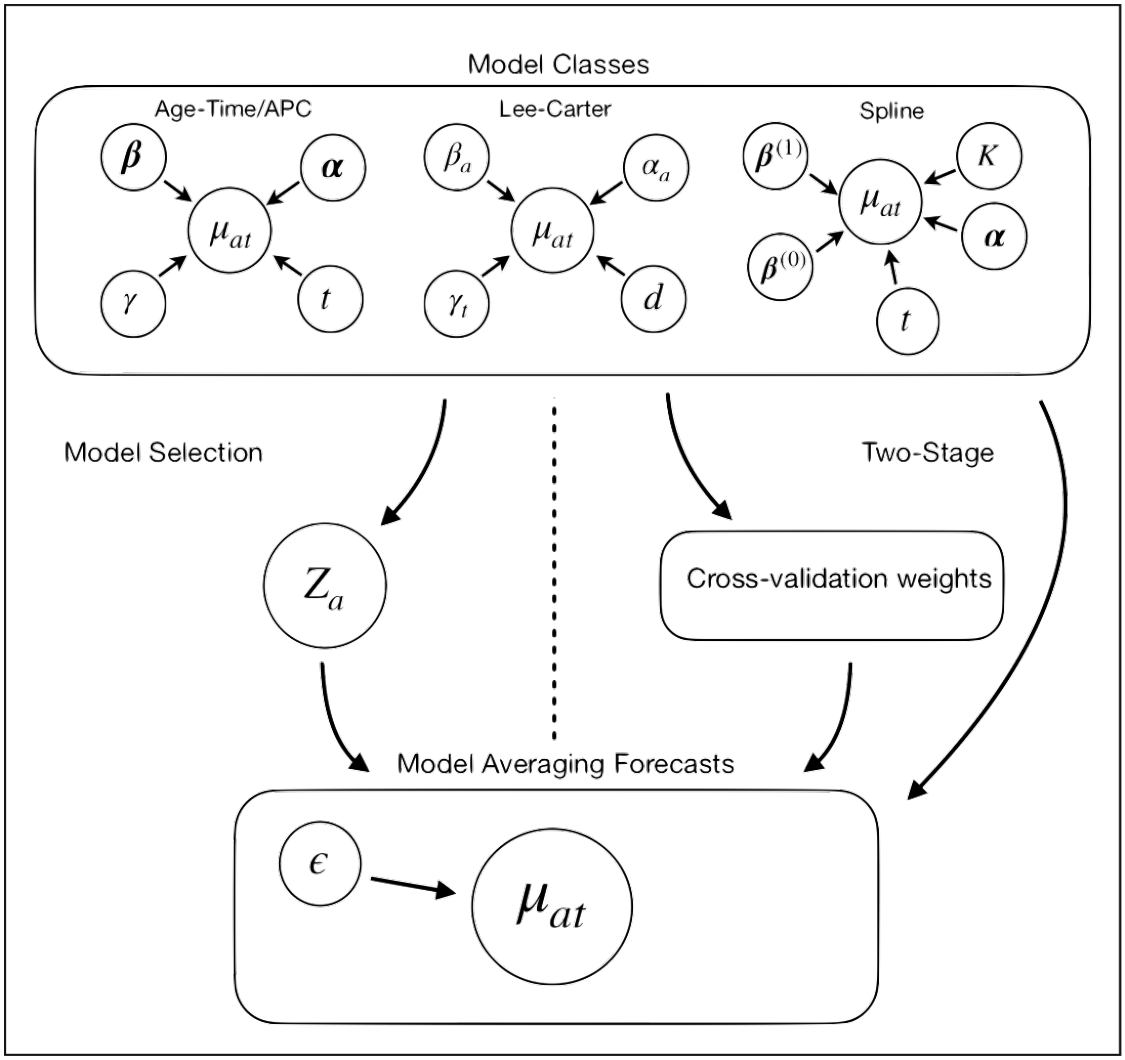


Table 2.2: Comparing key features of the model selection and two-stage approaches

Feature	Model selection approach	Two-Stage approach
Computational time	The single in-sample fitting of the model selection approach allows for a smaller computational burden, and therefore a shorter computational time than that of the two-stage approach.	The need for two stages, and two model fittings, means the computational time of using the two-stage approach will always be larger than when using the model selection approach. In practice the strength of this feature can vary depending on variables such as software usage, model complexity and dataset sizes. In this application the usage of the model selection approach saves approximately forty minutes per fitting.
Weighting flexibility	A feature of the using the model selection approach is the ability to estimate more flexible weights. This flexibility is utilised in this application where weights are estimated at the age group level, however if desired there is the option to be even more specific with the weighting such as incorporating spatial disparities into the weighting.	The option of using more flexible weighting is not available when using the two-stage approach as the weights depend on the scale of the bias used. If the age-specific mortality rate bias replaces the life expectancy bias in the weight calculation the smaller scale of the bias will result in the approach being even more lenient, and therefore assign more weight to the poorly performing models.
Preselecting models	The model selection approach is found to be a more strict weighting system, with any poorly performing model not being selected. This allows for less concern when initially choosing which models are to be used in the ensemble as any badly fitting model will not effect the model averaged forecast. The difference between the two approaches when preselecting models is shown in Chapter 3.	With the two-stage approach being more lenient, more care needs to be taken when preselecting models. When using this approach even a very poorly performing model is still able to influence the final model averaging forecast.

2.4.4 Rolling Window

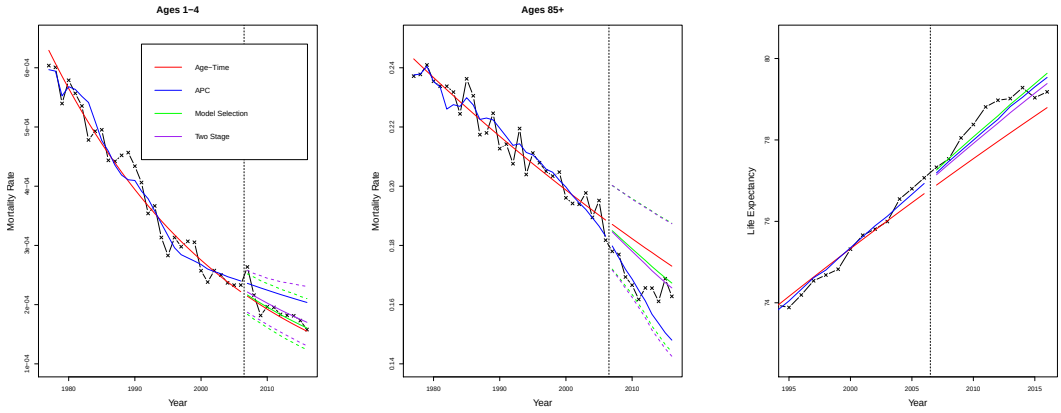
Due to the vast amount of data available through HMD for the UK, a rolling window technique is used to compare the forecasts of each model. This rolling window approach utilises time frames or 'windows' within the UK's history with a view to compare model forecasting performance in each window. The usage of the rolling window also allows for an overall review of model performance in which results from the three windows can be combined, showing which models were more consistent in producing strong forecasts. Three time frames were chosen within the UK's history, each holding 40 years of data, being Window 1: 1977-2016; Window 2: 1947-1986; and Window 3: 1962-2001. Data from each window was then split, hiding the last ten years from each model to then use to cross validate. This allows for thirty years in each window to fit each model. When using the two-staged approach, model weights are calculated through hiding the last ten years of the thirty.

2.4.4.1 Individual Window results

In the cross validation periods, two measures of forecast performance are used, the average forecast error and the coverage. The average forecast error is used to measure the accuracy of the point estimate forecast. It can be seen as the absolute mean error between the forecasts and the observed data (after mortality rates are converted into life expectancy), therefore the smaller the better. The 95% coverage is used as a measure of the uncertainty level surrounding each forecast. It shows how many observations lie inside the 95% credible intervals around each forecast, meaning the closer the nominal level of 0.95 the better. If the coverage statistic is less than the nominal level then it means that the uncertainty around the forecasts may not be

large enough. On the other hand if the coverage statistic is larger than the nominal level then the uncertainty range of the forecasts may be too large. The coverage statistic is calculated here using the age-group specific mortality forecasts. For clarity, Figure 2.5 shows two sets of mortality rate forecasts alongside the life expectancy forecasts for males in in Window 1. The life expectancy can be viewed as a combination of all of the age group specific mortality rates (at a given time) therefore to create a one number summary as to the point estimate accuracy of a forecast, the life expectancy is used. Using the coverage statistic a one number summary showing the accuracy of forecasted uncertainty can be formed. Therefore a higher quantity of forecasts is used by calculating the coverage with the mortality rate projections.

Figure 2.5: Two age group specific mortality rate forecasts alongside the life expectancy forecasts using the male Window 1 data



It should also be noted that when forecasting in the cross validation periods extra Poisson uncertainty is also added to each forecasted death count. This is due to the comparison against the observed data in which the data are assumed to follow a Poisson distribution.

Table 2.3 shows the comparisons of each single model forecast against the model averaging alternatives for each window. The table can be interpreted in two ways. Firstly the usage of model averaging can be compared against that of a single model approach. The mean forecast bias shows that using a single model approach can be beneficial when considering a single window, with the single models having a lower bias in three out of the six settings. On the other hand, when solely using the single model approach no single model consistently gives better forecasts than the others. This supports the assumption of not knowing the true forecasting model, leading to using a model averaging method to incorporate the model uncertainty. The decision to use model averaging over the single models can be found to be clearly advantageous when comparing the uncertainty around forecasts. In all but one of the settings (Window 2; Female) the model averaging approaches provide more appropriate uncertainty levels according to the coverage statistic. Noticeably the one occasion in which the single models provide a better coverage, both model averaging approaches are found to give too much uncertainty, sticking with the trend that the added model uncertainty typically leads to larger ranges of uncertainty in forecasts.

With the use of model averaging shown to be more beneficial than when using a single model approach, a second interpretation of the table can be the comparison between the two BMA approaches. A common pattern is found when comparing the model selection to the two-stage approach in that the model selection constantly gives more accurate point estimate forecasts, outperforming the two-stage approach in five of the six settings whilst having the most accurate forecasts out of any model in two of the settings. Alternatively the two-stage approach consistently provides larger uncertainty bounds, however a larger uncertainty interval is not always ideal as found in both Window 2 settings.

Table 2.3: Comparing forecasting performance of single models and model averaging models over three time windows

			Male		Female	
		Model	Mean Forecast Bias	95% Coverage	Mean Forecast Bias	95% Coverage
Window 1	Single Model Approach	Age-Time	0.708	0.52	0.454	0.73
		Age-Period-Cohort	0.298	0.73	0.299	0.73
		Lee-Carter	0.766	0.65	0.56	0.83
		Spline ($K = 20$)	**0.247**	0.72	0.36	0.71
		Spline ($K = 24$)	0.29	0.75	**0.271**	0.78
	BMA Approach	Two-Stage	0.342	**0.94**	0.363	**0.96**
		Model Selection	0.258	0.87	0.291	0.78
Window 2	Single Model Approach	Age-Time	0.865	0.91	**0.173**	0.86
		Age-Period-Cohort	1.92	0.44	1.952	0.52
		Lee-Carter	0.971	0.83	0.587	**0.96**
		Spline ($K = 20$)	1.081	0.83	0.92	0.87
		Spline ($K = 24$)	1.03	0.89	1.234	0.9
	BMA Approach	Two-Stage	1.02	0.99	0.74	1
		Model Selection	**0.648**	**0.97**	0.553	0.98
Window 3	Single Model Approach	Age-Time	1.035	0.47	0.408	0.68
		Age-Period-Cohort	0.337	0.67	0.458	0.56
		Lee-Carter	0.847	0.61	0.264	0.81
		Spline ($K = 20$)	0.13	0.81	0.229	0.64
		Spline ($K = 24$)	0.342	0.82	0.372	0.81
	BMA Approach	Two-Stage	0.272	**0.96**	**0.177**	**0.88**
		Model Selection	**0.128**	0.9	0.223	0.83

2.4.4.2 Overall results

Although the forecast comparisons for individual windows provide a more detailed look at each model, Table 2.4 shows overall results when combining each window alongside forecast performance encompassing both males and females too. The overall male and female results both show that when using a model averaging approach, not only do the point estimate forecasts show to be more consistent, the uncertainty intervals also provide a more consistently suitable range. The one single model shown to provide accurate forecasts is the linear age-time model for females, however the error of using this in a single model approach can be found when showing the same model is the worst overall model for male data.

The overall results provide a single number overview as to how each model performed. The overall advantage of using model averaging can be found in which both techniques provide more accurate forecasts as well as more suitable coverage in comparison to each single model. It is more difficult to decide which model averaging approach is superior as the model selection approach provides more accurate forecasts whilst giving a lower coverage statistic. On the other hand the two-stage approach gives excellent uncertainty ranges whilst sacrificing forecast accuracy.

Table 2.4: Overall mean forecast bias and 95% coverage after combining the three windows

		Male		Female		Overall	
	Model	Mean Forecast Bias	95% Coverage	Mean Forecast Bias	95% Coverage	Mean Forecast Bias	95% Coverage
Single Model Approach	Age-Time	0.869	0.63	**0.345**	0.76	0.607	0.69
	Age-Period-Cohort	0.851	0.61	0.903	0.6	0.877	0.61
	Lee-Carter	0.861	0.69	0.47	0.86	0.666	0.78
	Spline ($K = 20$)	0.486	0.78	0.503	0.74	0.495	0.76
	Spline ($K = 24$)	0.554	0.82	0.626	0.83	0.59	0.83
BMA Approach	Two-Stage Model Selection	0.546	**0.96**	0.427	**0.95**	0.486	**0.96**
		0.345	0.91	0.356	0.86	**0.35**	0.89

2.5 Discussion

In this chapter a new Bayesian model averaging method for forecasting life expectancy has been shown, and compared with forecasts from an alternative approach used by Kontis et al. (2017). In a cross validation setting the benefits of using model averaging can be found. Over a set of three windows both model averaging techniques have been shown to provide more consistently accurate forecasts than when using a single model approach. Similarly, the model averaging forecasts were also shown to provide more appropriate uncertainty levels across the three time periods. When comparing the two model averaging approaches, the model selection method has been shown to provide more accurate forecasts consistently, whilst retaining comparable levels of uncertainty.

There are many extensions available to follow this study. Firstly, in line with the research shown by Kontis et al. (2017), forecasts can be made using information from multiple countries. Such datasets are readily available from the Human Mortality Database and alongside the two stage approach, the model selection approach can be used to provide accurate life expectancy forecasts. Secondly a wider variety of models can be used in the model ensemble. Some considerations in models can include that of piecewise linear spline models with the knot placements in different positions, allowing for any uncertainty surrounding the best knot position. Other options available given a large time period are those that are able to penalise the older data, and subsequently prioritise the most recent data, such as weighted likelihood models or P-Spline models. Further considerations can be the Cairns-Blake-Dowd model (Cairns et al., 2006) as used by Benchimol et al. (2016) which is more typically suited to modelling mortality at older ages (60+).

The inclusion of additional models can also incorporate the usage of more model covariates. So far all models used include an overdispersion parameter, added to incorporate extra variability to the Poisson models and in turn capture any unknown effects. However, if the data are available, the use of model covariates can allow further insight into the precursors of mortality and help when creating life expectancy projections. There are however issues involving how the covariates can be used with forecasting. The main challenge is how to relate the covariate at time t to the observed mortality. One option is a simultaneous approach whereby the covariate value affects the mortality now. For example, a rise in GDP in a country immediately leads to improvement in the country's medical service which in turn helps reduce mortality. If all happen within the same year, then such a simultaneous approach is reasonable. Therefore when producing a mortality forecast, a forecast for the covariate is also required. A second

approach is to use a lag technique in which the mortality at time t is affected by a covariate that happened at a previous time $\hat{t}$. Using the GDP example again, if the impact of medical improvement on mortality takes a few years to realise, then a lag is more appropriate. To then produce mortality forecasts the covariate forecasts are not needed until the number of forecasted time units are greater than $t - \hat{t}$.

The extension that will be shown in Chapter 3 involves a more detailed comparison of the two BMA approaches shown here. In the chapter, two investigations are considered highlighting the importance of preselecting models using both approaches.

Chapter 3

Comparing Model selection against the Two-Stage approach

In Chapter 2 forecasts of two BMA strategies were shown alongside the performances of single model approaches used in the ensemble. The results show that the use of both BMA approaches provide a more reliable forecasting quality than when using a single model. Here the main focus is to compare the two BMA approaches, highlighting the importance of preselecting models to use in the ensemble. This is shown through the use of two investigations, both being an extension of the main application shown in Section 2.4.

3.1 Investigation 1: Prior knowledge of a poorly forecasting model

As a first extension to the application shown in Section 2.4, the case of having prior knowledge as to how each model performs is examined. For this case, the extension is shown using the second time window. In the second window there is a commonality between both males and females when considering the performance of each model. It is found that the forecast quality of the Age-Period-Cohort model is extremely poor, in both accuracy and coverage. Although this model is used in the model ensemble for both BMA approaches, it is found that the model selection approach provides better forecasts than the two-staged approach in both bias and coverage. This motivates the first application, in which the knowledge of understanding which models perform poorly can be used to improve the BMA forecasts when using the two-stage approach. These improved forecasts can then be compared against the model selection approach, in which the prior knowledge is not assumed.

Table 3.1 shows such a case in which the model selection BMA method, with the inclusion of the entire collection of models, is shown against the two-stage approach which is used in three settings. Firstly the two-stage approach is shown as in Section 2.4 where all models are included, secondly the forecasts of the two stage approach is shown without the inclusion of the Age-Period-Cohort model. A third setting of the two stage approach is shown by not only removing the Age-Period-Cohort model but also removing the Piecewise Linear Spline model with the knot placed $\frac{4}{5}$ through the time period, as that is also shown to give poor forecasts for both males and females. Across both males and females the forecast results show

that even with the prior knowledge of which models should not be used for forecasts, the two-staged approach does not provide as accurate forecasts as the model selection approach in terms of both point estimates and uncertainty levels. This can perhaps be considered a surprising outcome considering the weighting system of the two stage approach uses the model forecasts whilst the model selection approach assumes a strong in sample fit will result in a strong forecast performance.

Table 3.1: Overall mean forecast bias and 95% coverage after combining the three windows

		Male Window 2		Female Window 2	
	Model	Mean Forecast Bias	95% Coverage	Mean Forecast Bias	95% Coverage
Single Model Approach	Age-Time	0.865	0.91	**0.173**	0.86
	Age-Period-Cohort	1.92	0.44	1.952	0.52
	Lee-Carter	0.971	0.83	0.587	**0.96**
	Spline ($K = 20$)	1.081	0.83	0.92	0.87
	Spline ($K = 24$)	1.03	0.89	1.234	0.9
BMA Approach	Two-Stage	1.02	0.99	0.74	1
	Two-Stage without APC	0.972	0.99	0.74	0.92
	Two-Stage without APC and Spline ($K = 24$)	0.96	0.98	0.6	0.92
	Model Selection	**0.648**	**0.97**	0.553	0.98

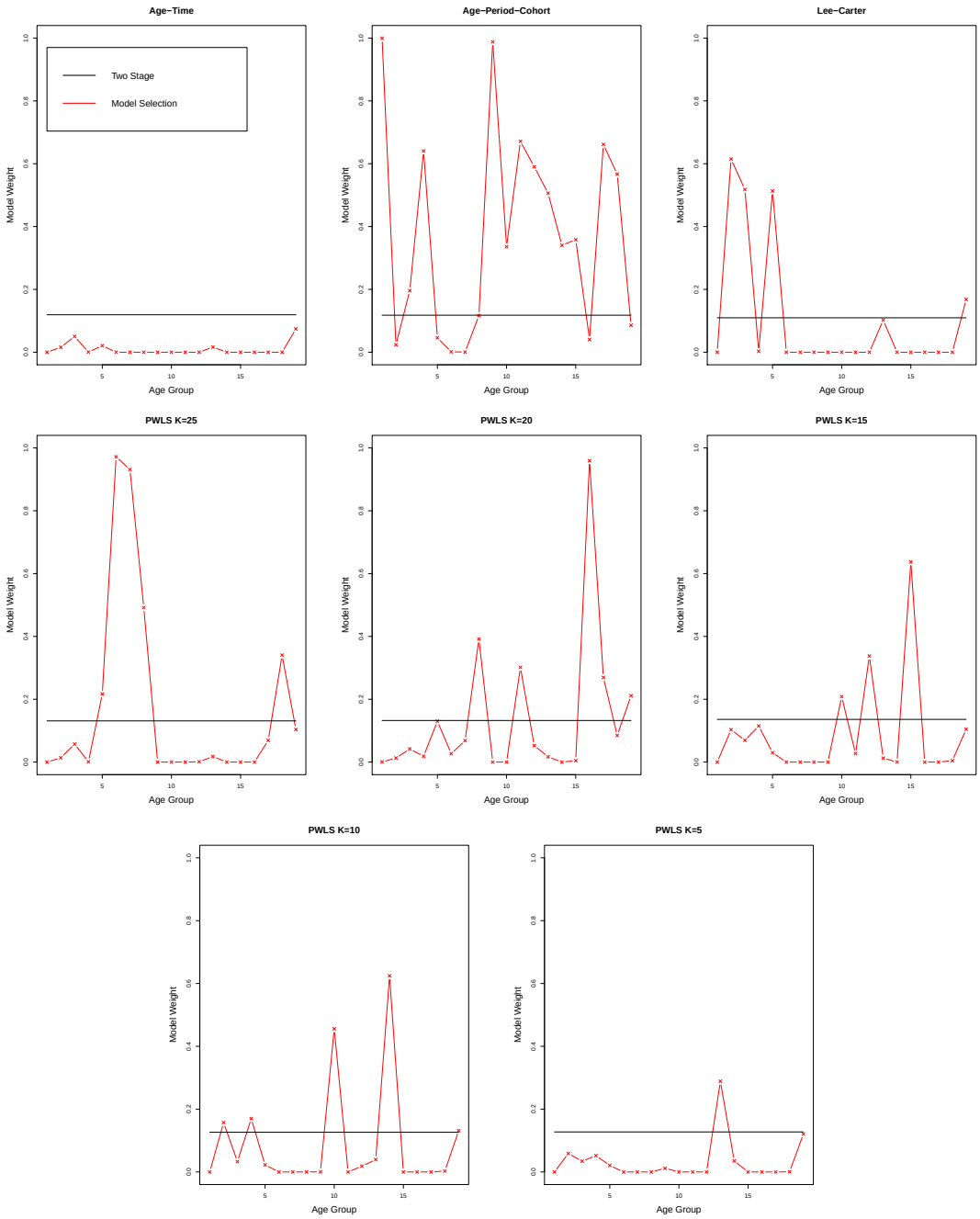
3.2 Investigation 2: Using a larger number of models

A second application is considered here to provide a case in which prior knowledge of model performance is not known, however more models are added to the ensemble, with less care in preselecting them. This is done through replacing both Piecewise Linear Spline models with a group of five spline models with the knots placed equidistantly at five year intervals. Although

using this strategy of multiple spline models with different knot positions can be considered as an alternative to estimating the true position of the knot, less prior thought is placed into which models are suitable to be used in the group. Therefore a suitable model weighting system is required which can adequately penalise the worse models and subsequently end up providing better forecasts.

Figure 3.1 shows the comparison of weights found for each model given both model selection (red) and two-stage approach (black) using the data of males during Window 1 (1977-2016). The first clear difference between the two weighting approaches is the application of age group specific weights for the model selection approach compared to the single weight given to each model through the two stage approach. Secondly, the more selective nature of the model selection approach can be found as some models, such as the linear Age-Time model, have very little weight given overall compared to the two stage approach which gives a very similar weight across all models. Noticeably, in this setting, the model selection weights are able to distinguish which models provide the better forecasts based on the in-sample fit. The best two examples of this are the weights given to the Age-Time and the Age-Period-Cohort models. Using Table 2.2, the Age-Time model can be found to give some of the worst forecasts when using the point estimate and uncertainty level accuracy. The Age-Period-Cohort model provides some of the best forecasts, alongside the two spline models. Figure 3.1 shows this performance of both models through model selection weights in which very little weight is given to the Age-Time model whilst a higher weight is given to the Age-Period-Cohort model more often than not. This is a feature which is not shown by the two-stage approach. This highlights key features of both approaches, in which the model selection approach is more strict in when estimating weights in comparison to the more lenient two-staged approach.

Figure 3.1: The model weights when considering more models from both BMA approaches for males in Window 1



In another key difference, the age specific weighting shows to be advantageous in forecasting as in all three windows for both males and females the model selection BMA approach provides more accurate point estimate forecasts than the two-stage approach. The more lenient two-stage approach does however provide more reliable coverage statistics, with all being closer to the nominal level of 95% than that of the model selection approach. This can also be explained by the differences in weighting approaches as the stricter model selection approach will lose the forecasting uncertainty from any model not selected. On the other hand with the two stage approach giving roughly equal weights to all models, it allows the forecasting uncertainty from all models to be included in the model averaging forecast. To show this, Table 3.2 provides the overall performance of each approach with the inclusion of the additional spline models given the three windows. Interestingly when compared to the overall results shown in Table 2.3, the performance of the model selection approach remains relatively consistent for both point estimate and coverage performance. In comparison the two stage approach provides less accurate point estimate forecasts whilst keeping similar coverage statistics.

Table 3.2: Overall mean forecast bias and 95% coverage after combining the three windows with the inclusion of five spline models

		Male		Female		Overall	
	Model	Mean Forecast Bias	95% Coverage	Mean Forecast Bias	95% Coverage	Mean Forecast Bias	95% Coverage
Single Model Approach	Age-Time	0.87	0.63	0.345	0.76	0.607	0.7
	Age-Period-Cohort	0.847	0.61	0.909	0.6	0.878	0.61
	Lee-Carter	0.86	0.7	0.47	0.86	0.665	0.78
	Spline ($K = 5$)	0.819	0.66	0.358	0.74	0.589	0.7
	Spline ($K = 10$)	0.718	0.68	0.365	0.68	0.542	0.68
	Spline ($K = 15$)	0.568	0.77	0.403	0.68	0.486	0.72
	Spline ($K = 20$)	0.486	0.78	0.503	0.74	0.494	0.76
	Spline ($K = 25$)	0.624	0.83	0.687	0.84	0.656	0.83
BMA Approach	Two Stage	0.672	**0.94**	0.413	**0.96**	0.543	**0.95**
	Model Selection	**0.454**	0.92	**0.298**	0.87	**0.376**	0.9

3.3 Discussion

In this chapter, two investigations are considered to show the importance of preselecting models for both BMA approaches. The findings of the investigations show that the model selection approach is able to provide well performing forecasts, even when a poorly forecasting model is included in the ensemble. Similarly, the model selection approach is also able to produce better forecasts when a larger number of models are used. In contrast, both scenarios are shown to alter the forecast performance of the two-stage approach. This suggests that more care is needed when preselecting models when using the two-stage approach as opposed to when using the model selection approach. Although, as in all forecasting, there should always be an element of care when deciding which types of model to use.

The inclusion of age group specific weights when using the model selection approach is also shown to be advantageous. This is expected as not all age groups exhibit the same mortality patterns over time. Therefore the age group specific weighting allows for the most appropriate models to be combined. Although the addition of age specific weights may make the comparisons between the two approaches unfair, the feature is unavailable when using the two-stage approach. If the projection bias on life expectancy is to be swapped with the projection bias on mortality rates at each age group in the weighting calculation, the scale of the mortality rates are far smaller than that of the life expectancy, making the projection bias smaller which subsequently makes the two-stage approach give even more lenient model weights.

More general comparisons can also be made between the two methods, by comparing the techniques used by both. A first difference being how the competing models are compared to

obtain model weights. In the two-stage approach shown model weights are based directly on the forecast bias of the model in a cross validation period. As the outcome of the application is to forecast, model weights are determined by forecast performance. However, the forecast bias used to determine a models weight neglects all uncertainty accompanying the forecast. In comparison, the model selection approach incorporates the in-sample uncertainty as the model selections are made at the iterative level. As each iteration can be seen as a sample from the posterior of each model, the full in sample uncertainty surrounding each model is included when using the approach over a large number of iterations. Therefore the model selection approach is more in line with the overarching goal of incorporating uncertainty. However, the two-staged approach does ultimately show to be good in terms of coverage due to the more lenient weights.

Another comparison can be found in the two-stage approach's necessity the split the data set to create a cross validation period. If the time series being used is short, the two-stage approach uses an even shorter time series to estimate model parameters during its first stage. Therefore when estimating model weights, the model parameters may not be reliable - resulting in the possibility of model weights being unreliable. In situations such as this, the model selection approach is preferable.

An extension to the project considered next is the transition from the country level, count based mortality data to individual level survey data, here provided by the Demographic and Health Surveys (DHS). A new way to interpret the DHS child mortality data is considered in Chapter 4 through a survival analysis approach, with a view to return to the model averaging framework in Chapter 5.

Chapter 4

Forecasting under-five mortality in a developing country context: A survival analysis approach

Under-five mortality is used to track the progression of many developing countries. To continue to improve in this area the United Nations has set a series of goals for each country to reach by 2030. Here we show a novel approach towards under-five mortality forecasting. This proposed method stems from the modelling of the Demographic and Health Surveys (DHS) data, in which the data are collected from developing countries through surveys and are available for a limited numbers of years. The limited number of survey years from the DHS data poses a challenge to the conventional modelling of mortality forecasting. This has been combatted by following each birth cohort to then use as a time component (Rutstein and Rojas, 2006) in a regression model.

Here we also use the birth cohorts as the time component. However, rather than modelling mortality rates through a more conventional Bernoulli or Poisson model, a novel survival analysis approach is taken.

In our approach, the time to event data are modelled using Weibull models, where the shape and scale parameters are specific to each birth cohort. Flexible time series models (such as linear and spline models) are considered in order to capture the variation of these cohort-specific parameters with a view to forecasting.

In this chapter, we also show a cross validation study to evaluate the forecast performance of the survival models against a more traditional approach. Forecasts to 2030 are also produced allowing us to compare against the Sustainable Development Goals set by the UN.

4.1 Introduction

Under-five mortality is a key statistic to measure the healthcare in low and middle income countries, therefore it is vital that available data can be both analysed and forecasted appropriately.

In 2000 the United Nations announced a set of Millennium Development Goals (MDG; UN 2000) including twenty one targets to help to achieve eight overarching goals by 2015. These goals involve improvements to healthcare such as reducing child mortality (Goal 4) and eradicating extreme poverty and hunger (Goal 1). The outcome from Goal 4 showed excellent results with substantial progress being made towards the reduction of child mortality, with

every region, apart from Oceania, showing a 50% (or more) reduction in under-five mortality (Way, 2015). It was also deemed from the report that further progress could be made, in particular by focusing on neonatal deaths. This was outlined in the next set of targets that the UN released called the Sustainable Development Goals (SDG; UN 2015). The SDG's are set in a similar manner the MDG's with a set of 17 goals, each with their own list of targets, hoped to be achieved by 2030. The specific target that is looked at here is Target 3.2:

"By 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births and under-5 mortality to at least as low as 25 per 1,000 live births." UN (2015)

Here, neonatal mortality is defined as death of children less than one month old and under-five mortality is death before the child's fifth birthday.

The SDGs are set for all countries within the UN. Here we focus on Sub-Saharan Africa and use the Demographic and Health Surveys (DHS; ICF International 1984). The DHS provide limited survey data which pose a number of challenges and allow different interpretations to modelling.

Recently, Mejía-Guevara et al. (2019) produced under-five mortality forecasts of 31 Sub-Saharan African countries and compared them with the SDG target. In this case, different modifications are made to DHS data, such as using the UN IGME estimates to adjust the data. After the modifications are made a variant of the Lee-Carter model (Li et al., 2004) is used to capture any age and time trends to then forecast both neonatal and under-five mortality to 2030. We follow a similar method in our application (Section 4.4) to show competing forecasts.

A different approach to analysing the under-five DHS data has been shown as using Bernoulli models including smoothing spatial effects within countries (Mercer et al., 2015). Here adjustments are also made to the DHS data, such as an HIV adjustment, before random effect Bernoulli models are used to estimate age and spatial effects. Projections are also made from the model, showing the probability of countries achieving the MDG. Here, a different view of the the data is shown through using a novel survival analysis approach.

4.1.1 Survival Background

Survival data analysis is used to study 'time-to-event' data, in which observations can be followed for a certain length of time until an event of interest occurs. Due to the wide varieties of contexts that can be modelled using survival analysis, the definition of the event can vary. For example in engineering Khalaf et al. (2013) use the failure of medical equipment to show that the age of the equipment does not affect its survival, however when preventative maintenance is performed the lifespan of the equipment improves. The usage of the survival analysis concept can even be shown in tourism management as Falk (2013) shows the survival of ski-lift companies, with the conclusions that if snowmaking facilities are available from the start of the project the business is more likely to survive. In our setting of under-five mortality we can measure the survival of children under different settings in which the event is death. Of course, in any survival analysis, there will be some observations which never experience the event. These observations can remain in the study using them as a censored time. Censoring can occur for many reasons such as, if a patient drops out of a study they may not experience the event. Alternatively, they may experience a different event making further analysis impossible (Clark et al., 2003).

Survival analysis techniques are commonly applied to compare the different characteristics of populations under different settings. An example could be comparing the effects of different treatments (Branson and Whitehead, 2002) in which two (or more) groups using specific treatments are observed and then modelled to show the different advantages/disadvantages within treatments. Another example is shown by Min et al. (2011) who examine how different student backgrounds affect the dropouts in an undergraduate engineering major.

Another way of viewing a survival analysis approach is to find whether any covariates are related to the event in question. Kardaun (1983), for example, uses the stage of cancer and age as main covariates to show how both effect the survival of male cancer patients. Alternatively, in the mortality setting Ayele et al. (2017) investigate contributing factors to under-five mortality in Ethiopia, in order to attempt to help continue the already declining mortality rate.

Here the predictive ability of survival models is used in which parameters can be estimated using a regression setting, allowing for an increase in time to be used to produce future forecasts. An example of this is shown by Humble et al. (2006). Humble et al. (2006) use the area level Human Mortality Database data to model the probabilities of dying given a certain age. In this case Weibull models, using different functions of time (e.g. linear across calendar year) were used to capture the survival patterns of each of the ages.

To model survival times we start with a time T assuming it is random and continuous the cumulative distribution function is defined as

$$F(T) = Pr(T \leq t), \qquad t > 0$$
$$f(t) = \frac{dF(t)}{dt}$$

$F(t)$ can be described as the probability that a candidate experiences the event before time t with $f(t)$ being its density function. Two common descriptive probabilities are used in survival analysis, the survival function and the hazard rate. The survival function, $S(t)$, is the probability that a random candidate will survive until time t, therefore at $t = 0$ $S(0) = 1$ and over time $S(t)$ decreases until eventually $S(\infty) = 0$. This is shown as

$$S(T) = Pr(T > t) = 1 - F(t)$$

The density function, survival function and hazard function are related as

$$f(t) = h(t)S(t)$$

therefore,

$$h(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} = -\frac{d}{dt}\log(S(t))$$

The hazard function (also known as the instantaneous failure rate), $h(t)$, is the probability that the candidate experiences the event at time t given that they have survived to time t. This can also be transformed to a cumulative hazard function, $H(t)$, as

$$H(t) = \int_0^t h(u)du = -\log(S(t))$$

In survival analysis the models used mainly fit into two classes - Proportional Hazards (PH) and Accelerated Failure Time (AFT) models. The PH models use a structure in which the model covariate effects are found through modelling the hazard function whilst the AFT models apply covariate effects to the log survival time (Ali et al., 2015). In this study the Weibull distribution is used to model survival of under fives. This distribution is unique as it is a fully parametric, flexible model that can be represented as both a PH and a AFT model (Wang et al., 2018). The representation used here is using a Weibull Accelerated Failure Time model in which the survival time of under-fives are modelled. It is noted that the semi-parametric Cox Proportional Hazards model (Cox, 1972) is widely used in the field, however in this study the Weibull distribution provides a nice functional form from which a cohort structure for forecasting can be applied.

4.1.2 DHS Data

The data used is obtained from the Demographic and Health Surveys (DHS). The main objective of the DHS Program is to improve the collection, analysis, and dissemination of population, health, and nutrition data and to facilitate use of these data for planning, policy-making and program management in a developing country setting (Croft et al., 2018). For information relating to child health the DHS interview women aged between 15-49 who slept in the household the night before. To identify which households to interview, the DHS use a two stage

cluster sampling procedure. In the first stage a sample of enumeration areas (EAs) that cover the particular country are chosen with probabilities proportional to the size of the EA. All of the occupied residential households within each of the selected EAs are then listed in which, to reduce the costs involved with listing the households, each selected EA is reduced to segments with a population of no less than 500 ($\approx$ 100 households) (ICF International, 2012). In the second stage households are selected using a systematic sampling technique, with the chosen households interviewed for the surveys.

Across the DHS there is information from over 90 countries however the number of surveys for a lot of these countries is very limited. Overall the average range of surveys for a given country is approximately 3-6 and we focus on those with more surveys, and therefore more observations.

Within the DHS the full birth histories of the participating children are recorded and are used to study child mortality. This framework allows for individual level observations to be listed. The types of survey questions asked involving child health are the date of birth of the child and the age at death of the child if they have died. Other variables recorded are the gender of the child, date which the survey was taken as well as more individual level covariates such as whether the household is in a rural or urban area and the education level of the mother.

It is noted that there are risks when using survey data and therefore to use this kind of data we must make assumptions. Firstly we have to trust the questions asked (Whether the child is alive, Age at death, Date of birth) have been answered accurately. Secondly we assume that the survival of the mother does not have an effect on the mortality of the child. The latter

assumption is linked with survivor bias in which the mothers surveyed have to be alive to be questioned. A well known example of survivor bias is the influence of HIV. Mothers who test positively for HIV have a higher risk of death and are less likely to be surveyed. This means that children of HIV positive mothers are less likely to be included in the data and have also been known to have a higher risk of dying before the age of five (Wakefield et al., 2019). Another risk of using the survey data is that it is not a registry based dataset and is not fully representative of the country in question. Sample weights are included in the survey to help combat this.

It has been shown that a difference in mortality patterns exist between genders in under-five mortality within low/middle income countries (Costa et al., 2017).

To check this in the DHS setting, gender was used as a covariate in a Weibull model using two surveys from each country. The surveys used were from the 1993 and 2014 Kenya surveys and the 1995 and 2016 Uganda surveys to show that the gender of the child has an effect on the mortality pattern in both more recent and older time periods. The results are shown in Table 4.1. Using this method it can be shown that the inclusion of the gender covariate is significant for both countries during both earlier and later survey years. To account for these changes in mortality between males and females, the data for the two genders are split and modelled independently.

Table 4.1: Output from Cox PH models over Kenya and Uganda showing the difference in mortality patterns between males and females over time

Survey	Covariate	Mean	Std. Deviation	2.5%	97.5%
Kenya 1993	Gender	-0.1270	0.0632	-0.2512	-0.0030
Kenya 2014	Gender	-0.1573	0.0276	-0.2115	-0.1032
Uganda 1995	Gender	-0.1000	0.0457	-0.1898	-0.0103
Uganda 2016	Gender	-0.1721	0.0266	-0.2244	-0.1199

Due to the survey structure the challenge of missing data is encountered here. This missing data accounts for a very small proportion of observations ($\frac{19}{157575}$ for Kenya and $\frac{20}{151953}$ for Uganda) and is assumed that removing these missing entries will not produce bias in the parameter estimates. The missing entries are therefore omitted from the study.

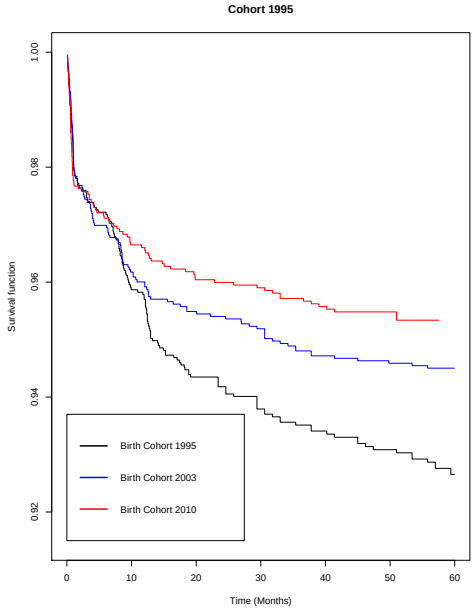
4.2 Methodology

4.2.1 Survival Approach

Our technique involves separating each child depending on the year in which they were born. Each birth cohort can then be analysed through using a survival analysis approach. The use of this survival context allows for the observations with deaths at under 60 months old to remain uncensored whilst the observations that survive can be censored at their age when the survey was taken. It should also be noted that any observations that are tracked and have survived longer than five years are censored at 59 months. An easy way to visualise each cohort of data is through survival functions. Figure 4.1 shows Kaplan-Meier plots of three separate birth

cohorts over the five year age range using data from Kenya. Kaplan-Meier plots are used as an estimator to show the true survival curve for a given dataset (Goel et al., 2010). It can be seen clearly from the figure that in this selection of cohorts, as birth cohorts increase, the survival rates of the under-five children increase (most notably for the ages after 10 months), and thus the mortality rate is dropping.

Figure 4.1: Kaplan-Meier plots of three separate birth cohorts showing the increased survival of under 5 children over time



4.2.2 Exploratory Weibull plots

Here we use the parametric Weibull distribution to model the data. The full density of the distribution is

$$f(y|r,\mu) = r\mu y^{r-1}e^{-\mu y^r} \quad \text{where} \quad \mu, r > 0 \tag{4.1}$$

and

$$h(y|r,\mu) = r\mu y^{r-1} \qquad S(y|r,\mu) = e^{-\mu y^r}$$

in which the shape and scale parameters are r and μ respectively, and the data (in our case time until death or censored age) is shown as y. To further explore the data, a Bayesian Weibull model is fitted independently to each of the 31 birth cohorts from the Kenya surveys, taking the form:

$$\begin{gathered} y_i \sim Weibull(r,\mu) \\ log(\mu) = \alpha \\ \alpha \sim N(0, 10000) \qquad r \sim Gamma(1,1) \end{gathered}$$

where i represents each survival time and $i = 1, ..., N$.

To understand the effects of the shape and scale over time, the interpretation of both parameters must be known. Firstly, the shape parameter controls how the hazard rate changes over time. The value of the shape can be interpreted as when $r < 1$ indicates that the hazard rate decreases over time, a value $r = 1$ shows that the rate is constant (and the Weibull reduces to

an exponential) whilst when $r > 1$ the hazard rate increases over time. In under-five mortality we find that deaths are more frequent the younger the child is and therefore the older the child gets the more likely they are to survive. Since this shows that the hazard rate decreases over time, we find that the cohorts have estimated shape parameters less than 1, justifying the $Gamma(1, 1)$ prior. As the name suggests, the scale parameter is used to adjust the model to the scale of the data. When using the same time ranges (as here in which the age range is 0-60 months) the scale parameter can be used to show any changes in mortality level over time. The interpretations of these two parameters are shown visually through survival functions in Figure 4.2 in which the left plot shows changes to the shape with μ fixed at 2, and the right shows changes to the scale with r fixed at 0.5.

Figure 4.2: Showing the effects of changing the shape and scale parameters in a Weibull distribution

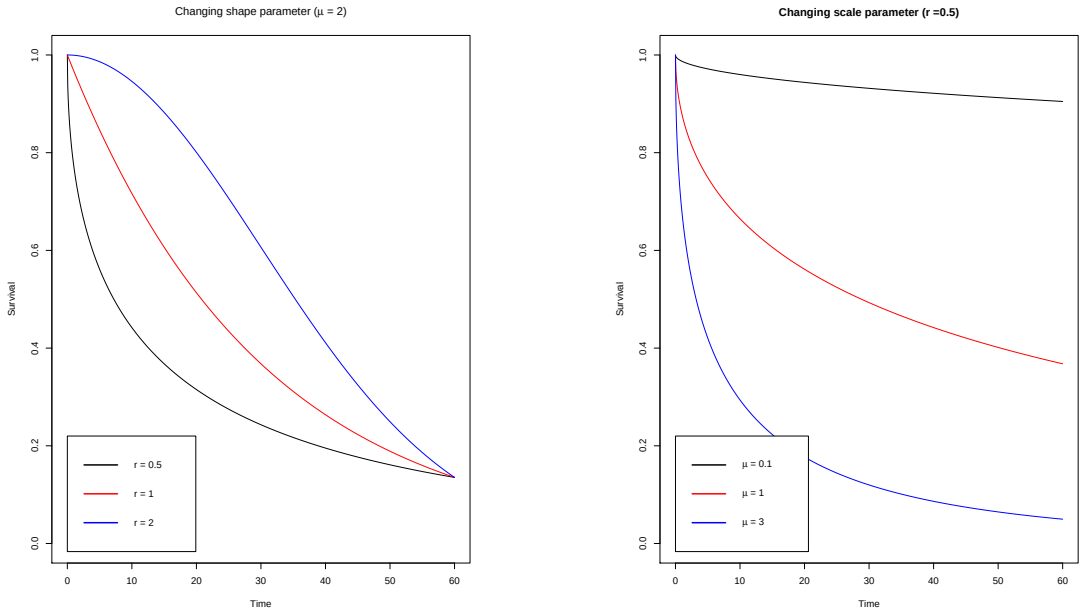


Figure 4.3 (left) shows the posterior distributions of the scale parameters estimated for each birth cohort. The drop in mortality rate can be clearly seen when following each cohort. It is also shown that the last two cohorts show a rise in scale parameter as well as increased uncertainty around the estimates. Figure 4.3 (right) is a visual representation of how many years each birth cohort can be observed for, given the surveys available. The figure shows that the most recent cohorts are not able to hold a full five years of data and this is shown in the model fitting. Figure 4.4 shows the difference in model fitting when using less observed time. Because the model has no information from the life experience of the child after a certain age, it can only fit to the data available. Therefore when fitting a model to a single cohort, if the cohort is not fully observed, the unobserved ages are assumed to follow the same trend as the observed ages. Another challenge can also be found here in that the less time a cohort has to be observed, the lower the frequency of observations that it tends to have. Therefore when modelling the cohorts with lower numbers of observations, the uncertainty around the parameter estimates are greater. The number of observations, as well as median ages of death are shown in Table 4.2.

Table 4.2: The frequency of observations as well as the total number of deaths per birth cohort for female Kenya data

Birth Cohort	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Deaths	212	215	233	252	231	203	250	209	260	197	212	169
Median Age at Death	5.95	5.65	6.95	6.30	6.05	8.05	6.45	5.55	7.65	6.45	8.35	8.45
Total Observations	2532	2349	2816	2837	2947	2533	2550	2311	2797	2224	2351	2371
Birth Cohort	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Deaths	270	235	201	193	240	186	214	135	151	132	147	144
Median Age at Death	7.85	4.25	6.25	6.65	5.90	6.30	4.80	3.55	2.55	4.05	2.55	3.10
Total Observations	2876	2722	2538	2489	2932	2480	3256	2638	2618	2473	2777	2816
Birth Cohort	2008	2009	2010	2011	2012	2013	2014					
Deaths	147	97	98	86	74	75	30					
Median Age at Death	3.45	2.15	0.58	3.60	0.65	0.50	0.07					
Total Observations	3079	1961	2147	2092	2032	2132	1190					

Figure 4.3: Scale parameter estimates for birth cohorts fitted individually

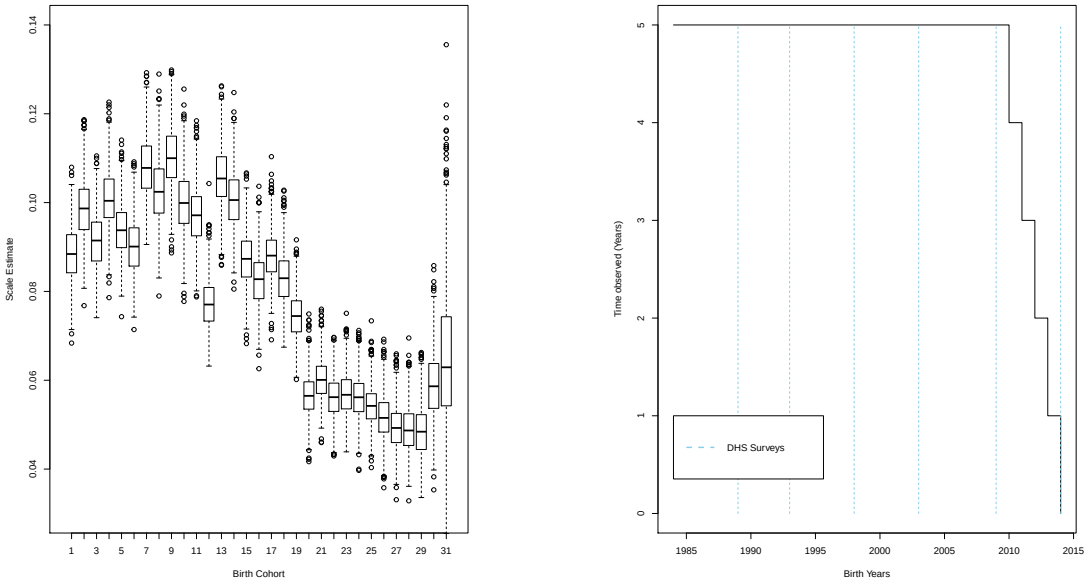
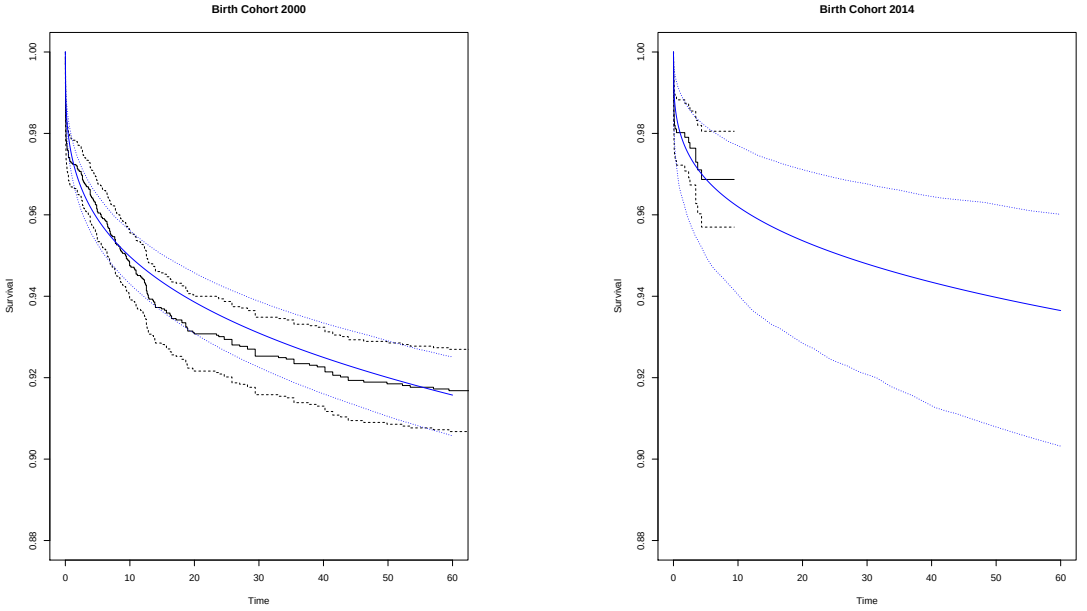
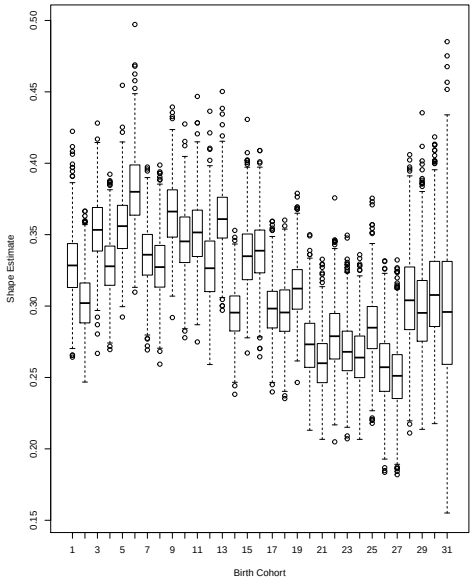


Figure 4.4: Showing the fitting of a Weibull model when using a reduced observation time



The equivalent plot for the shape parameters is shown in Figure 4.5 where the posterior distributions of the individual cohort shape parameters are shown. Similarly to the scale, the most recent cohorts, with less amount of time to be observed, show higher uncertainty in the posterior distributions. However even with the changes in uncertainty, it can be shown that the trend of the scale parameters is relatively constant, meaning that the main trend of the reducing mortality rates is captured by the scale.

Figure 4.5: Posterior distributions of shape parameter estimates over birth cohorts



4.2.3 Age Groups

One of the findings that Mejía-Guevara et al. (2019) show highlights the importance of monitoring the child survival of different age groups. As we also find the different age groups to show different mortality patterns, we split the data into the same groupings being the neonatal ([0-1) months), post neonatal ([1-12) months) and child ([12-60) months) ages. In fitting the Weibull model to the data without using age groups, due to the neonatal age group having a high mortality rate in a short space of time, the model overfits the earlier ages and therefore sacrifices the model fit to the later ages. If, for example, a linear trend model is then placed on the scale it will be found to have a small gradient due to the neonatal age groups not changing as much over time. In turn this is not representative of the entire under-five lifespan and so for a better fitting model the age groups are used. It should be noted that the shape parameter is found to be constant over time for both when using age group specific shape parameters and when using an overall shape parameter, allowing only the scale parameter to be forecasted.

Figure 4.6 (top left, top right) shows the difference in survival function when using the age groups. Here a vague Weibull model, as shown in Section 4.2.2, has been fitted to a single cohort of data (birth cohort 1995). One model fitting has been done with no grouping (top left) whilst the other uses the three age groups to fit the shape and scale (top right). When the age groupings are used the Weibull model is fit to each group independently. Due to the independent fitting, each age group has its own survival function, leading to three separate model estimates for survival curves. In comparison to using no age groups, Figure 4.6 (top right) shows the three separate survival curves, estimated from the three independent age groups, in which the survival curves are shown with their corresponding age group. It is important when

viewing these curves not to see them as a single combined survival function (notice the overall curve is not a cumulative function), however when viewing the three age groups sequentially it provides an excellent comparison to the original method.

Although it is noticeable from the two plots, Figure 4.6 (bottom) shows a Cox-Snell residual plot acting as a model diagnosis tool. The calculation of Cox-Snell residuals in Bayesian analysis is shown by Chaloner (1991) and is defined as

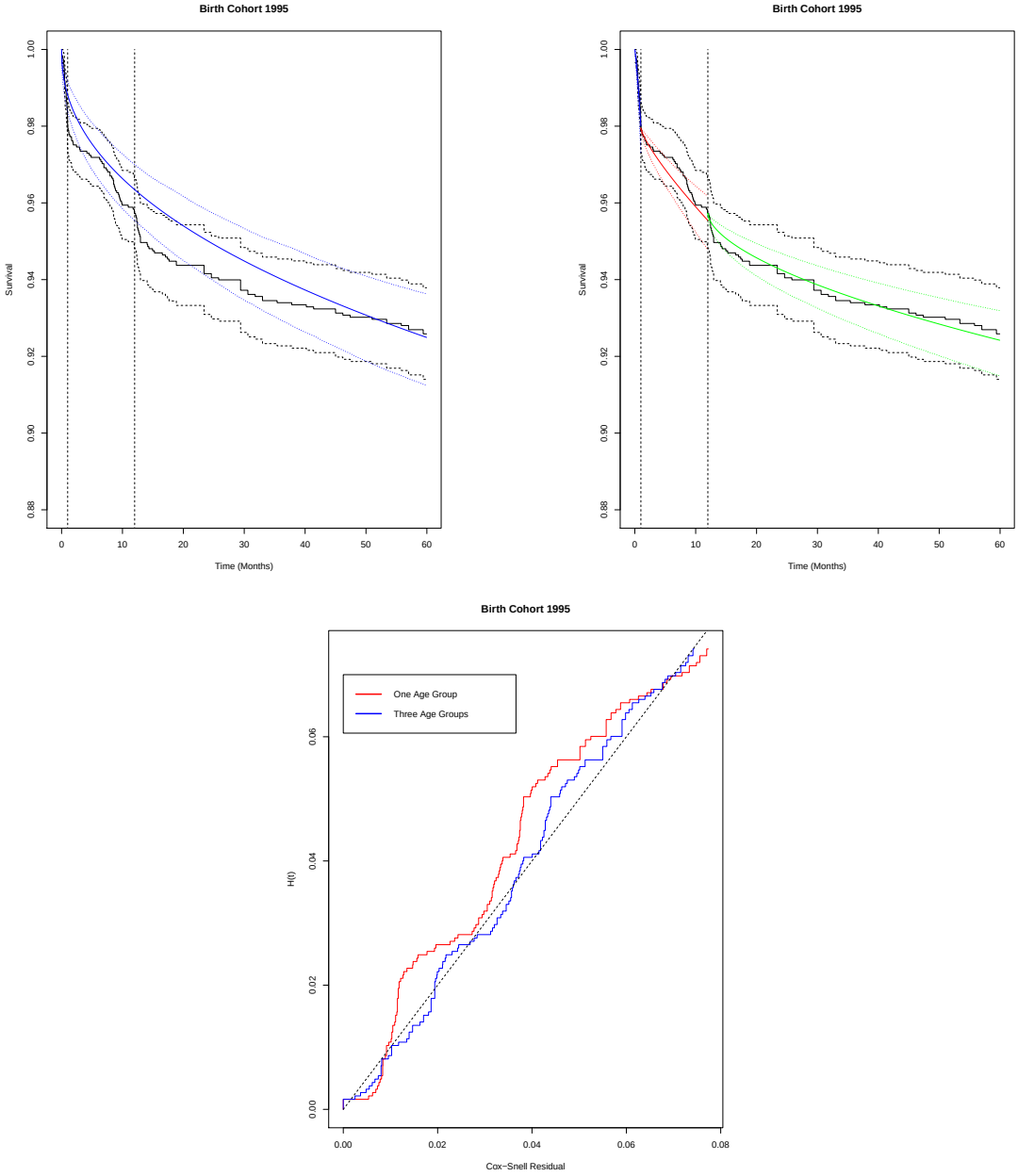
$$r_{C_i} = H_i(t_i, \mathbf{\Phi}|\mathbf{x}_i), \qquad i = 1, ..., n$$

and therefore, as shown in Section 4.1.1,

$$= -\log(S(t_i, \mathbf{\Phi}|\mathbf{x}_i)), \qquad i = 1, ..., n$$

in which $\mathbf{\Phi}$ is a vector of the parameters used in the model and $\mathbf{x}_i$ is the vector of covariates. If the model shows a perfect fit, the Cox-Snell residuals would match the cumulative hazard function of the data. This is then shown in the plot where the Cox-Snell residuals of both the model fits are plotted against the estimated cumulative hazard function. A perfect model fit would follow the $y = x$ line (shown as a black dotted line) and therefore the closer the model fits are to that line, the better the model is performing. Figure 4.6 (bottom) shows the Cox-Snell residuals when using no age groups, and three age groups (when using posterior means) plotted against the cumulative hazard function of the data. As can be seen from a visual comparison of the two survival functions, splitting the data into age groups provides not only a more accurate fitting but also can then be used to capture the separate mortality trends within the groups.

Figure 4.6: Plots showing the benefits of splitting the observations into three groups before modelling (Neonatal, infant and under five)



4.2.4 Weibull Models

To forecast from the distribution a log-linear trend can be put onto the age-specific scales whilst using the birth cohorts as time components. Using Y as the lifetimes of the under-five children, the linear model can be expressed as:

$$\begin{aligned} y_i &\sim Weibull(r_{A_i}, \mu_{A_i C_i}) \\ r_{A_i} &\sim Gamma(1,1) \end{aligned}$$

$$log(\mu_{A_i C_i}) = \alpha_{A_i} + \beta_{A_i} \times C_i \tag{4.2}$$

$$\begin{aligned} \alpha_{A_i} &\sim N(0, 10000) \\ \beta_{A_i} &\sim N(0, 10000) \end{aligned}$$

where A_i defines the age group of the ith observation and the birth cohort is defined by C_i in which C is the numerical value of the cohort, going in ascending order with the most recent cohort being 0. Here α_A can be seen as the intercept whilst β_A the gradient of the linear trend. As each age group is modelled independently, the vague priors are assigned separately to each of the age specific parameters, and so there is no smoothing across ages.

For a second model, a piecewise linear spline model is chosen to help capture any breaks in the mortality patten, as shown in Figure 4.3 around birth cohort 18. The form of the spline is chosen to be piecewise linear as splines with higher orders (e.g. cubic spline) can lead to overfitting, i.e. providing a better in-sample fitting resulting in unrealistic forecasts. As shown by Foreman et al. (2017) where the non-linear spline models used to forecast epidemic diseases result in larger forecasting errors in a cross validation setting due to these non-linear terms. The spline is chosen to have one knot (K) and is to be $\frac{3}{5}$ up the birth cohorts, allowing for the most recent cohorts to determine the change in gradient whilst also providing enough information from the cohorts that are able to be observed for the full five years. It should be noted that given the current timescale of data, this choice of knot is also able to capture any changes to mortality that the millennium development goals have contributed to. The spline model is shown as

$$
\begin{aligned}
y_i &\sim Weibull(r_{A_i}, \mu_{A_i C_i}) \\
r_{A_i} &\sim Gamma(1,1)
\end{aligned}
$$

$$
log(\mu_{A_i C_i}) = \alpha_{A_i} + \begin{cases} \beta^{(0)}_{A_i} \times C_i, & \text{if } C_i \leq K \\ \beta^{(0)}_{A_i} \times K + \beta^{(1)}_{A_i} \times (C_i - K), & \text{if } C_i > K \end{cases} \tag{4.3}
$$

$$
\begin{aligned}
\alpha_{A_i} &\sim N(0, 10000) \\
\beta^{(0)}_{A_i} &\sim N(0, 10000) \\
\beta^{(1)}_{A_i} &\sim N(0, 10000)
\end{aligned}
$$

4.3 Inference using INLA

Both Weibull models were fitted using Integrated Nested Laplace Approximation (INLA; Rue et al. 2009) within the R statistical software (R Core Team, 2013).

INLA is restricted to a class of models called latent Gaussian models of which covers a wide range of models and encompasses the additive regression models being used here for forecasts. Similarly, for computational efficiency, the latent field has to be a Gaussian Markov Random Field (GMRF) meaning that, in this case, the priors on the regression covariates must follow a multivariate normal distribution. Again, this assumption does not change the model fitting as the covariate priors have been uninformative Normal distributions.

Rather than using MCMC iterative methods, INLA utilises the Laplace approximation. The Laplace approximation allows for a density function to be approximated by a Normal distribution. The mean of the approximating Normal distribution is the mode of the density, found through solving $\frac{\partial log f(x)}{\partial x} = 0$, where $f(x)$ is the density to be approximated. In the INLA package this is done using a quasi-Newton method. The variance of the Normal distribution can then also be found by evaluating $-1/\frac{\partial^2 log f(x)}{\partial x^2}$ at the point of the mode. This approximation is made on the joint posterior distribution $Pr(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y})$ where $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are the sets of parameters and hyperparameters respectively and $\mathbf{y}$ is the data. It is expected that the approximation will be accurate given that $Pr(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y})$ is almost Gaussian given the assumption of Normally distributed priors. Using this approximation, firstly the joint posterior of the hyperparameters $Pr(\boldsymbol{\psi}|\mathbf{y})$ can be approximated through

$$Pr(\boldsymbol{\psi}|\mathbf{y}) \approx \frac{Pr(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\psi})Pr(\boldsymbol{\theta}|\boldsymbol{\psi})Pr(\boldsymbol{\psi})}{\tilde{Pr}(\boldsymbol{\theta}|\boldsymbol{\psi},\mathbf{y})}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\psi})} = \tilde{Pr}(\boldsymbol{\psi}|\mathbf{y}) \tag{4.4}$$

where $Pr(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\psi})$ is the likelihood, $Pr(\boldsymbol{\theta}|\boldsymbol{\psi})$ is the multivariate normal prior on $\boldsymbol{\theta}$, $Pr(\boldsymbol{\psi})$ is the priors on the hyperparameters and $\tilde{Pr}(\boldsymbol{\theta}|\boldsymbol{\psi},\mathbf{y})$ is the Laplace approximation of $Pr(\boldsymbol{\theta}|\boldsymbol{\psi},\mathbf{y})$. This is evaluated at $\boldsymbol{\theta}^*$ which is the mode for a given $\boldsymbol{\psi}$. Using this grid of points $\tilde{Pr}(\psi_i|\mathbf{y})$ can be used to approximate the full conditional distribution $\tilde{Pr}(\theta_i|\boldsymbol{\psi},\mathbf{y})$ and $\boldsymbol{\psi}$ can be integrated out to obtain the marginal posterior for $\boldsymbol{\theta}$

$$\tilde{Pr}(\theta_i|\mathbf{y}) \approx \int \tilde{Pr}(\theta_i|\boldsymbol{psi},\mathbf{y})\tilde{Pr}(\boldsymbol{\psi}|\mathbf{y})d\boldsymbol{\psi} \approx \sum_j \tilde{Pr}(\theta_i|\boldsymbol{\psi}^{(j)},\mathbf{y})\tilde{Pr}(\boldsymbol{\psi}^{(j)}|\mathbf{y})\Delta_j \tag{4.5}$$

for a set of integration points j, and a set of weights Δ_j. For further detail, explanation and a worked example on the workings of INLA the reader is referred to Blangiardo and Cameletti (2015).

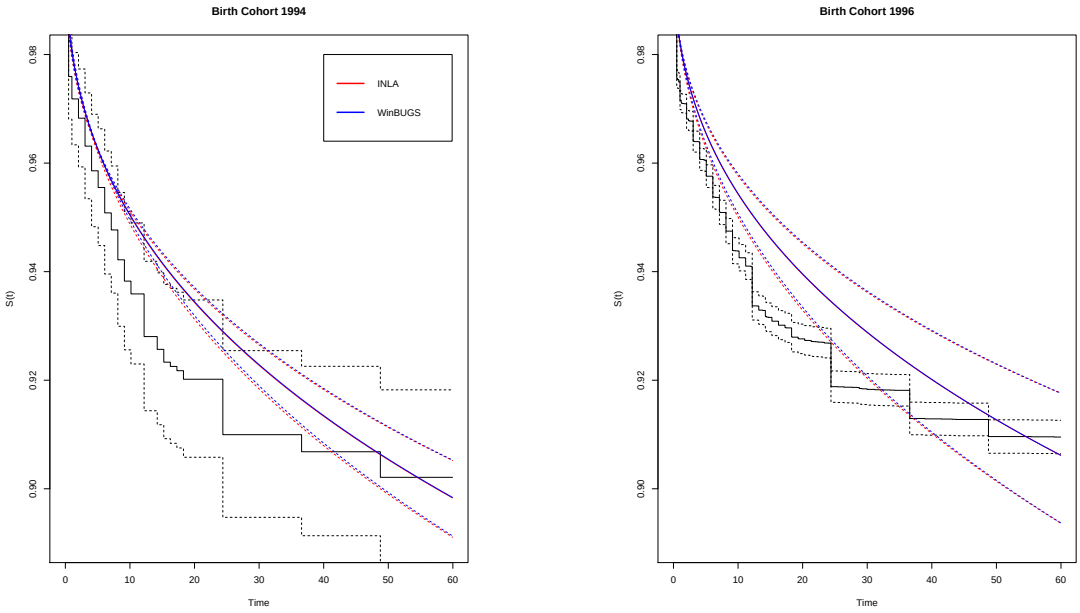
The INLA approximation offers a much faster alternative to the MCMC method of WinBUGS whilst retaining the accuracy of parameter estimation not only for posterior means but also for uncertainty. To test this the same models were fitted using both INLA and WinBUGS to compare the model fits. Five birth cohorts were selected (1991-1996) out of the Kenya dataset and a linear Weibull model was used (without the splitting of age groups) to compare any differences between the fits. The model used is defined similarly to that in Equation 4.2 although the parameters are not age group specific. Table 4.3 shows the parameter estimates found through using the methods. It shows that INLA can provide very similar estimates to the

MCMC alternative with the advantage of approximating the posterior distributions, making the computation much faster. To show that the differences in estimation do not have an effect on the model fits, the survival functions of two of the five cohorts are shown in Figure 4.7 along with the survival functions of the model fits. It is clear that INLA is able to match WinBUGS in providing accurate posterior distributions of each parameter.

Table 4.3: Comparing parameter estimation between WinBUGS and INLA

Parameter	WinBUGS Estimate (95% CI)	INLA Estimate (95% CI)
α	-2.32 (-2.454, -2.186)	-2.32 (-2.456, -2.188)
β	-0.043 (-0.094, 0.009)	-0.043 (-0.095, 0.008)
r	0.42 (0.39, 0.45)	0.42 (0.39, 0.45)

Figure 4.7: A visual comparison of an INLA fitting against a WinBUGS fitting



When using INLA in this setting we find a disadvantage in which some complexity is lost when modelling the shape parameter. An example of this is that if it is perceived that the shape parameter in a Weibull model may change over time, a trend model like a linear/spline model cannot be placed on the shape (unlike the scale). For the datasets that have been used here this is not a problem as the shape parameters remain relatively similar over time (for Kenya this is shown in Figure 4.5) however for datasets from other countries this may need to be considered. Due to the limitations when using the shape parameter it was decided the age groups needed to be completely independent of each other when the survival models were fitted. This allows a separate shape parameter estimation for each age group and is essential for an appropriate fitting, as the three age groups show different changes to the hazard rate over time. To show the difference the change makes, a single shape parameter (assuming a constant shape over the three age groups) is estimated to have a posterior mean of $r = 0.53$ with 95% credible interval $(0.51 - 0.55)$. However, when using independent shape parameters per age group gives posterior means of $r_1 = 0.36$, $r_2 = 0.91$ and $r_3 = 0.52$ with credible intervals $(0.34 - 0.38)$, $(0.86 - 0.96)$ and $(0.49 - 0.55)$ when both fittings use a linear Weibull model. To make this change the age groups have to be assigned their own datasets.

To do this the full dataset needs to be replicated three times for each age group. The first age group can use the entirety of the data with the adjustment of any deaths with an age of death greater than (or equal to) one month were changed to be observed as alive and censored at 0.99 months, the limit of this age range. Any censored observations with an age equal to or over one month were also censored at 0.99 months. For the second age group, any observations said to have died or been censored at less than a month old were removed. Then since the survival model assumes that each survival time starts from time zero, each age at

death or censored age equal to or greater than one month and less than twelve months were subtracted by one month. The remaining ages at death and censored ages (≥ 12 months) were censored at 10.99 months, as this is the oldest age after the one month subtraction. Similarly for the final age group, any times of death or censored times less than twelve months old were removed, with the remaining times subtracted by twelve months.

To avoid numerical overflow it is recommended in the INLA documentation to scale all times to make the maximum 1. With the three age groups being fitted separately, each is individually scaled down to run the models then any simulations from the model can be scaled back appropriately. This should not be classed as a disadvantage to using INLA as it only requires a transformation to have any output on the preferred scale.

4.4 Application

4.4.1 Transformation to mortality rate

Since the survival analysis approach outputs data in which the birth cohorts are used as the time component, for a model comparison it is necessary to transform the output to match that of a more traditional mortality approach in which the year of death is used to measure time.

This is possible by taking the shape and scale parameters (at the iterative level) and sample from the Weibull distribution (Equation 4.1) to receive simulated ages at death. For each birth cohort, a population of new borns of size 2000 is followed over time and the time of death of each baby is simulated using the estimated shape and scale parameters. Since the shape

and scale parameters are also age-specific, the times at death are simulated based on the ages of the children as follows. A set of event times for the 2000 new borns is generated from the Weibull with scale = μ_{1C} and shape = r_1. Individuals whose event times are less than 30 days are considered to have died within the first month and their event times are the times of death. Another set of event times for the remaining individuals (there are 2000-n_1 of them with n_1=number of deaths within the first month) is then generated with scale = μ_{2C} and shape = r_2 and the event times that are above 1 and less than 12 months are recorded as times of death. Finally, a third set of event times is generated using scale = μ_{3C} and shape = r_3 for the surviving individuals (now the number is 2000-n_1-n_2 with n_2=number of deaths within 1-12 months). The times between 12-60 months are recorded as times at death. Here, the values for μ_{AC} and r_A (for $A = 1, 2$ and 3) are drawn from their marginal posterior distributions over 3000 times (thus there are 3000 sets of such parameters) and the above procedure is carried out 3000 times to reflect the uncertainty in the resulting mortality rates. In practice we find that any changes to this does not effect the forecasts as long as the frequency of simulations are not too low.

The other information we need to transform the ages at death into the mortality rate is the date of birth. As the cohort specific age parameters are used to simulate ages at death, the birth year of each simulation is already known. To allow for any skewness in the dates of birth throughout the year, each date of birth is sampled from the in-sample data. Due to the different frequencies of observations across birth cohorts, here date of births are sampled from the specific cohort as well as its two neighbouring cohorts. This is done for each cohort apart from the most recent five, which show the lowest number of observations. Dates of births for these cohorts are sampled from a pool of the last five cohorts.

Once we have these simulations the transition to mortality rate is performed in the same way as that of the true data. This is shown by Hill (2011) and is also the method used by Mejía-Guevara et al. (2019). The information needed to transition to the mortality rate is whether the observation has died, the age at death and the date of birth. Once these are obtained, the transition can be done through a series of steps. These steps are shown with an example as to how the data are manipulated.

Step 1:

The DHS surveys record dates through the use of Century-Month Coding (CMC) (DHS et al., 2013). This is used so the specific months are represented clearly no matter when the surveys are taken. The coding is structured so that January 1900 is given a value of 1 in CMC, February 1900 is given a value of 2, January 1901 is given a value of 13 etc. The CMC equivalent to a month can be easily found through using the equation

$$\text{CMC(Month,Year)} = 12(\text{Year} - 1900) + \text{Month}$$

Using this system to measure dates is useful however the exact times of death and dates of birth (within the month) are not known therefore Hill (2011) add pseudo-random noise to the times. To do this, two variables are taken from the surveys, being the household number and the day of the interview. They are then used to create random noise as they are assumed to have no correlation to age at deaths or dates of birth. The two random variables, random1 and random2, are found through splitting the household numbers and days of interviews into deciles (labelled 0-9) then dividing these by 10 and adding 0.05. This way each both the random variables will be ranged between 0.05 and 0.95.

To add the noise, random1 is added to the dates of birth creating variable 'dob'. random2 is used with the age at death depending on the scale on which age is measured. The ages of death from the DHS surveys are measured in days when the death occurs within 28 days, months when the death is within 1 to 23 months and years when the deaths occur after 23 months. The notation used to distinguish days, months and years is the first value of the three letter code given (1, 2 or 3) whilst the age at death is shown as the last two values given. For example if an age at death of 107 is supplied then the observation has died at the seventh day, whereas if a value of 207 is given then the observation has died within the seventh month. If the age is measured in days, the new age at death (aad) is $(\text{'age at death'} + \text{random2})/31$, if its measured in months aad becomes 'age at death' + random2 and if its measured in years the aad is $(\text{'age at death'} + \text{random2}) * 12$.

This step is only necessary to perform on the DHS data and not the Weibull simulations as the age at death simulations are already specific and the dates of birth are sampled from the DHS pseudo-random dates of birth. It is also performed before any models have been fitted in order to use the same data across both types of model. A slight adaptation is used here to allow for noise to be added to the censored ages when using the survival analysis approach. Here noise from random2 is not only added to the age at death but also the date of the survey and so when the censored ages are calculated (age at survey) they are also including the same type of noise added to the ages at death. If the censored age has then taken a negative value or a value equal to or greater than 60 it is relabelled as 0.05 or 59.95 respectively. It should be noted that day of interview was not recorded during the 1989 Kenya survey and so month of interview was used for these observations.

To show an example, a small sample of the real Kenya female data using all of the surveys is shown with the motive of finding the neonatal mortality rate for the year 1989. This step is shown in Table 4.4 where the two random variables are produced and added to the original data.

Table 4.4: An example of adjusted DHS data from female Kenya data

Alive	Age at death	Date of Birth (CMC)	Date of Survey (CMC)	random1	random2	aad	dob
yes	NA	Dec 1988 (1068)	June 2003 (1243)	0.65	0.95	NA	1068.65
no	209	Feb 1989 (1070)	May 2003 (1242)	0.95	0.85	9.85	1070.95
yes	NA	July 1989 (1075)	May 2003 (1242)	0.85	0.15	NA	1075.85
no	107	Feb 1989 (1070)	April 1989 (1072)	0.15	0.75	0.25	1070.15
yes	NA	Sep 1989 (1077)	May 2003 (1242)	0.85	0.95	NA	1077.85

Step 2:

Here any observations which are born in times which do not have any influence on the mortality rate are removed. Since we are looking for any deaths aged between 0 and 1 month old in the year 1989, any birth before 1st December 1988 and after 31st December 1989 are not relevant to this calculation. Therefore any observations with date of birth CMC less than 1068 or greater than 1081 (31st of December 1989 is assumed to be 1st January 1990) are removed.

Step 3:

The exposure to risk is used as the population of the mortality rate. To find the exposure to risk, each observation is placed into a category (scenario). These categories are labelled a, b, c, d and e and are determined by the amount of time the observation has in the wanted time period at the wanted age. Each scenario can be shown as:

a: The observation is born too early to experience the age range within the time period

b: The observation experiences some of the age range before the required time period

c: The observation experiences the full age range within the time period

d: The observation experiences some of the age range after the required time period

e: The observation is born too late to experience the age range within the time period

Although categories a and e are mentioned, it should be noted that in application if Step 2 is performed correctly there should be no observations in these scenarios that remain in the subset of data.

Using the notation that x_U, x_L, t_U, t_L are the upper and lower bounds of the age range and time period respectively, the exposure to risk can be calculated for each scenario as follows:

Scenario	Exposure for Survivors	Exposure for Deaths
b	$x_U - x_{t_L}$	$x_d - x_{t_L}$
c	$x_U - x_L$	$x_d - x_L$
d	$x_{t_L} + (t_U - t_L) - x_L$	$x_d - x_L$

where x_{t_L} is the age of the observation at the lower bound of the time period and x_d is the age of death.

An example of the exposure calculation is shown in Table 4.5. The notable observations here are the first, second and the fourth rows. In the first observation the exposure is shown as 0.65 as they are shown to be born 65% of the way through December 1988 and so are categorised in scenario b. This means that the first 35% of the first month alive was in 1988 and the remaining 65% was in 1989 and therefore the exposure to risk is 0.65. The second observation

is shown to have an exposure of 1 with the observation still alive, this is because although they did have a death it was aged 9.85 months and therefore they were alive between the 0-1 month age range. For the fourth observation, the age and time of death are within the specified ranges and therefore the death is shown as 1.

Table 4.5: An example of exposure calculated using DHS data from female Kenya data

Alive	aad	dob	Scenario	Exposure to Risk	Deaths
yes	NA	1068.65	b	0.65	0
no	9.85	1070.95	c	1	0
yes	NA	1075.85	c	1	0
no	0.25	1070.15	c	0.25	1
yes	NA	1077.85	c	1	0

It is at this step in which survey weights can be applied to the exposure to risk and deaths. However as this is not used in the survival analysis framework it has not been included here.

Step 4:

Next the mortality rates can be calculated using the summed exposure to risk and number of deaths. This is shown as

$$M(x,t) = \frac{\sum_{i=1}^{N} D(i,x,t)}{\sum_{i=1}^{N} E(i,x,t)} * \text{length}(x)$$

where $M(x,t)$ is the mortality rate for age range x and time period t while $D(i,x,t)$ and $E(i,x,t)$ are the deaths and exposure to risk of the i^{th} observation for age range x and time period t. A difference to the calculation here which is not mentioned by Hill (2011) is the multiplication of the length of the age range. This is added to allow for proper scaling between the

exposure and deaths. For example when using the age range 1-12, a survivors exposure to risk in scenario c would be $x_U - x_L = 11$ whereas any death would still be denoted as 1. Multiplying the mortality rate (or dividing the exposure) by the length of the age range, in this case 11, scales the exposure to be 1 at its maximum and can then be used in conjunction with the deaths.

In the hypothetical example of the five observations the mortality rate is found as

$$M(x, 1989) = \frac{\sum_{i=1}^{5} D(i, x, 1989)}{\sum_{i=1}^{5} E(i, x, 1989)} * \text{length}(x) = \frac{1}{3.9} * 1 = 0.2564$$

It should be noted that although this transformation from the survival approach to mortality rate is used, it is not necessary as the forecasted survival curves show how the mortality patterns are changing over time. The transformation to mortality rate is used here in order to compare the survival approach against models that provide mortality rate forecasts.

4.4.2 Lee-Carter Model

To evaluate the performance of the survival approach, it is compared against a commonly used model in mortality forecasting named the Lee-Carter model (Lee and Carter, 1992). The well documented Lee-Carter model is fitted as

$$log(\mu_{at}) = \alpha_a + \beta_a \gamma_t + \epsilon_{at} \tag{4.6}$$

in which μ_{at} represents the mortality rate for age a and time t in which the mortality rates

are calculated through the procedure shown in 4.4.1. An overall age intercept is α_a whilst multiplicative age and time effects are β_a and γ_t respectively. So a key difference between the survival and Lee-Carter approaches is that the survival models use individual level data whilst the data for the Lee-Carter models are aggregated.

In the traditional fitting of the model $\hat{\alpha}_a$ is calculated through getting the age specific means of the log mortality rates matrix. To calculate the multiplicative terms Singular Value Decomposition is used on the centred matrix $log(\mu_{at}) - \hat{\alpha}_a$ with the constraints of $\sum \beta_a = 1$ and $\sum \gamma_t = 0$ to ensure the solution is unique. In the traditional model the independent errors ϵ_{at} are assumed to be distributed normally with a mean of zero and variance σ^2.

To then forecast from the model it is assumed that β_a is constant over time and only γ_t has to be forecasted. Lee and Carter suggest that a random walk with drift is the most appropriate model for their data. Although they also state that other ARIMA models may be more appropriate for other datasets, the random walk with drift is almost exclusively used in applications (Girosi and King, 2008). The random walk with drift is shown as

$$\begin{aligned} \hat{\gamma}_t &= \hat{\gamma}_{t-1} + \hat{d} + \epsilon_t \\ \epsilon_t &\sim N(0, \sigma^2_{rw}) \end{aligned}$$

in which $\hat{\gamma}_t$ are the estimates of the time components γ_t, $\hat{d}$ is the drift parameter, defined as

$$\hat{d} = \frac{\hat{\gamma}_T - \hat{\gamma}_{t_0}}{T - 1}$$

and σ^2_{rw} is the standard deviation defined as

$$\hat{\sigma}^2_{rw} = \frac{1}{T-1}\sum_{t=1}^{T-1}(\hat{\gamma}_{t+1} - \hat{\gamma}_t - \hat{d})^2 \tag{4.7}$$

This model was fitted as one of the competing models used to compare against the survival approach. Four other variants of the Lee-Carter model were also fitted, using two different R packages.

Firstly a Lee-Carter model was fitted using the *ilc* package (Butt et al., 2019) using Gaussian errors. The *ilc* package uses an iterative Newton-Raphson method to estimate the model parameters. This leads to finding similar parameter estimates as when using the standard Lee-Carter model. A Lee-Carter model was also fitted with the *ilc* package using Poisson errors, adjusting equation 4.6 to

$$\log(y_{at}) = \log(n_{at}) + \alpha_a + \beta_a\gamma_t + \epsilon_{at} \tag{4.8}$$

in which y_{at} and n_{at} are the number of deaths and total exposure for each age and time respectively. For this reason the model with Poisson errors is shown to give different parameter estimates to both of the previous Lee-Carter variations and it uses the extra information of exposure to distinguish any fluctuations in frequency of observations. The use of the *Demography* package (Hyndman, 2012) allows for another representation of the Lee-Carter model. The difference between this and the original fitting of the model is that here a two step approach is taken to estimate the time trend γ_t. In this case, after β_a and γ_t have been estimated in the SVD step, γ_t is refitted to the number of deaths using the already found parameters $\hat{\alpha}_a$ and $\hat{\beta}_a$. The use of a second stage fitting is recommended originally by Lee and Carter however some researchers skip this phase (Girosi and King, 2008).

Two other adaptations of these models are also used for comparison. As the survey form of the data provides fewer numbers of observations as time increases, the mortality rates in the most recent years become more unstable. When using the standard form of the Lee-Carter model this may be an issue as the forecasts are a random walk with drift starting from the estimate given to the last time component (γ_T). As this problem is down to the frequency of observations (and therefore total exposure), the main model this affects is the model using Gaussian errors with no second stage fitting. To try to combat this the two other models are fitted in the exact same way as the Gaussian error *ilc* model and the *Demography* model using a second stage fitting process. However rather than using the entire set of mortality rates, the models are fitted to a reduced set in which the most recent three years have been removed. Forecasts then start from the third last 'in-sample' year and continue to match the full fitted models. It should also be noted that as the second stage fitting refits $\hat{\gamma}_t$ to the number of deaths, reducing this model is not entirely necessary however in the case of it producing competing forecasts it is included for comparison. All five Lee-Carter variations are modelled with 10000 simulations to provide forecasts.

After adjustments are made to the data, Mejía-Guevara et al. (2019) use a variation of the Lee-Carter model first shown by Li, Lee and Tuljapurkar (2004). Here we show the forecasts of the traditional styled Lee-Carter model as the Li-Lee-Tuljapurkar variant used by Mejía-Guevara et al. (2019) is adjusted for when the data experiences missing time points. When no years are missing, as in our settings, the Li-Lee-Tuljapurkar variant is designed to revert to the original Lee-Carter model.

4.4.3 Results

4.4.3.1 Kenya

In the application the DHS data is used from two countries, Kenya and Uganda. The country specific results are shown first whilst a overall comparison between the approaches are shown in Section 4.4.3.3. This Section will focus on the results from the Kenya data whilst Section 4.4.3.2 looks at the output from the Uganda surveys.

To evaluate how the models perform a cross validation setting is applied. Here all of the information used for running the models has been acquired from all of the surveys apart from the most recent which is withheld to compare forecasts. The Kenya standard DHS surveys that are currently available are from 1989, 1993, 1998, 2003, 2008/2009 and 2014.[1] Therefore the data from the 2014 Kenya survey are only used to review the forecasting performance of the models whilst the rest are used to fit the models. For a comparison as to how losing the last survey affects the frequency of observations, Table 4.6 shows the number of observations alongside median ages of death. Comparing this with the full female Kenya data (shown in Table 4.2) the information lost can be shown clearly when not using the 2014 survey. As each of the surveys here hold approximately the same number of observations, given that the later surveys can also add information to earlier birth cohorts, this creates an overall decreasing trend in the number of observations per birth cohort as time increases. This decreasing trend combined with the latest cohorts being observed for less time creates more uncertainty around the most recent birth cohorts and subsequently is more beneficial for the survival models or the Lee-Carter model with Poisson errors as it includes a population (exposure) term.

[1] Data accessed 13th December 2018

Table 4.6: The frequency of deaths and total observations for female Kenya data when not including the 2014 survey

Birth Cohort	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995
Deaths	197	190	208	232	185	164	193	163	183	131	142	113
Median Age at Death	5.95	6.35	6.85	6.25	5.95	7.35	6.15	4.95	6.75	6.25	8.15	8.05
Total Observations	2367	2144	2490	2533	2463	2061	1963	1761	2032	1489	1403	1439
Birth Cohort	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Deaths	164	131	107	96	112	88	85	34	38	32	34	38
Median Age at Death	7.05	3.65	5.45	6.35	4.40	5.80	3.95	0.42	3.85	4.80	1.65	1.00
Total Observations	1656	1537	1224	1047	1187	1028	1287	791	569	587	607	572
Birth Cohort	2008	2009										
Deaths	19	3										
Median Age at Death	1.31	0.03										
Total Observations	600	32										

The three plots shown in Figure 4.8 show the forecasts of the two Weibull models as well as the best fitting variant of the Lee-Carter model for female observations in Kenya over the three age groups. In this case the overall best fitting model is shown to be the Lee-Carter model using Poisson errors. An effect of losing a survey to create a cross validation period found in the 1-12 months age group. For this age group the estimated mortality rate from the three models seem to be higher than the true rate during the in-sample fit. This is due to the 2014 DHS survey also adding information to earlier birth cohorts than 2009 and therefore slightly changing the earlier mortality rates whilst the model fits are estimated solely from information gained up to the 2008/2009 survey. The mortality patterns in the three age groups are similar to that of the males shown in Figure 4.9 where the neonatal age group (0-1 month) show a small downwards trend over time whilst the decline in mortality for the other two groups are far stronger. The strong declines in mortality look to start roughly at the year 2000, potentially showing the improvements gained from the MDGs (UN, 2000). This drop in mortality rate is captured by the Weibull spline model whilst although the less complex linear trend model shows a decline for the later ages, the earlier years restrict the model from having as accurate forecasts.

Figure 4.8: Mortality rate forecasts for Female Kenya DHS data withholding the 2014 survey for cross validation

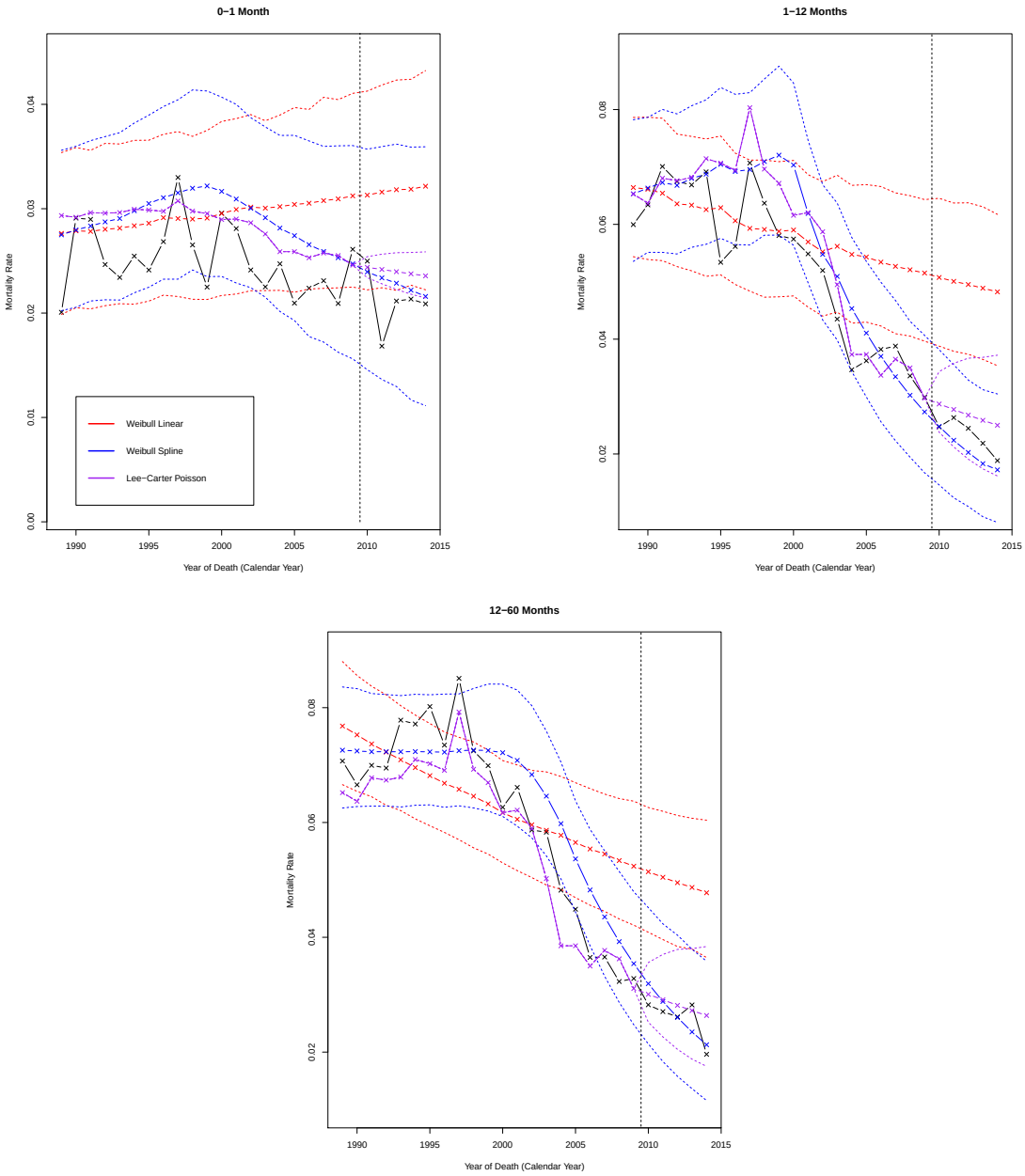
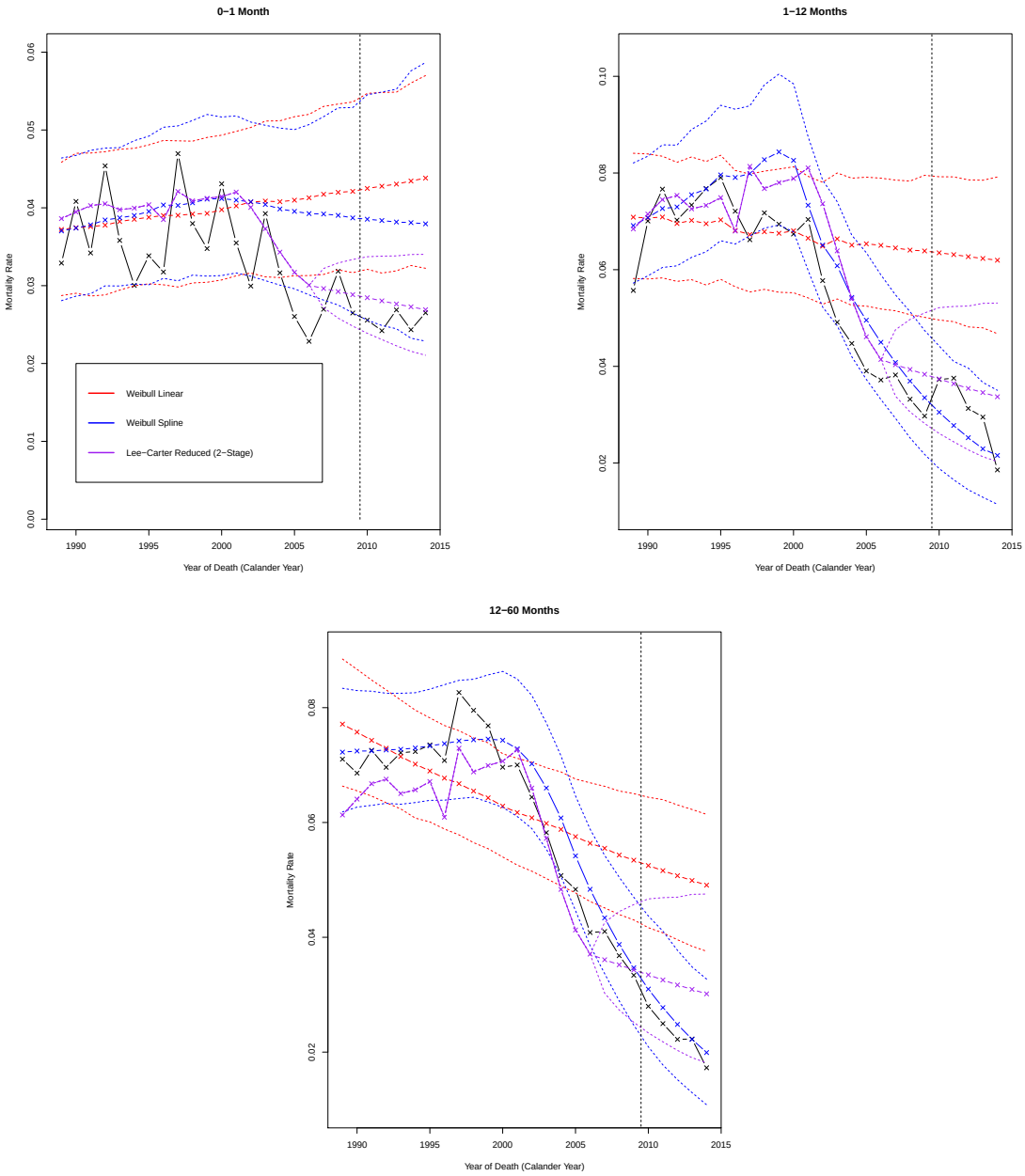


Figure 4.9: Mortality rate forecasts for Male Kenya DHS data withholding the 2014 survey for cross validation



Numerical results from the Kenya application are shown in Table 4.7. Here the mean forecast bias is used to measure the ability of the point estimate forecasts from the two Weibull models and the Lee-Carter variants. Here we define the mean absolute forecast bias as mean(|Forecast Error|) in which the forecast error can be found by finding the difference between the (posterior mean) forecasts and the observed mortality rate. A better forecast provides a lower bias, so the lower the mean forecast bias the more accurate forecasts are attached with that model.

Table 4.7: Mean bias for Kenya over the 3 age groups

		0-1	1-12	12-60	Total Mean Error
Male	Weibull Linear	0.018	0.032	0.028	0.026
	Weibull Spline	0.013	0.006	0.002	0.007
	Lee-Carter	0.030	0.017	0.003	0.017
	Lee-Carter (iterative Poisson)	0.009	0.011	**0.002**	0.007
	Lee-Carter (iterative Gaussian)	0.030	0.017	0.003	0.017
	Lee-Carter Reduced (iterative Gaussian)	0.004	0.009	0.013	0.008
	Lee-Carter (2-stage)	0.027	0.013	0.003	0.014
	Lee-Carter Reduced (2-stage)	**0.002**	**0.005**	0.009	**0.005**
Female	Weibull Linear	0.011	0.027	0.024	0.02
	Weibull Spline	0.002	0.003	0.002	**0.002**
	Lee-Carter	0.009	**0.002**	0.002	0.004
	Lee-Carter (iterative Poisson)	0.003	0.004	0.003	0.003
	Lee-Carter (iterative Gaussian)	0.009	0.002	**0.002**	0.004
	Lee-Carter Reduced (iterative Gaussian)	0.006	0.021	0.010	0.012
	Lee-Carter (2-stage)	0.009	0.003	0.004	0.005
	Lee-Carter Reduced (2-stage)	**0.002**	0.012	0.012	0.009

When comparing the Weibull spline model against the Lee-Carter forecasts for females we find that within each age group a different version of the Lee-Carter model performs best. Due to the mortality rates being on similar scales across the age groups, the total mean forecast error can be used without the value being dominated by a particular age group. Using this overall mean forecast error, despite the spline model being outperformed at the age group level, the consistent forecasts produced are shown as the best.

The males in Kenya are shown to have similar mortality patterns as females across the three age groups. A slight difference is found in the neonatal age group where the males show a larger decline over time when all of the data is used, although the data for cross validation shows the females to have more of a reduction. In the oldest two age groups, the spline model is shown to forecast just as well as the best of the Lee-Carter variants. The neonatal age group is where it is shown to lose performance as (when using the in-sample data) the full extent of the drop in mortality is not captured as well, and so it over predicts the true mortality rate. Despite this, when comparing the overall performance of each model there is only one model with more accurate forecasts which is the reduced Lee-Carter model with a second stage fitting on the time component. Because of this the Weibull spline model can still be shown as one of the best forecasting models out of the selection.

For a more concise summary of the Kenya forecasts, the pooled male and female outputs are shown in Table 4.8. The mean forecast bias is shown here alongside the coverage of the models. With the forecast bias being a measure of the point estimates, it is also important to show that the models provide an appropriate amount of uncertainty with forecasts and this is tested using the coverage probability. The 95% coverage is defined as if an observation lies within

the 95% credible interval of the forecast then that entry is assigned a 1. If the observation lies outside the interval then it is assigned a 0. The frequency of how many 1's is then found and divided by the total number of forecasts. A perfect 95% coverage would see the value match its nominal level and will be 0.95. The aggregated table shows that compared to all of the Lee-Carter models the Weibull spline model may not be the best performing model at the age group level, however the consistent forecasts allow it to be the best performing model overall. The spline model is also shown to give wider uncertainty bounds than the Lee-Carter models however the coverage statistics show this may be necessary as the spline model has the closest coverage to 0.95.

Table 4.8: Mean forecast bias and 95% coverage for Kenya data combined for males and females

	Mean Forecast Bias	95% Coverage
Weibull Linear	0.0231	0.033
Weibull Spline	**0.0048**	**0.933**
Lee-Carter	0.0105	0.533
Lee-Carter (iterative Poisson)	0.0051	0.6
Lee-Carter (iterative Gaussian)	0.0104	0.533
Lee-Carter Reduced (iterative Gaussian)	0.0104	0.733
Lee-Carter (2-stage)	0.0097	0.533
Lee-Carter Reduced (2-stage)	0.0070	0.9

4.4.3.2 Uganda

Similarly as for Kenya, a cross validation study is performed using Ugandan surveys to test the survival approach against the more traditional approach. The available Uganda standard DHS surveys available are 1989/1990, 1995, 2000-2001, 2006, 2011, and 2016.[2] To compare the forecasts of each model the 2016 survey is withheld during the model fits and reintroduced to evaluate model performance.

Figures 4.10 and 4.11 show the male and female forecasts over the three age groups respectively. The models shown are both the survival models (Weibull linear and spline) as well as the best fitting Lee-Carter variant according to its point estimate accuracy. As found with Kenya, the more concerning plots shown here are regarding the neonatal age group. Although the models are only fitted using surveys up to 2011, there is shown to be only a small downwards trend in mortality rates when compared to the other age groups. Although the downwards trend across all three groups is positive, the focus on neonatal mortality in the SDG means we would much rather see this group start to match the progress made by the other two.

[2]Data accessed 13th December 2018

Figure 4.10: Mortality rate forecasts for Female Uganda DHS data withholding the 2014 survey for cross validation

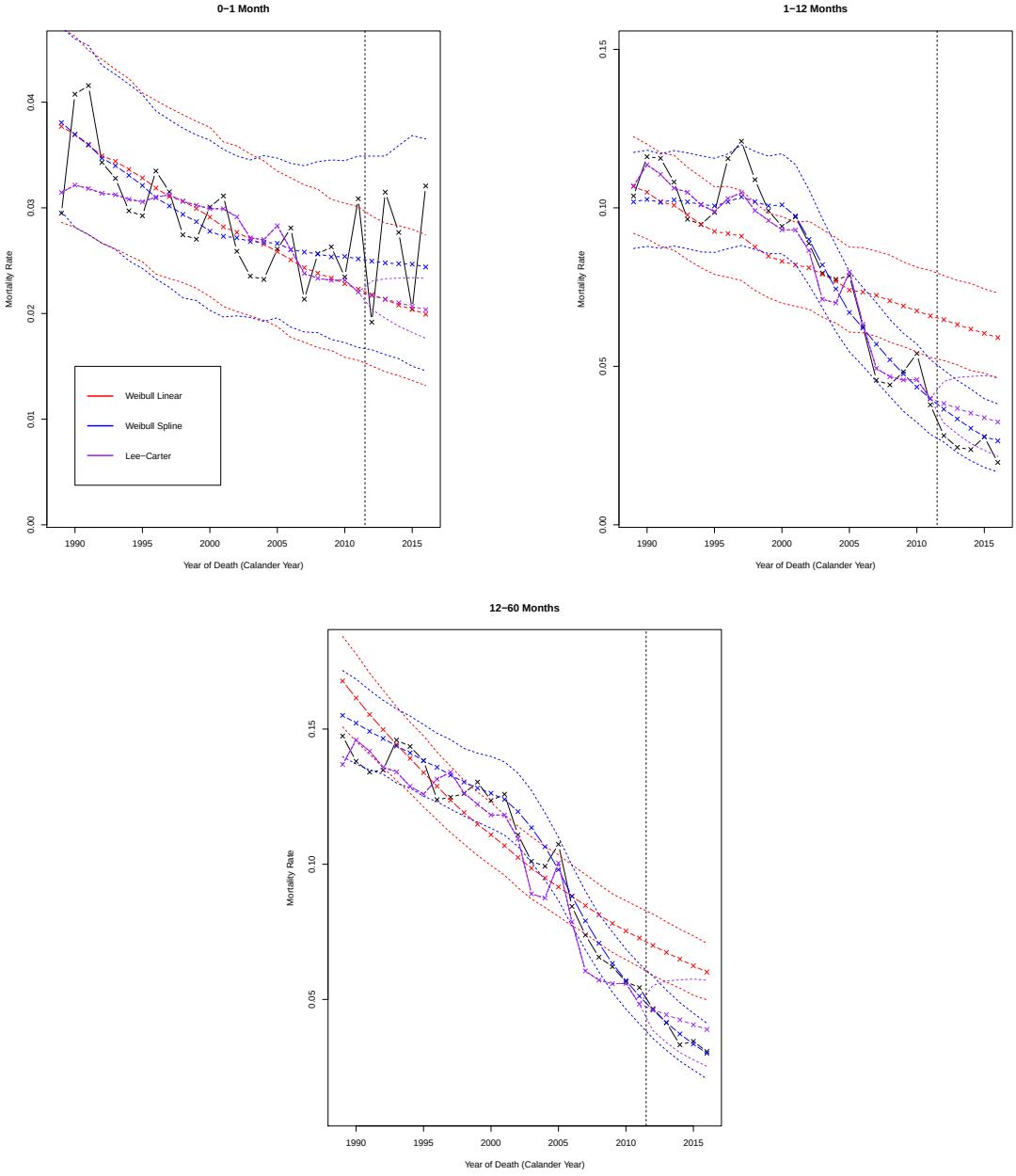


Figure 4.11: Mortality rate forecasts for Male Uganda DHS data withholding the 2014 survey for cross validation

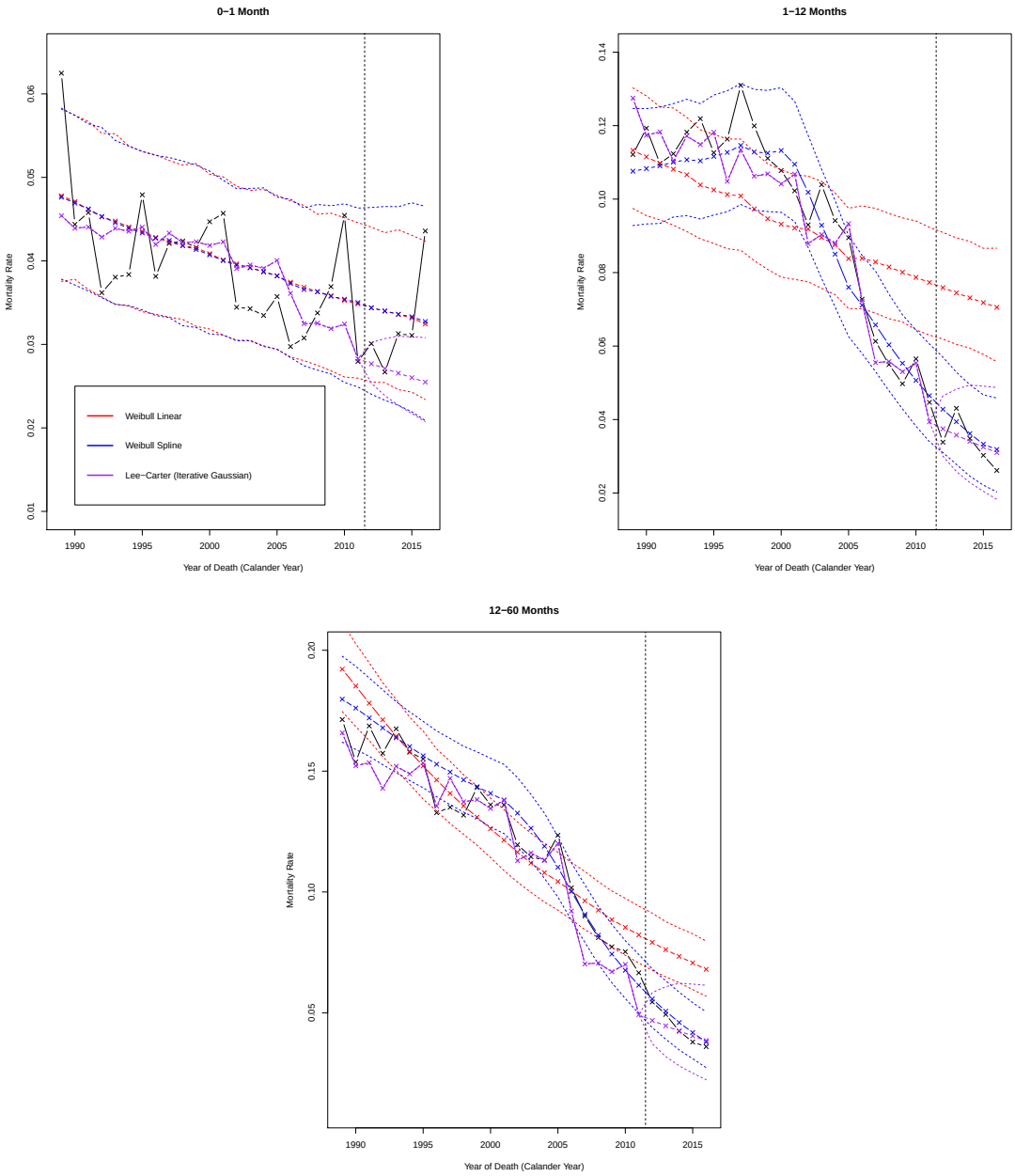


Table 4.9 shows the mean forecast error across the three age groups for both males and females. For female forecasts the Weibull spline model is shown to be the most accurately forecasting model in each age group. The model here is shown to capture the mortality rate decline in the last two groups exceptionally resulting in the forecasts excelling. This results with it being the overall best fitting model, but with some of Lee-Carter models also producing competitive forecasts. For males the Weibull spline model is the most accurately forecasting model in the last age group. With this added performance in the child mortality age group as well as the models showing a similar forecast performance in the first two age groups, the overall mean forecast error shows the spline model to be the most accurate.

Table 4.9: Mean bias for Uganda over the 3 age groups

		0-1	1-12	12-60	Total Mean Error
Male	Weibull Linear	0.006	0.033	0.022	0.020
	Weibull Spline	0.006	0.004	**0.004**	**0.005**
	Lee-Carter	0.004	0.004	0.007	0.005
	Lee-Carter (iterative Poisson)	**0.004**	**0.003**	0.008	0.005
	Lee-Carter (iterative Gaussian)	0.004	0.004	0.007	0.005
	Lee-Carter Reduced (iterative Gaussian)	0.005	0.008	0.009	0.007
	Lee-Carter (2-stage)	0.004	0.004	0.007	0.005
	Lee-Carter Reduced (2-stage)	0.005	0.006	0.008	0.007
Female	Weibull Linear	0.006	0.031	0.024	0.020
	Weibull Spline	**0.005**	**0.007**	**0.002**	**0.005**
	Lee-Carter	0.006	0.009	0.004	0.006
	Lee-Carter (iterative Poisson)	0.007	0.009	0.004	0.007
	Lee-Carter (iterative Gaussian)	0.006	0.009	0.004	0.006
	Lee-Carter Reduced (iterative Gaussian)	0.008	0.011	0.008	0.009
	Lee-Carter (2-stage)	0.006	0.009	0.004	0.006
	Lee-Carter Reduced (2-stage)	0.009	0.009	0.006	0.008

The overall model performance for Uganda across both males and females is shown in Table 4.10. Here (as with Kenya) the mean forecast error is used to test the point estimate fit of forecasts while the 95% coverage is used to evaluate the uncertainty produced. As expected from the gender specific results, the Weibull spline model is shown to be the most accurately forecasting model. The main differences between forecasts are found in the coverage. The Lee-Carter models all show lower coverage rates than the optimal 0.95 level. This could suggest that the uncertainty ranges associated with the forecasts are too small. On the other hand, the Weibull spline model has each observed forecast within its 95% credible intervals. This could suggest that the spline model provides too much uncertainty however, it is clear that of the figures that the uncertainty level around the spline model can be justified as appropriate.

Table 4.10: Mean forecast bias and 95% coverage for Uganda data combined for males and females

	Mean Forecast Error	95% Coverage
Weibull Linear	0.0202	0.6
Weibull Spline	**0.0046**	**1**
Lee-Carter	0.0055	0.633
Lee-Carter (iterative Poisson)	0.0060	0.667
Lee-Carter (iterative Gaussian)	0.0055	0.633
Lee-Carter Reduced (iterative Gaussian)	0.0082	0.833
Lee-Carter (2-stage)	0.0057	0.633
Lee-Carter Reduced (2-stage)	0.0073	0.866

4.4.3.3 Overall Comparisons

With the application of the survival approach showing positive results for both Kenya and Uganda, an overall one-number summary of model performance is shown in Table 4.11 to compare the models after pooling the male and female results for both countries. This is useful for a quick overview as to how the models compete, however the more detailed summaries shown in Sections 4.4.3.1 and 4.4.3.2 are more reliable when comparing models. Due to its success in the country specific analysis, it is no surprise to see that the Weibull spline model, fit through using the novel survival approach, has the overall lowest mean forecast error and the best coverage out of the group. It is also notable that the more traditional fitted Lee-Carter models, being the standard model as well as the iterative versions (with gaussian errors) show poorer fitting forecasts. The more complex models, fitting the parameter estimates to deaths rather than the rates, allow for instances when the frequency of observations are low, like the 2009 cohort for Kenya data (see Table 4.6), not to have as much of an impact in the fitting of the model. What is very noticeable from the results is that the Weibull linear model has poor forecasts in comparison to the other models, showing that a thorough exploratory analysis of data is still necessary regardless of the approach used.

Table 4.11: Overall mean forecast bias and 95% coverage combining the Kenya and Uganda results

	Mean Forecast Bias	95% Coverage
Weibull Linear	0.0217	0.317
Weibull Spline	**0.0047**	**0.967**
Lee-Carter	0.0080	0.583
Lee-Carter (iterative Poisson)	0.0055	0.633
Lee-Carter (iterative Gaussian)	0.0080	0.583
Lee-Carter Reduced (iterative Gaussian)	0.0093	0.783
Lee-Carter (2-stage)	0.0077	0.583
Lee-Carter Reduced (2-stage)	0.0071	0.883

The goal of this study was to show that the survival approach can used as a viable option in this setting. Given the results from both the Kenya and Uganda comparisons we find that using the survival setting we can not only provide competing forecasts but we have the ability to provide forecasts with more consistent point estimates and more appropriate uncertainty intervals.

4.4.4 2030 Forecasts

In 2015 the United Nations released a set of sustainable development goals (UN, 2015) to attempt to further reduce child mortality after the success of the millennium development goals (UN, 2000). The goals set were to reduce the neonatal mortality to a maximum of 12 deaths for every 1000 live births and reduce under-five mortality to a maximum of 25 deaths every 1000 live births by 2030. With the survival analysis approach to forecasting DHS data being shown to be a suitable alternative to the more traditional methods, the Weibull spline model was used in conjunction with the full amount of data available to compare 2030 forecasts with the goals set.

Similarly to Section 4.4.1, when the age at death simulations are made for the forecasting years, the dates of births are sampled from the pooled group dates from the last five observed cohorts. After the transformation the neonatal mortality rates are found as being any deaths within the first month (identical to the first age group) whilst the under-five mortality rates account for any deaths within the first five years of life (pooling the three age groups). Figures 4.12-4.15 show the model fits and forecasts in full over Kenya and Uganda for males and females whilst Table 4.12 shows the exact forecasts with 95% credible intervals.

Our forecasts show that both males and females in Kenya are on track to reaching the SDGs relating to under-five mortality. Both are also showing strong declines in neonatal mortality but further reduction is needed to meet the specific goal. An interesting development in Kenya is that with the inclusion of the 2014 survey a stronger decline in neonatal mortality can be found in males. This strong decline shows that when forecasting these rates as far ahead as 2030, males are projected to overtake females and result in a lower mortality rate.

Both Ugandan males and females are not predicted to reach the SDG target using this data. For both males and females under-five mortality shows strong reductions over time, showing positive steps for the future, and although they are not predicted to reach the SDG target, the under-five goal is within the 95% credible intervals of both. With the SDG focus on the neonatal mortality rate we find that males show a small decline over time. More worryingly, when modelling the females the spline model picks up a small upwards trend in neonatal mortality using this data. This shows further work may be needed to try and reduce neonatal mortality.

It should be noted that this study is not reflective of the entire populations of Kenya and Uganda as sampling weights have not been applied and the data used is the raw DHS survey data.

Table 4.12: Weibull spline model forecasts for neonatal and under-five deaths (per 1000) in 2030 compared with the sustainable development goal

	NMR per 1000 (95% CI)	U5MR per 1000 (95% CI)
SDG-3 (Goal)	12	25
Kenya Female	14.6 (7.5 - 23.9)	23.1 (10.5 - 40.1)
Kenya Male	12.5 (6.5 - 20.2)	20.1 (9.2 - 34.7)
Uganda Female	28.4 (17.2 - 42.6)	39 (21.8 - 61.2)
Uganda Male	30.3 (18.8 - 44.4)	44.5 (26 - 67.8)

Figure 4.12: Full neonatal and under-five mortality rate forecasts up to 2030 for females in Kenya

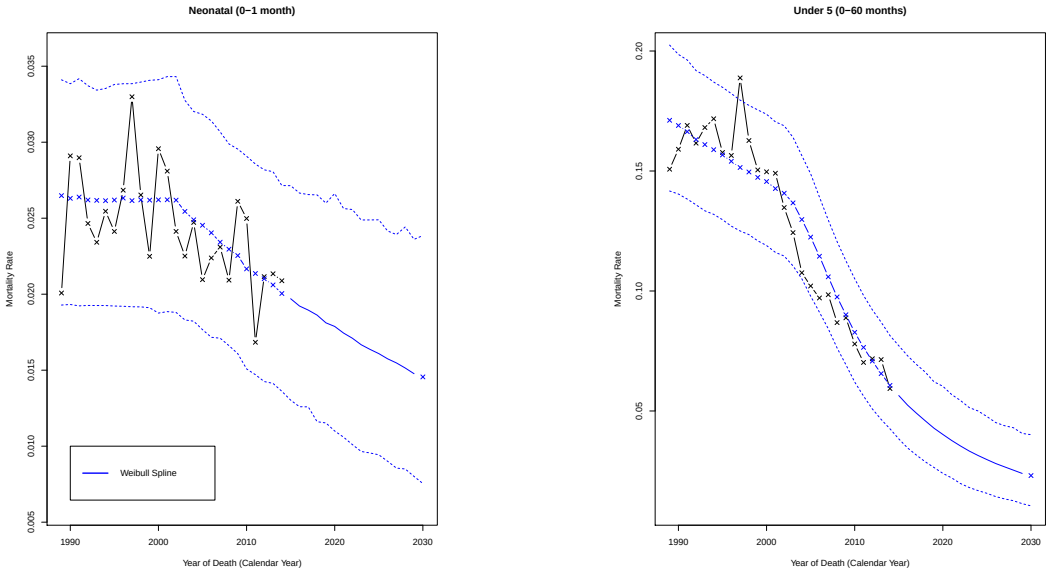


Figure 4.13: Full neonatal and under-five mortality rate forecasts up to 2030 for males in Kenya

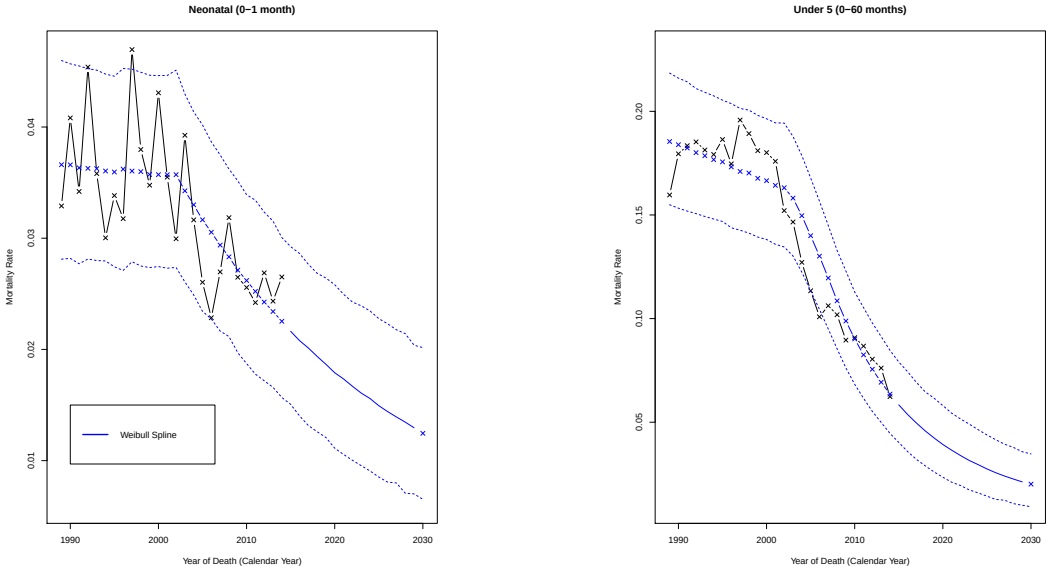


Figure 4.14: Full neonatal and under-five mortality rate forecasts up to 2030 for females in Uganda

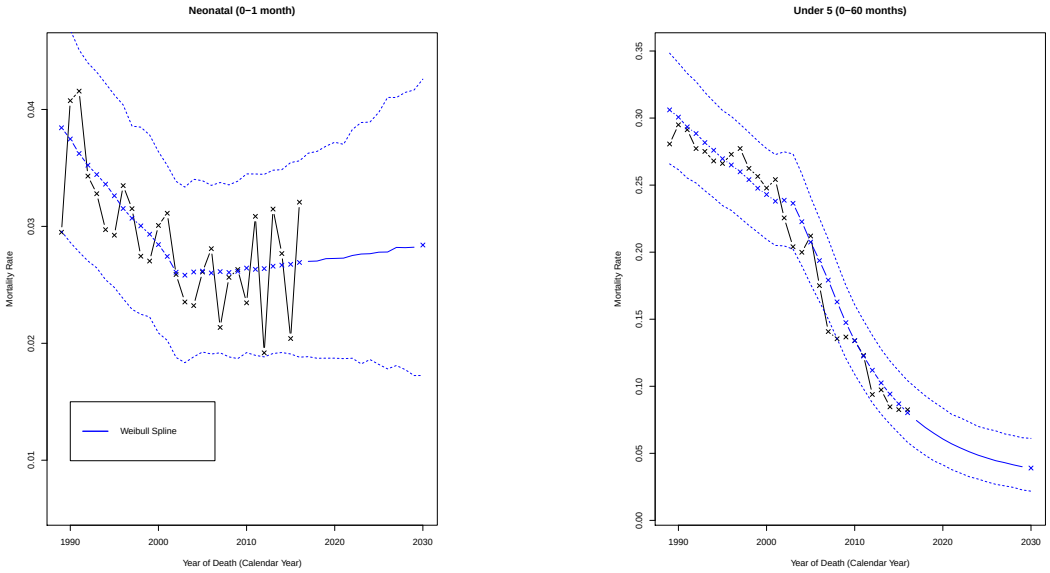
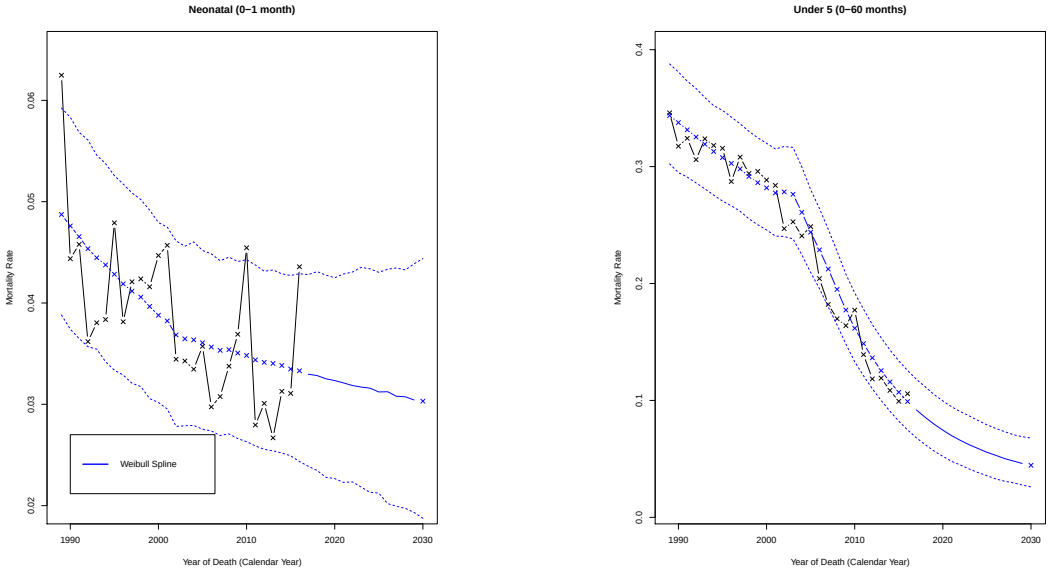


Figure 4.15: Full neonatal and under-five mortality rate forecasts up to 2030 for males in Uganda



4.5 Discussion

In this study we have shown a novel approach to forecasting using the individual level DHS data. This survival approach has been compared with a more traditional mortality setting over two countries with both male and female observations separated. The outcome from the cross-validation setting showed the survival approach is capable of providing forecasts that not only match, but outperform that of a traditional approach. The best performing survival model was then used to forecast to 2030 allowing us to show if the countries are on the track to achieving the SDG target (UN, 2015). Even without adjusting the data, our projections for Kenya show similar findings to that of Mejía-Guevara et al. (2019) in which they are on track to reach the under-five mortality target but not the neonatal target. With the SDG goals specifying neonatal mortality we hope to see a similar effect as the MDGs had on under-five child mortality.

This new framework for mortality forecasting provides many paths for future work. A clear improvement to the method is to take into account various data adjustments. This has been done through using the DHS data by using adjustments to account for survivor bias (Wakefield et al., 2019) or adjusting the data to match the UN IGME estimates (Mejía-Guevara et al., 2019). However whether similar methods are possible when using the survival approach has not been looked into. Another improvement to the data used can be the incorporation of the sample weights which accompany each observation in the dataset. These sample weights are added with the goal to use with the data to account for the probability of each participant being selected (Croft et al., 2018). The inclusion of the sample weights would be a necessary

improvement, especially when attempting to analyse the mortality patterns of entire countries. When they have not been applied, as in this instance, we must be mindful that the survey population is not fully representative of the entire country population.

Another extension to the study is to include more model covariates. More covariates can be included to capture any unnoticed trends and find any lifestyle changes that may positively or negatively affect child mortality and subsequently improve model performance. The usage of the individual level survey data allows for more detailed information on the lifestyle of the children and so there are a number of covariates which we could consider that would usually be unavailable at the area level. Firstly Way (2015) found that children born in rural areas are 1.7 times more likely to die before their fifth birthday than those born in urban areas, showing a need for a rural/urban indicator may be useful. Another useful indicator can be a measure of the mothers education as Way (2015) also found that mothers with secondary or higher education are almost three times as likely to have their children survive than mothers without any education. It has also been shown that smoking histories and alcohol consumption can affect mortality (Wang and Preston, 2009; Jarl et al., 2008) therefore it may be possible that the tobacco/alcohol consumption of the mother can change the mortality pattern of the child. However, tobacco and alcohol consumption of the mother is currently not available through the DHS and so when using this data source we have no means of tracking it. To account for these types of covariates that cannot be tracked, a frailty model could be considered. A frailty model is a random effects model that can be used to describe the influence of unobserved covariates (Hougaard, 1995) and so could be used (in conjunction) with other known covariates to have a more descriptive model.

Another area of future work can be an application of the models to other countries. So far the countries that have been used are Kenya and Uganda. Under-five and neonatal mortality rates could be forecasted using more countries to show how Sub-Saharan Africa has responded to the MDG target and is responding to the SDG targets. As mentioned in Section 4.3, if we are to continue using this exact approach each individual country will need to be analysed separately showing that there no need for a trend model to be placed on the shape parameter. If it is found that a trend model is required for both shape and scale the same methodology can be followed however the model fitting will need to be done using a different software that allows for more control in the model specification, such as WinBUGS.

In line with the study shown in Chapter 2, a model averaging approach could be applied to the survival models. The results in the cross validation show that there is a need for a model averaging technique as different models are shown to perform best amongst different countries, genders and age groups. This leads to a more general question as to whether there is a way the Lee-Carter models (using aggregated data) can be used in a model averaging approach with the survival models (using individual level data). What the results also show is that the linear model forecasts poorly in comparison to the other models, showing that (as always) the models chosen to be used in a model averaging approach have to be carefully selected. An option for other survival models when using a model averaging approach can be a number of spline models in which the knots are positioned in different places. This will then remove some uncertainty about the true position of the knot, with a combination of the best fitting models being used. In the next chapter we show a progression to this extension of the work being how model averaging weights can be found in this setting.

Chapter 5

Model Selection in a survival setting

After the results from the application in Section 4.4 one of the many possible extensions is the introduction of a model averaging framework for the survival models. In this chapter this extension is explored through the means of showing a weighting technique that can be applied to the survival setting. Along with the description of the weighting technique, a simulation setting is shown in which weights assigned to each model can be numerically assessed.

5.1 Introduction

With a novel method for using a survival technique for forecasting under-five mortality shown in Chapter 4 a clear extension in line with the overall project is to look into a model averaging approach in this context. Throughout the application in Chapter 4 the two survival models used take the form of linear and piecewise linear spline trends placed on the scale of a Weibull distribution. The findings show that in post neonatal ([1-12) months) and child ([12-60)

months) mortality levels the added flexibility of the spline is a necessity to fully capture the more severe mortality drops over time. In the neonatal age range ([0-1) month) a more subtle mortality drop over time is found (particularly in Uganda) allowing for the linear model to show some more competitive forecasts and showing it as a viable option to include in a model averaging approach.

As shown in Chapter 2 one of the important steps in a model averaging approach is the formulation of the weights assigned to each model. The model selection method shown can be found as an adaptation of a finite mixture model (McLachlan and Basford, 1988). Here the application of a mixture of Weibull models is shown, with the adjustments needed to transition the mixture to obtain model weights via a model selection approach.

In the survival setting steps have been taken to implement a mixture model with the usage of parametric survival models (Erişoğlu et al., 2011; Mohammed et al., 2014, 2015) in which model parameters are evaluated through means of the Expectation Maximisation (EM) algorithm. Although the EM algorithm can be extended into the Bayesian framework, Marín et al. (2005) show an alternative Bayesian method for parameter estimation through a Gibbs sampling approach. Here the Gibbs sampling framework is adjusted and used to show how it can appropriately estimate model weights for a model averaging approach, rather than a traditional mixture model approach.

5.2 Methodology

Firstly, the form of the models can be shown in a more general form as

$$y_i \sim Weibull(r, \mu(C_i))$$

where y_i are the observed times of events, $\mu(C_i)$ is the scale parameter being modelled by a function across cohorts C_i and r representing the shape parameter which for all models considered is constant across birth cohorts. This general form of the models encompasses the models shown in Section 4.2.4, for example the linear Weibull model gives the structure $log(\mu_i(C_i)) = \alpha + \beta * C_i$ (here $\mu_i(C_i) \equiv \mu(C_i; \alpha, \beta)$) and for the spline model $\mu_i(C_i) \equiv \mu(C_i; \boldsymbol{\alpha}, \boldsymbol{\beta})$. For the remainder of the chapter the dependence of μ on the model parameters are suppressed for simplicity and so notation of $\mu_i \equiv \mu(C_i; \cdot)$ is simplified to μ_i. This form of model can also represent one following an Exponential distribution with the constraint that r is fixed at 1. Under the Weibull distribution the hazard and survival functions are

$$h(y|r, \mu) = r\mu y^{r-1} \qquad S(y|r, \mu) = e^{-\mu y^r}$$

To calculate the likelihood of a single Weibull model in which no observations are censored, each observation contributes to the full density and therefore the likelihood will be

$$L(\boldsymbol{\Phi}|\{y_1, ..., y_N\}) = \prod_{i=1}^{N} f(y_i|\boldsymbol{\Phi}), \qquad f(y_i|\boldsymbol{\Phi}) = h(y_i|\boldsymbol{\Phi})S(y_i|\boldsymbol{\Phi})$$

where $\mathbf{\Phi}$ is the model parameter vector and N represents the total number of observations. With right censored observations not experiencing the event yet, the information a censored observation brings only contributes to the probability that y_i exceeds the event. This is also the definition of the survival function $S(y_i|\mathbf{\Phi})$. To show whether an event time is censored, an indicator variable δ is introduced where

$$\delta_i = \begin{cases} 1 & \text{if the age of death is observed (uncensored)} \\ 0 & \text{if the observation is censored} \end{cases}$$

Therefore the likelihood of a survival model experiencing right censored data can be shown as

$$L(\mathbf{\Phi}|\{(y_1, \delta_1), ..., (y_N, \delta_N)\}) = \prod_{i=1}^{N} h(y_i|\mathbf{\Phi})^{\delta_i} S(y_i|\mathbf{\Phi})$$

Given this definition, the likelihood of the general form of the forecasting Weibull model can be shown as

$$L(r, \mu_i|\text{data}) = \prod_{i=1}^{N} h(y_i|r, \mu_i)^{\delta_i} S(y_i|r, \mu_i) = \prod_{i=1}^{N} (r\mu_i)^{\delta_i} y_i^{(r-1)\delta_i} e^{-\mu_i y_i^r}$$

Building on this, the single model approach to forecasting mortality can be extended into a model averaging approach in which the description of the data comes from a group of k models

$$g(y_i, \delta_i|k, \mathbf{r}, \boldsymbol{\mu}) = \sum_{j=1}^{k} w_j f(y_i, \delta_i|r_j, \boldsymbol{\mu}) \tag{5.1}$$
$$\mathbf{w} \sim Dirichlet(\boldsymbol{\phi})$$

in which $\mathbf{r}$ and $\boldsymbol{\mu}$ are vectors of shape and scale parameters respectively, with both corresponding to each of the k Weibull models. The model weights are $\mathbf{w}$ which have priors of $\boldsymbol{\phi} = (\phi_1, ...\phi_k)$. The likelihood of the combined models can then be shown as

$$L(\mathbf{w}, \mathbf{r}, \boldsymbol{\mu}|k, \{(y_1, \delta_1), ..., (y_N, \delta_N)\}) = \prod_{i=1}^{N} \sum_{j=1}^{k} w_j (r_j \boldsymbol{\mu})^{\delta_i} y_i^{(r_j-1)\delta_i} e^{-\boldsymbol{\mu} y_i^{r_j}} \tag{5.2}$$

with the likelihood of the combination of models known, priors can be assigned to each parameter to form a joint posterior distribution from which posterior inference is carried out.

Taking a closer look at the model selection weighting method shown in Section 2.2.1 a similarity can be found with that of a mixture model approach. A key difference between the mixture model and model selection approach involves how the model parameters are estimated, a comparison further explored in Sections 5.2.1 and 5.2.2.

5.2.1 Mixture Modelling when using Weibull models

All material in this subsection (5.2.1) is from Marín et al. (2005). It is shown here how to set up the mixture framework for a set of Weibull models, which can then be adapted in Section 5.2.2 to a model selection approach to estimate model weights in a model averaging approach.

Marín et al. (2005) proposed a similar, albeit simpler (parameters are scalar, not a function of birth cohort), mixture of Weibull distributions for modelling heterogeneous survival data. Placed in the Bayesian paradigm, they developed a Gibbs sampling algorithm for parameter estimation. Gibbs sampling is a method of MCMC used to approximate posterior distributions and uses an iterative procedure that updates each parameter based of the parameter values found in the previous steps. For an example of Gibbs sampling consider a parameter set $\mathbf{\Theta} = \{\Theta_1, \Theta_2, ..., \Theta_N\}$ and the full conditional distribution of each parameter $Pr(\Theta_i|\Theta_{j\neq i}, \mathbf{y})$ is known for $i, j = 1, 2, ..., N$. Firstly set of initial values $\mathbf{\Theta^{(0)}} = \{\Theta_1^{(0)}, \Theta_2^{(0)}, ..., \Theta_N^{(0)}\}$ is given for all parameters, then the iterative procedure can proceed as

1. $\Theta_1^{(t)} \sim Pr(\Theta_1|\Theta_2^{(t-1)}, \Theta_3^{(t-1)}, ..., \Theta_N^{(t-1)})$

2. $\Theta_2^{(t)} \sim Pr(\Theta_2|\Theta_1^{(t)}, \Theta_3^{(t-1)}, ..., \Theta_N^{(t-1)})$

3. ...

4. $\Theta_N^{(t)} \sim Pr(\Theta_N|\Theta_1^{(t)}, \Theta_2^{(t)}, ..., \Theta_{N-1}^{(t)})$

5. t = t+1. Go to Step 1

In this section, $\boldsymbol{\theta}$ is introduced as the set of scalar scale parameters for each model ($\boldsymbol{\theta} = (\theta_1, ...\theta_k)$ for $j = 1, ..., k$) not to be confused with the scale parameters μ_j which are some function of the birth cohort.

The key starting point of their sampling scheme is the introduction of a model indicator z_i, which allocates each individual i to one of the k models (e.g. $z_i = j$ means the observation of individual i is modelled by the j^{th} model). In fact, using these model indicators in estimating the mixture weights is a standard practice in mixture modelling (Diebolt and Robert, 1994).

With the introduction of z_i, the likelihood shown in Equation 5.2 simplifies to

$$L(k, \mathbf{w}, \mathbf{r}, \boldsymbol{\theta}|\text{data}) \propto \prod_{j=1}^{k} (\theta_j r_j)^{\bar{n}_j} \exp\Big(r_j \sum_{i:z_i=j}^{N} \delta_i \log(y_i) - \theta_j \sum_{i:z_i=j}^{N} y_i^{r_j} \Big) \tag{5.3}$$

where $\bar{n}_j = \#\{i : z_i = j \text{ and } \delta_i = 1\}$ is the total number of uncensored observations assigned to model j for $j = 1, ..., k$.

Prior distributions are allocated to the model weight and each model parameter

$$\begin{aligned} \mathbf{w}|k &\sim Dirichlet(\phi_1, ..., \phi_k) \\ r_j &\sim Gamma(\alpha_{r,j}, \beta_{r,j}), \qquad \text{for } j = 1, ..., k \\ \theta_j &\sim Gamma(\alpha_{\theta,j}, \beta_{\theta,j}), \qquad \text{for } j = 1, ..., k \end{aligned}$$

Here the notation for the priors of the model specific parameters ($\alpha_{r,j}$ etc) allows for flexibility when setting different priors to each model. For simplicity the dependence of the model is

dropped ($\alpha_{r,j} = \alpha_r$ etc) however it can be easily brought back into the conditional posterior distribution for a given model parameter. Using the likelihood function and prior distributions, four full conditional distributions can be derived:

$$\mathbf{w}|k, \mathbf{z}, \boldsymbol{\theta}, \mathbf{r}, \text{data} \sim Dirichlet(\phi_1 + n_1, ..., \phi_k + n_k) \tag{5.4}$$

where $n_j = \#\{i : z_i = j\}$ is the total number of observations assigned to model for $j = 1, ..., k$.

$$\theta_j|k, \mathbf{z}, \mathbf{r}, \text{data} \sim Gamma(\bar{n}_j + \alpha_\theta,\ \beta_\theta + \sum_{i:z_i=j}^{N} y_i^{r_j}) \tag{5.5}$$

$f(r_j \mid k, \mathbf{z}, \boldsymbol{\theta}, \text{data}) \propto g(r_j)$ where

$$g(r_j) = r_j^{\bar{n}_j+\alpha_r-1} \exp\Big\{ - r_j\Big(\beta_r - \sum_{i:z_i=j}^{N} \delta_i \log(y_i)\Big) - \theta_j \sum_{i:z_i=j}^{N} y_i^{r_j}\Big\} \tag{5.6}$$

$$\Pr(z_i = j|k, \mathbf{w}, \mathbf{r}, \boldsymbol{\theta}, \text{data}) \propto w_j (r_j\theta_j)^{\delta_i} y_i^{(r_j-1)\delta_i} e^{-\theta_j y_i^{r_j}} \tag{5.7}$$

Using these full conditionals, Marín et al. (2005) use the following Gibbs sampling procedure to sample each model parameter

1. t = 0. Set initial values $\mathbf{w}^{(0)}, \boldsymbol{\theta}^{(0)}, \mathbf{r}^{(0)}$

2. $z_i^{(t+1)} \sim z_i|k, \mathbf{w}^{(t)}, \boldsymbol{\theta}^{(t)}, \mathbf{r}^{(t)}$, data, for $i = 1, ..., N$
3. $\mathbf{w}^{(t+1)} \sim \mathbf{w}|k, \mathbf{z}^{(t+1)}, \boldsymbol{\theta}^{(t)}, \mathbf{r}^{(t)}$, data
4. $\theta_j^{(t+1)} \sim \theta_j|k, \mathbf{z}^{(t+1)}, \mathbf{w}^{(t+1)}, \mathbf{r}^{(t)}$, data, for $j = 1, ..., k$
5. $r_j^{(t+1)} \sim r_j|k, \mathbf{z}^{(t+1)}, \mathbf{w}^{(t+1)}, \boldsymbol{\theta}^{(t+1)}$, data, for $j = 1, ..., k$
6. $t = t + 1$. Go to Step 2

A key feature of the mixture model is how the model parameters are estimated through Equations 5.5 and 5.6. The two conditional posterior distributions show that both shape and scale estimates are dependent on the subset of data assigned to them (through the model selection parameter z_i).

5.2.2 Model Selection weights when using Weibull models

To estimate weights for a model averaging approach using the model selection technique (shown in Chapter 2) the goal is to choose between models fitted to the entire dataset and give a higher weight to those which provide better descriptions of the data. Notice that here for a forecasting model, the scale is no longer scalar and returns to being a function of birth cohort. The estimation of μ_j and r_j is to be done independently of the data assigned to each model with two options available as to how this can be done.

Firstly the Gibbs sampling procedure can be performed as shown in Section 5.2.1 if the full conditionals of each model parameter is derived. However in Steps 4 and 5 a dummy variable $\mathbf{D}$ is to be used instead of $\mathbf{z}$ in which $D_i = j$, for $i = 1, ..., N$ and $j = 1, ..., k$. Replacing Steps 4 and 5 with

4. $\mu_j^{(t+1)} \sim \mu_j | k, \mathbf{D}, \mathbf{r}^{(t)}$, data, for $j = 1, ..., k$
5. $r_j^{(t+1)} \sim r_j | k, \mathbf{D}, \boldsymbol{\mu}^{(t+1)}$, data, for $j = 1, ..., k$

Now the estimation of parameters in each model is carried out using the full dataset.

In this forecasting setting, and the use of the dummy variable $\mathbf{D}$, the full conditional of r_j can remain the same (given a Gamma prior). However with μ_j being some function of birth cohort its conditional can be more complex to derive. Due to the independence amongst models, each model can be estimated separately from the Gibbs sampling algorithm, with each model using

the full dataset. Model iterations/posterior samples can be used at the iteration level instead of each simulation in Steps 4 and 5. Note that for both methods only Equations 5.5 and 5.6 are affected and Equations 5.4 and 5.7 remain the same.

All that is then required to start the iterative process are initial values for the weights $\mathbf{w}$. Here using equal model weights as initial values is recommended to allow for no bias when assigning model weights.

5.3 Simulation setting

To show an application of the model selection weighting method for survival models, a simulation setting is used to compare the performance of model weights assigned to three models. Here five birth cohorts with 2000 observations each are simulated through the posterior means of both shape and scale parameters found using the spline model in Section 4.4.3.1 for the twelve to sixty months age group. This age group is chosen as the mortality rate for that group shows a clear drop, favouring a spline model. The posterior mean of the shape parameter was found as 0.52, therefore limiting how well an exponential model (with its shape fixed at 1) can follow the data. The posterior means of the five birth cohorts chosen to simulate from are cohorts 1997-2001 as the knot of the spline model is at 1999, i.e. the centre of the five cohorts. After simulating from the models, to keep the structure of the mortality data, any simulation time found to be greater than 1 is right censored at 1 (note that when using INLA the observations are scaled to which 1 is the maximum - Section 4.3) as a simulation of with a value of 1 is equivalent to a child dying at 60 months old (age five).

The model ensemble used consists of three models. Two Weibull models, one with a linear trend across birth cohort and the other with a piecewise linear spline on birth cohorts. The third model used follows an exponential distribution with a linear trend placed across birth cohort. The specification of the models are

$$\begin{aligned} &\textit{Linear Model} \\ y_i &\sim Weibull(r, \mu_{C_i}) \\ r &\sim Gamma(1,1) \\ \\ log(\mu_{C_i}) &= \alpha + \beta \times C_i \\ \\ \alpha &\sim N(0, 10000) \\ \beta &\sim N(0, 10000) \end{aligned}$$

$$\begin{aligned} &\textit{Spline Model} \\ y_i &\sim Weibull(r, \mu_{C_i}) \\ r &\sim Gamma(1,1) \\ \\ log(\mu_{C_i}) &= \alpha + \begin{cases} \beta^{(0)} \times C_i, & \text{if } C_i \leq k \\ \beta^{(0)} \times k + \beta^{(1)} \times (C_i - k), & \text{if } C_i > k \end{cases} \\ \\ \alpha &\sim N(0, 10000) \\ \beta^{(0)} &\sim N(0, 10000) \\ \beta^{(1)} &\sim N(0, 10000) \end{aligned}$$

The linear exponential model can be shown as a special case of the the linear Weibull model with the shape r being fixed at 1. These three models are chosen as it is expected there will be a clear ordering of which model is better suited to the data. This is expected as the data are simulated using a shape parameter of 0.52 favouring the Weibull models over the exponential model. Similarly, as the data are simulated from the posterior means found from the spline model in Section 4.4.3.1, it is expected the spline model is better suited to fit the data than the linear. Therefore the expected ordering of models is the spline model to be assigned the most weight, followed by the linear model with the less flexible exponential model given the smallest weight. The three models are fitted using INLA, with posterior samples from

each model parameter used as iterations for obtaining model weights. The prior used for the probability of each model being selected is equal for each model ($\phi_1, \phi_2, \phi_3 = \frac{1}{3}$) allowing for no prior bias towards any weighting preferences.

The posterior means of model weights are shown for three simulations in Table 5.1. These estimates were obtained after running the Gibbs sampling algorithm for 40,000 iterations over two chains in which the first 10,000 iterations of each chain were removed as burn in. This raises a key characteristic of estimating weights separately from model fitting in which extra allowance is needed when gathering iterations of model parameters to allow for a burn in when estimating model weights. The results here show the expected outcome in which over the three simulations the spline model is preferred to the other two, with the linear model being assigned slightly less weight. Whilst doing so the weighting technique is also shown to penalise the exponential model, giving it 8% of the weight in each simulation.

Table 5.1: Posterior means of model weights over three simulations using the survival model averaging approach

	Simulation 1	Simulation 2	Simulation 3
Linear	0.41	0.36	0.43
Spline	0.51	0.56	0.49
Exponential	0.08	0.08	0.08

The estimated scale parameters are shown against the original scale parameters for both linear and spline models in Figure 5.1. With the corresponding shape parameters of $r^{(\text{Sim 1})} = 0.50$ $(0.45, 0.56)$, $r^{(\text{Sim 2})} = 0.53$ $(0.47, 0.6)$ and $r^{(\text{Sim 3})} = 0.53$ $(0.48, 0.6)$ for the linear model and $r^{(\text{Sim 1})} = 0.48$ $(0.43, 0.54)$, $r^{(\text{Sim 2})} = 0.51$ $(0.45, 0.57)$ and $r^{(\text{Sim 3})} = 0.52$ $(0.46, 0.58)$ for the spline model. With each estimated shape parameter being close to the original 0.52, Figure 5.1

shows some explanation for the small differences in weights over the three simulations. When comparing the weights assigned for the linear and spline models across the three simulations, the third simulation views the performance of both models as close to equal whilst the second simulation clearly favours the spline model. This discrepancy across simulations is explained as Figure 5.1 (right) shows that the spline model in the third simulation does not capture as much of a drop as the other two, making the overall model more comparable to the linear model and therefore making it more difficult for the weighting method to distinguish between the two models, justifying the more even weights in this simulation. For an example of some of the output available from a simulation, Figure 5.2 shows the five simulated survival functions plotted with the in-sample fitting of each model, alongside the history and density plots of the model weights (in which the density has accounted for burn in).

Figure 5.1: Showing the fitted scale parameters against the original scales for the linear and spline models over the three simulations

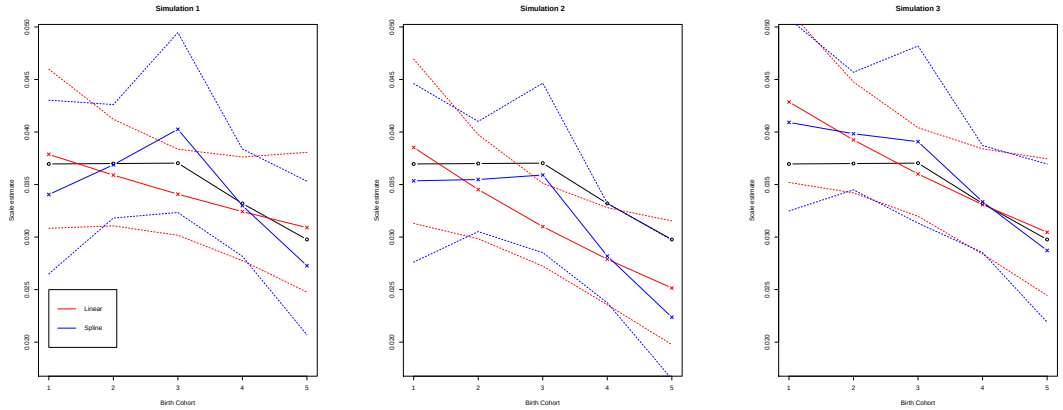
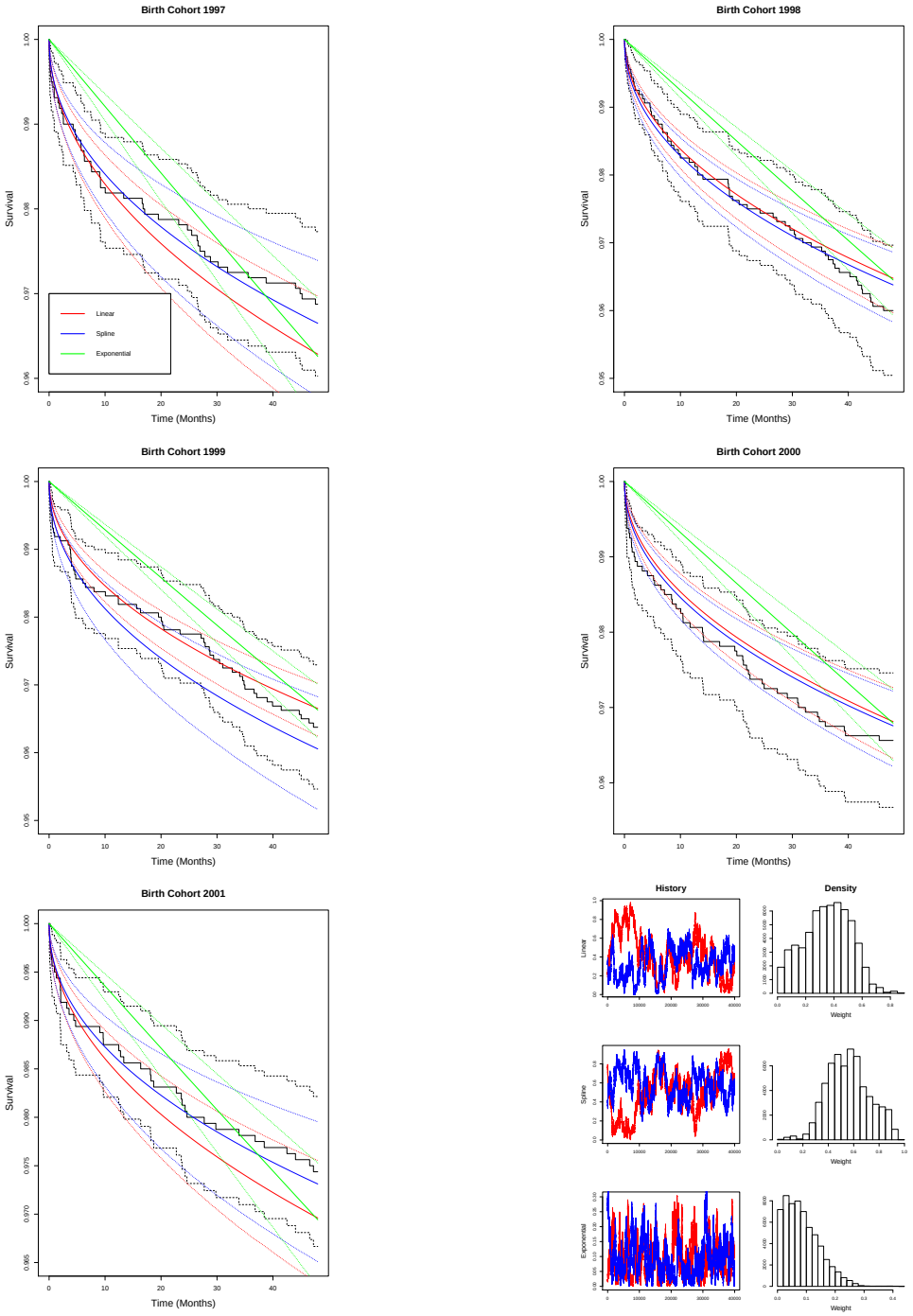


Figure 5.2: The model weights and fits from the first simulation when using a survival model averaging approach



As an extension to this simulation setting, the weighting system was tested to see how the weights perform when simulating the dataset from not only the Weibull spline model, but also the linear Weibull and Exponential models. Similarly as for when using the spline model to simulate the dataset, the parameter values for the linear Weibull and Exponential models are the posterior means from when fitting the models to the Kenya twelve to sixty age group. This allows the the parameter estimates to be similar to as found in the under-5 mortality forecasting setting. To allow for any more complete picture as to how the weights are assigned each dataset is simulated 50 times, with the Gibbs sampling weighting method running for 40,000 iterations and discarding the first 10,000 for burn in. The average of the posterior means of the model weights over the 50 simulations are reported in Table 5.2 for each of the simulated datasets. It can be found that when simulating the dataset from the exponential model both Weibull models are flexible enough to be able to perform similarly to the Exponential model, resulting in roughly equal weights across the three models. As the Exponential dataset is transitioned to the Weibull linear dataset, the need for the Weibull shape parameter can be found as both Weibull models are able to better capture the changes made to the survival times when using a more realistic shape parameter of $r = 52$. Since the Weibull models are able to capture this change and the Exponential is not, the Exponential model is given a low weight of 7% whilst the two Weibull models remain with roughly equal weights. When the further layer of complexity is added by means of simulating the dataset from the spline model, rather than the linear model, the ability to capture this change is represented in the model weights as the spline model starts to consistently be assigned a higher weight than the linear model. When viewing the different weightings across all three dataset types, it is clear to see that the weighting system is able to distinguish the more flexible models from the least flexible, as when the dataset is considered

Exponential all three models perform relatively equally, whereas with the Weibull spline dataset both the Exponential and Weibull linear models tend to not be able to capture the data as well as the spline model.

Table 5.2: Average posterior means of model weights when simulating the data from different models

	Exponential Data	Linear Data	Spline Data
Exponential weight	0.32	0.07	0.07
Linear weight	0.33	0.45	0.42
Spline weight	0.35	0.48	0.5

5.4 Discussion

With this approach to model averaging when using survival models still being a work in progress it leaves further challenges ahead in not only application but also with the weighting method. A first note is that if this method is to be applied to the Weibull models shown in Chapter 4, an obvious feature is to have age group specific model weights due to the independent fitting of each age group. Note that this is also shown to be the preferred option in the Poisson framework.

Another challenge involves how the models are combined after having an assigned weight for each model. The combination of models through using proportions of iterations equal to the models assigned weight is shown in Chapter 2. To follow the same procedure in the survival setting there are a number of options available as to how the weights can be applied. When using the Weibull models (as used here) the first thing to keep in mind is the dependence

between the shape and scale parameters. Since for any model the two parameters are estimated jointly, it is vital to keep the same pairings when combining models. One of the options available here is to sample a proportion of model iterations equivalent to the model weight in which the selected iterations are fixed for both shape and scale parameter. If the model averaging survival curve is then wanted, each survival curve can be calculated at the iteration level in which the resulting set of survival functions can be used as the full model averaging distribution. It should be noted that this procedure is only applicable at the age group level. If the mortality rate is wanted the ages at death can be simulated at the iterative level, as shown in Section 4.4.1, through using the combined iteration pool for model averaging.

However if the mortality rate is the only wanted output there are two other options that can be considered. Firstly the model weights could be applied at the simulation stage in which the number of simulations used from each model is proportional to the model weight. For example when simulating for the first age group, if 2000 simulations are to be done and a model has an assigned weight of 0.5 that model is only used to simulate for 1000 of the total set. Then the same approach can be taken with the remaining simulations for the other age groups. Secondly the model weights can be introduced at the final stage (after the mortality rates have been calculated for each model). This method would be the closest to the application of weights shown in Chapter 2 in which each model has iterations of mortality rate forecasts in which samples of these iterations can be taken with size proportional to each model weight, resulting in the pooled iteration set representing the model averaging forecasts.

With the weighting system and then application of the weights in place, an application of the model averaging can be performed. A integral part of the model averaging process is the selection of models to use. As mentioned in Section 4.5 an option that can be used is a series of spline models, similar to as shown here, can be used in the model averaging group in which each spline model has the knot placed in a different location. The combination of this set of models allows for any uncertainty around the true knot position to be encompassed in the model uncertainty the model averaging accounts for.

Chapter 6

Discussions and Conclusions

In this study an improved method to forecast life expectancy has been developed. This method incorporates the three main sources of uncertainty, being data, parameter and model uncertainty. In cross validation studies, this has allowed for not only consistently accurate forecasts but also forecasts with more appropriate uncertainty bounds than that of the single model approaches. This method is also compared to the BMA approach shown by Kontis et al. (2017) and is shown to have the following desirable features

- More flexible weighting specification, allowing for age group specific weights
- Model weights can be estimated using a single model fitting
- Less care is needed when preselecting models

A new approach to forecasting under-five mortality is then developed for when using individual level survey data. Forecasts from this survival analysis approach are then compared with

forecasts from different variations of the popular Lee-Carter model, with the survival approach shown to provide the better forecasts overall. Finally, steps are taken to incorporate model uncertainty in the survival analysis approach, with a similar method to obtaining model averaging weights as shown when using the area level data.

Throughout the thesis, further extensions specific to the corresponding part of the study are shown at the end of the relevant chapter. Here three more general extensions are mentioned, highlighting their role in progressing this field of research.

6.1 Different specifications in weighting systems

Throughout the study three different weighting systems have been considered, two of them being very similar. The two similar methods can be considered as two different specifications of the model selection approach whilst the third is the two-staged approach shown by Kontis et al. (2017). The two model selection methods share similarities in that they are both based on the same model selection framework. However both are applied differently. The main difference between the two approaches is that the model selection variable (Z_i) is assigned at the age group level when using the area level data and at the observation level when using the individual level data. A full investigation outlining the effects of this difference is still to be done. However it is expected that allowing the model selection to be at the observation level will be a more lenient weighting system than selecting a model at the age group level. This is expected as when assigning at the iteration level the overall best fitting model for that iteration is to be selected, whereas at the observation level although the overall best fitting

model will have the higher proportion of assignments, other models may also be selected for some observations. In terms of forecast performance, a more lenient weighting system is more likely to provide larger uncertainty intervals (as found when using the two-staged approach). Therefore using the model selection at the observation level may help to improve the coverage of the forecasts.

6.2 Incorporating individual level information

One of the areas that the project can be extended is by means of including more individual levelled information, which can be incorporated using the proposed survival approach. This information can be included by means of using additional covariates and, in the case of the DHS survey data, the usage of the survey weights. The inclusion of individual level covariates can allow for the model to capture different mortality patterns, depending on the lifestyle of the child. As mentioned in Section 2.5, the inclusion of additional covariates can bring challenges in terms of forecasting. For example the need to forecast the covariate, or use a lagged approach to acquire the overarching mortality forecast. However these added covariates can also allow for more specific mortality forecasts, such as the usage of spatial information allowing for subnational forecasts.

However (in the context of using survey data) to obtain accurate forecasts, not only at the subnational level but also at the national level, it is vital to incorporate survey weights. Survey weights are allocated to each observation in the surveys, and are intended to be used to allow for information found using the surveyed population to become representative of the overall

population. When using survey weights, an observation with a higher weight is viewed to be more important than a one with lower weight. In this line of thought, the survey weight can be incorporated into the likelihood at the given observation. Using the Cox PH model for time to event data the same logic is used, in which sample weights are incorporated through multiplying each weight to the partial likelihood (Gardiner, 2015). In the same process, one way to incorporate sample weights here can be to weight the Weibull likelihood at the observation level.

6.3 Further exploration of the model selection framework

The findings in Chapter 4 show that the Weibull spline model produced the best overall forecasts. Despite this a model averaging approach, which can include the spline model, can be used to incorporate model uncertainty into the forecasts. When considering other competing models to use in the ensemble, one option is to consider other specifications of Weibull models. This step is taken in Chapter 5 in which it is shown how to estimate model weights for the survival models. However the results in Chapter 4 also show the linear Weibull model to provide poor forecasts in comparison to the different variations of Lee-Carter model. This raises the question as to whether it is possible to combine forecasts from the Lee-Carter models with the Weibull models when using a model selection approach. The key difference between the two model types is that the Weibull models use individual level data, whilst the Lee-Carter models use aggregated data. This difference means that one observed datapoint when using the aggregated data can be considered to be a combination of multiple individual level observations.

A possible step to being able to use both types of model in a model selection framework can be to estimate weights at the aggregated data level. The output from the Weibull models can be transformed into death counts. This can be done using the method outlined in Section 4.4.1. Then the model selection method can be carried out in the Poisson setting, as in Chapter 2. In the same manner, forecasts from different types of models can be combined when using a two-staged style approach. All that is required is that each model used has to be able to forecast on the same scale. For example, in the case of the two-staged approach shown by Kontis et al. (2017) each model must be able to provide life expectancy forecasts. Overall, the ability to include these different types of model allows for a wider variety of models to be considered, bringing the potential for even better mortality forecasts.

Bibliography

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE transactions on automatic control* **19**(6), 716–723.

Ali, Z., Hosseini, M., Mahmoodi, M., Mohammad, K., Zeraati, H. and Naieni, K. H. (2015), 'A comparison between accelerated failure-time and Cox proportional hazard models in analyzing the survival of gastric cancer patients', *Iranian journal of public health* **44**(8), 1095.

Ayele, D. G., Zewotir, T. T. and Mwambi, H. (2017), 'Survival analysis of under-five mortality using cox and frailty models in ethiopia', *Journal of Health, Population and Nutrition* **36**(1), 25.

Basnayake, B. and Nawarathna, L. (2017), 'Modeling and forecasting norway mortality rates using the lee carter model', *Biom Biostat Int J* **6**(1), 00158.

Benchimol, A. G., Alonso, P. J., Marín Díazaraque, J. M. and Albarrán Lozano, I. (2016), 'Model uncertainty approach in mortality projection with model assembling methodologies'.

Bennett, J. E., Li, G., Foreman, K., Best, N., Kontis, V., Pearson, C., Hambly, P. and Ezzati, M. (2015), 'The future of life expectancy and life expectancy inequalities in england and wales: Bayesian spatiotemporal forecasting', *The Lancet* **386**(9989), 163–170.

Blangiardo, M. and Cameletti, M. (2015), *Spatial and spatio-temporal Bayesian models with R-INLA*, John Wiley & Sons.

Booth, H. and Tickle, L. (2008), 'Mortality modelling and forecasting: A review of methods', *Annals of actuarial science* **3**(1-2), 3–43.

Branson, M. and Whitehead, J. (2002), 'Estimating a treatment effect in survival studies in which patients switch treatment', *Statistics in medicine* **21**(17), 2449–2463.

Butt, Z., Haberman, S. and Shang, H. L. (2019), 'The ilc package', *Retrieved on* **27**.

Cairns, A. J., Blake, D. and Dowd, K. (2006), 'A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration', *Journal of Risk and Insurance* **73**(4), 687–718.

Chaloner, K. (1991), 'Bayesian residual analysis in the presence of censoring', *Biometrika* **78**(3), 637–644.

Clark, T. G., Bradburn, M. J., Love, S. B. and Altman, D. G. (2003), 'Survival analysis part I: basic concepts and first analyses', *British journal of cancer* **89**(2), 232–238.

Clayton, D. and Schifflers, E. (1987), 'Models for temporal variation in cancer rates. ii: age–period–cohort models', *Statistics in medicine* **6**(4), 469–481.

Costa, J. C., da Silva, I. C. M. and Victora, C. G. (2017), 'Gender bias in under-five mortality in low/middle-income countries', *BMJ global health* **2**(2).

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2), 187–202.

Croft, T. N., Aileen, M., Courtney, K. et al. (2018), 'Guide to DHS statistics: DHS-7', *Rockville, Maryland, USA2018* .

DHS, M. et al. (2013), 'Standard recode manual for DHS 6', *Calverton, MD: MEASURE DHS, USAID* .

Diebolt, J. and Robert, C. P. (1994), 'Estimation of finite mixture distributions through bayesian sampling', *Journal of the Royal Statistical Society: Series B (Methodological)* **56**(2), 363–375.

Efron, B. (1992), Bootstrap methods: another look at the jackknife, *in* 'Breakthroughs in statistics', Springer, pp. 569–593.

Erişoğlu, Ü., Erişoğlu, M. and Erol, H. (2011), 'A mixture model of two different distributions approach to the analysis of heterogeneous survival data', *International Journal of Computational and Mathematical Sciences* **5**(2), 75–79.

Falk, M. (2013), 'A survival analysis of ski lift companies', *Tourism Management* **36**, 377–390.

Fang, X., Li, R., Kan, H., Bottai, M., Fang, F. and Cao, Y. (2016), 'Bayesian model averaging method for evaluating associations between air pollution and respiratory mortality: a time-series study', *BMJ open* **6**(8), e011487.

Foreman, K. J., Li, G., Best, N. and Ezzati, M. (2017), 'Small area forecasts of cause-specific mortality: application of a Bayesian hierarchical model to US vital registration data'.

Gardiner, J. (2015), 'Survival analysis with Survey Data'.
URL: *https://support.sas.com/resources/papers/proceedings15/2040-2015.pdf*

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995), 'Efficient parametrisations for normal linear mixed models', *Biometrika* **82**(3), 479–488.

Gelman, A. et al. (2006), 'Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)', *Bayesian analysis* **1**(3), 515–534.

Girosi, F. and King, G. (2008), *Demographic forecasting*, Princeton University Press.

Goel, M. K., Khanna, P. and Kishore, J. (2010), 'Understanding survival analysis: Kaplan-Meier estimate', *International journal of Ayurveda research* **1**(4), 274.

Goldring, S., Henretty, N., Mills, J., Johnson, K. and Smallwood, S. (2011), 'Mortality of the 'golden generation': What can the ons longitudinal study tell us?', *Population trends* **145**(1), 203–232.

Gompertz, B. (1825), 'On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies', *Philosophical transactions of the Royal Society of London* (115), 513–583.

Hansen, B. E. and Racine, J. S. (2012), 'Jackknife model averaging', *Journal of Econometrics* **167**(1), 38–46.

Hill, K. (2011), 'Direct estimation of child mortality from birth histories', *Tools for D* p. 166.
URL: *http://demographicestimation.iussp.org/content/direct-estimation-child-mortality-birth-histories*

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999), 'Bayesian model averaging: a tutorial', *Statistical science* pp. 382–401.

Hougaard, P. (1995), 'Frailty models for survival data', *Lifetime data analysis* **1**(3), 255–273.

Humble, R., Love, H. and Ryan, D. (2006), Analysis of trends in mortality near or during retirement for four European countries, *in* '28th International Congress of Actuaries, May'.

Hyndman, M. R. J. (2012), 'Package 'demography''.

ICF International (1984), 'The DHS program'. [Online; accessed 13-December-2018].
URL: *https://dhsprogram.com*

ICF International (2012), 'Demographic and health survey sampling and household listing manual'.

Jacobson, T. and Karlsson, S. (2004), 'Finding good predictors for inflation: A bayesian model averaging approach', *Journal of Forecasting* **23**(7), 479–496.

Jarl, J., Johansson, P., Eriksson, A., Eriksson, M., Gerdtham, U.-G., Hemström, Ö., Selin, K. H., Lenke, L., Ramstedt, M. and Room, R. (2008), 'The societal cost of alcohol consumption: an estimation of the economic and human cost including health effects in Sweden, 2002', *The European Journal of Health Economics* **9**(4), 351–360.

Kang, H. (2013), 'The prevention and handling of the missing data', *Korean journal of anesthesiology* **64**(5), 402.

Kardaun, O. (1983), 'Statistical survival analysis of male larynx-cancer patients-a case study', *Statistica neerlandica* **37**(3), 103–125.

Khalaf, A., Hamam, Y., Alayli, Y. and Djouani, K. (2013), 'The effect of maintenance on the survival of medical equipment', *Journal of Engineering, Design and Technology* .

Kontis, V., Bennett, J. E., Mathers, C. D., Li, G., Foreman, K. and Ezzati, M. (2017), 'Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble', *The Lancet* **389**(10076), 1323–1335.

Kontis, V., Bennett, J. E., Rashid, T., Parks, R. M., Pearson-Stuttard, J., Guillot, M., Asaria, P., Zhou, B., Battaglini, M., Corsetti, G. et al. (2020), 'Magnitude, demographics and dynamics of the effect of the first wave of the COVID-19 pandemic on all-cause mortality in 21 industrialized countries', *Nature Medicine* pp. 1–10.

Lee, R. D. and Carter, L. R. (1992), 'Modeling and forecasting US mortality', *Journal of the American statistical association* **87**(419), 659–671.

Li, N., Lee, R. and Tuljapurkar, S. (2004), 'Using the Lee-Carter method to forecast mortality for populations with limited data', *International Statistical Review* **72**(1), 19–36.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000), 'WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility', *Statistics and computing* **10**(4), 325–337.

Marín, J. M., Rodriguez-Bernal, M. and Wiper, M. P. (2005), 'Using weibull mixture distributions to model heterogeneous survival data', *Communications in Statistics-Simulation and Computation* **34**(3), 673–684.

Mathers, C. D., Sadana, R., Salomon, J. A., Murray, C. J. and Lopez, A. D. (2001), 'Healthy life expectancy in 191 countries, 1999', *The Lancet* **357**(9269), 1685–1691.

McLachlan, G. J. and Basford, K. E. (1988), *Mixture models: Inference and applications to clustering*, Vol. 38, M. Dekker New York.

Mejía-Guevara, I., Zuo, W., Bendavid, E., Li, N. and Tuljapurkar, S. (2019), 'Age distribution, trends, and forecasts of under-5 mortality in 31 sub-Saharan African countries: A modeling study', *PLoS medicine* **16**(3), e1002757.

Mercer, L. D., Wakefield, J., Pantazis, A., Lutambi, A. M., Masanja, H. and Clark, S. (2015), 'Space-time smoothing of complex survey data: small area estimation for child mortality', *The annals of applied statistics* **9**(4), 1889.

Min, Y., Zhang, G., Long, R. A., Anderson, T. J. and Ohland, M. W. (2011), 'Nonparametric survival analysis of the loss rate of undergraduate engineering students', *Journal of Engineering Education* **100**(2), 349–373.

Mohammed, Y. A., Yatim, B. and Ismail, S. (2014), A parametric mixture model of three different distributions: An approach to analyse heterogeneous survival data, *in* 'AIP Conference Proceedings', Vol. 1605, American Institute of Physics, pp. 1040–1045.

Mohammed, Y. A., Yatim, B. and Ismail, S. (2015), 'Mixture model of the exponential, gamma and weibull distributions to analyse heterogeneous survival data', *Journal of Scientific Research and Reports* **5**(2), 132–139.

Murray, C. J., Lopez, A. D., Mathers, C. D. and Stein, C. (2001), 'The Global Burden of Disease 2000 project: aims, methods and data sources', *World Health Organization* **36**, 1–57.

O'Brien, R. M. (2011), 'The age–period–cohort conundrum as two fundamental problems', *Quality & Quantity* **45**(6), 1429–1444.

Parkinson, D. and Liddle, A. R. (2013), 'Bayesian model averaging in astrophysics: a review', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **6**(1), 3–14.

Plummer, M. (2015), 'Cuts in bayesian graphical models', *Statistics and Computing* **25**(1), 37–43.

Preston, S. H., Guillot, M. et al. (2000), *Demography Measuring and Modeling Population Processes*, Blackwell Publishing.

R Core Team (2013), 'R: A language and environment for statistical computing'.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005), 'Using bayesian model averaging to calibrate forecast ensembles', *Monthly weather review* **133**(5), 1155–1174.

Remund, A., Camarda, C. G., Riffe, T. et al. (2017), A cause-of-death decomposition of the young adult mortality hump, Technical report, Max Planck Institute for Demographic Research, Rostock, Germany.

Rue, H., Martino, S. and Chopin, N. (2009), 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations', *Journal of the royal statistical society: Series b (statistical methodology)* **71**(2), 319–392.

Rutstein, S. O. and Rojas, G. (2006), 'Guide to DHS statistics', *Calverton, MD: ORC Macro* **38**.

Schwarz, G. et al. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Shair, S., Purcal, S. and Parr, N. (2017), 'Evaluating extensions to coherent mortality forecasting models', *Risks* **5**(1), 16.

Shang, H. L. and Haberman, S. (2018), 'Model confidence sets and forecast combination: an application to age-specific mortality', *Genus* **74**(1), 19.

Sloughter, J. M., Gneiting, T. and Raftery, A. E. (2010), 'Probabilistic wind speed forecasting using ensembles and bayesian model averaging', *Journal of the american statistical association* **105**(489), 25–35.

Tuljapurkar, S., Li, N. and Boe, C. (2000), 'A universal pattern of mortality decline in the g7 countries', *Nature* **405**(6788), 789–792.

UN (2000), 'Millennium development goals'. [Online; accessed 23-August-2020].
URL: *https://www.un.org/millenniumgoals/*

UN (2015), 'Sustainable development goals'. [Online; accessed 23-August-2020].
URL: *https://sdgs.un.org*

University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany) (2000), 'Human mortality database'. [Online; accessed: 12th June 2018].
URL: *https://www.mortality.org*

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K. and Clark, S. J. (2019), 'Estimating under-five mortality in space and time in a developing world context', *Statistical methods in medical research* **28**(9), 2614–2634.

Wang, H. and Preston, S. H. (2009), 'Forecasting United States mortality using cohort smoking histories', *Proceedings of the National Academy of Sciences* **106**(2), 393–398.

Wang, X., Yue, Y. R. and Faraway, J. J. (2018), *Bayesian regression modeling with INLA*, CRC Press.

Way, C. (2015), *The millennium development goals report 2015*, UN.

Wiśniowski, A., Smith, P. W., Bijak, J., Raymer, J. and Forster, J. J. (2015), 'Bayesian population forecasting: extending the lee-carter method', *Demography* **52**(3), 1035–1059.

Wright, J. H. (2008), 'Bayesian model averaging and exchange rate forecasts', *Journal of Econometrics* **146**(2), 329–341.

Yao, Y., Vehtari, A., Simpson, D., Gelman, A. et al. (2018), 'Using stacking to average bayesian predictive distributions (with discussion)', *Bayesian Analysis* **13**(3), 917–1007.